Classification of Science, Technology and Medicine (STM) Domains with PSO and NBC

Erfian Junianto*  
Universitas BSI  
Bandung, Indonesia

Mayya Nurbayanti Shobary  
Universitas BSI  
Bandung, Indonesia

Rizal Rachman  
STMIK Nusa Mandiri Jakarta  
Jakarta, Indonesia

Ai Ilah Warnilah  
AMIK BSI Tasikmalaya  
Tasikmalaya, Indonesia

Bambang Kelana Simpony  
AMIK BSI Tasikmalaya  
Tasikmalaya, Indonesia

*Corresponding Author: erfian.ejn@bsi.ac.id

Abstract - Science, Technology, and Medicine (STM) is a field of research that has a characteristic in each document. These characteristics are different from most documents that are used as a corpus in mining text research. Documents derived from Newswire are more dominant in previous research. However, in this study will try to classify documents from STM field. Complex technical terms, symbols, position information, and the number of citations would be a challenge itself. Previous studies have used the Naive Bayes Classifier (NBC) classification method. There are also those who apply Particle Swarm Optimization to assist its classification. From the Newswire field generated a fairly high accuracy Therefore, it would be applied to the optimization method with PSO and combine it with NBC method. This study produced accuracy value in classification model without using PSO equal to 82.73%. While in the classification model using PSO, the accuracy value is 87.27%. This shows that the use of PSO optimization is very influential on the classification.

Keywords: STM, Classification, Naive Bayes Classifier, Particle Swarm Optimization.

I. INTRODUCTION

Text mining can be defined as a process of digging up information which a user interacts with a set of text documents using analysis tools. The process also is components of data mining is classification. The purpose of the text mining itself is to obtain useful information from a collection of documents. Therefore, the data source used in text mining is a collection of texts that have a format of unstructured or at least semi-structured [1].

A previous study has used many text mining techniques. However, the category under study is still within the scope of the newswire. Such as entertainment, politics, organization, products, and others [2]–[8]. In the case of Science, Technology and Medicine (STM) is very different from the Newswire in general [9]. Content from STM is full of citations, complex technical terms, including symbols and position information, as well as other specific features. This is natural for the reader, but never appears in a text intended for the general public [9].

This research will be carried out classification for STM domains. This classification will use a similar model to previous studies, which is the classification of English-language news documents from 20newsgroup [5]. The method is same as the previous study to use the Naive Bayes Classifier (NBC) as a classification algorithm, and Particle Swarm Optimization (PSO) as an optimization algorithm for feature selection. The study resulted in a fairly high accuracy, that is 85.42%.

Basically, the classification that has involved process Information Retrieval (IR), Natural Language Processing (NLP), Data Mining (DM), and Information Extraction (IE). The process stages are combined in a single workflow [10]. However, the previous feature selection stage needs to be implemented, because it influences the classification [11]. This stage will drop the word which is the feature, but are not relevant or may occur redundant.

Unstructured text document form results in a too large dimension. This issue will affect the classification. To overcome this, it is necessary to apply Particle Swarm Optimization (PSO) method [12]. The advantage of this method is its ability to reduce dimensions, such as the social behavior of animals, for example, birds are flocking to promising positions to reach the right destination in multi-dimensional space [13].

After the implementation of PSO, the next step is to use NBC to classify it. Why NBC is chosen is because of its simplicity and computing speed, but it has high accuracy [14], [15]. In terms of performance, NBC is also quite good against the classification of documents containing numbers and text. However, the data from the document must be represented in vectors first. This is done because generally the classification algorithm can not process documents directly [1].

The most commonly used method is the Term Frequency-Inverse Document (TF-IDF) [16]–[23]. The TF-IDF method is one of the most well known weighting word algorithms, and the accuracy of TF-IDF in classifying articles is very promising, because TF-IDF assesses each word-value by using two approaches, the frequency of terms and how many terms can be found in the [19].
The dataset will be used comes from OA-STM-Corpus Elsevier with a total of 110 training documents and 10 testing documents. The 110 documents consist of 10 categories, which is each category has 11 documents in the form of full text journal articles. The categories of the STM document consist of Agriculture, Astronomy, Biology, Chemistry, Computer Science, Earth Science, Engineering, Materials Science, Mathematics, and Medicine [9]. Based on these data, the documents will be processed using RapidMiner as a tool of analysis. This software was chosen because of its open-source, full-featured, and user-friendly interface [24]–[26].

II. RESEARCH METHOD

The research method that will be used in this research is described in the following figure:

![Fig. 1. Research Method](image)

A. Document Collection

The initial phase of the study was the collection of documents, which came from OA-STM-Corpus Elsevier [9]. There are 110 documents in 10 categories with the following details:

<table>
<thead>
<tr>
<th>No.</th>
<th>Category Name</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Mathematics</td>
<td>10</td>
</tr>
<tr>
<td>2.</td>
<td>Astronomy</td>
<td>10</td>
</tr>
<tr>
<td>3.</td>
<td>Materials_Science</td>
<td>10</td>
</tr>
<tr>
<td>4.</td>
<td>Engineering</td>
<td>10</td>
</tr>
<tr>
<td>5.</td>
<td>Computer_Science</td>
<td>10</td>
</tr>
<tr>
<td>6.</td>
<td>Biology</td>
<td>10</td>
</tr>
<tr>
<td>7.</td>
<td>Agriculture</td>
<td>10</td>
</tr>
<tr>
<td>8.</td>
<td>Chemistry</td>
<td>10</td>
</tr>
<tr>
<td>9.</td>
<td>Earth_Science</td>
<td>10</td>
</tr>
<tr>
<td>10.</td>
<td>Medicine</td>
<td>10</td>
</tr>
</tbody>
</table>

The categories and amounts are pure results from OA-STM-Corpus, the authors do not make changes either in the amount or the contents of the documents.

B. Preprocessing

The next stage is preprocessing, which is an important starting point and stage [27]. Documents should be prepared in such a way as to make it easier to process. At this stage, the documents that will be processed are completely clean and ready for the next stage. The preprocessing stage consists of:

1. Transform to Lower Cases

Entering the initial stage of the process, all characters, especially letters, will be changed into lower case. This is done so that all text data entered into the process becomes equivalent and standard. For example the words "Mathematics", "MaThemaTics", "MatheMatics", and so on will be transformed into "mathematics". So all words have the same meaning when entering the next stage.

2. Tokenization

This stage will break the group of sentences into words, while eliminating punctuation. So that will form a collection of words that are unique and meaningful. For example, there is a sentence as follows:

```
the PETM was a period of geologically-rapid global warming that punctuated a warming Eocene climate 55.8Ma ago (Charles et al., 2011), and saw sea surface temperatures rise by 5-8°C from background levels (Zachos et al., 2005; Sluijs et al., 2007). It was associated with a substantial injection of δ13C-depleted carbon into the ocean-atmosphere system (see Pagani et al., 2006a) over <20ka (Cui et al., 2011), causing a negative carbon isotope excursion (CIE) between -2 and -7‰ in marine and terrestrial sediments (see overview in Schouten et al., 2007) lasting 170ka (Röhl et al., 2007), and a prominent dissolution horizon in the deep sea signifying deep ocean acidification (Kennett and Stott 1991; Zachos et al., 2005).
```

3. Filter Token by Length

After the tokenization stage, the next step is to choose a token from a set of words. This election is intended to remove words that are less meaningful, or do not qualify as words. This stage also allows a reduction of the dimensions in the document. For example, words with less than 3 letters and more than 30 letters will be removed. So the words that are less than 3 letters such as "a", "an", "my", "rd" or words with more than 30 letters like "wwwhhhhhhhhhiiiiiiiiiiiiiiiiii" will be removed so that tokens will be more unique and clean.

4. Stopwords

The stopwords stage is the stage that completes the previous stage where there are words that are between 3 letters up to 30 letters will be discarded if it has no relation with the document. Although the actual words at this stage still appear and still have meaning. Moreover, these words do support a sentence. However, these words cannot be used as tokens, because it is a conjunctive or a description. For example, "the", "this", "that", "his", "the,"
"her," "his", "their", "my", "our", "their", "your" "All", "few", "many", and others.

5. **Porter Stemming**

After going through the stages that make up the token, and the selection of a token that clean of a word that has little meaning or relation to the document, the next step in preprocessing is stemming word. The word that has become the token will be returned to the root. For example, the words "disease", "diseases", "diseased", will be transformed into a root form of "disease".

**C. Vector Creation with TF-IDF**

After the preprocessing stage is complete, the next step is to convert the word set to vector using weighting TF-IDF [19], [28]. From this vector will begin the process of optimization and classification. Illustrations of all stages that can also be referred to as indexing, as follows:

![Indexing a Document](image)

At this stage will be done two stages of term weighting, which will calculate the frequency of occurrence of terms in the document (term frequency), and the total occurrence. Then calculated inverse document frequency (IDF), and generate weight with multiplication between TF and IDF. Example calculation is to use TF-IDF as follows:

<table>
<thead>
<tr>
<th>Term</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>D5</th>
<th>D6</th>
<th>D7</th>
</tr>
</thead>
<tbody>
<tr>
<td>binary</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>computer</td>
<td>0.845</td>
<td>0.845</td>
<td>0.845</td>
<td>0.845</td>
<td>0.845</td>
<td>0.845</td>
<td>0.845</td>
</tr>
<tr>
<td>computer system</td>
<td>0.000</td>
<td>0.845</td>
<td>0.000</td>
<td>0.845</td>
<td>0.845</td>
<td>0.845</td>
<td>0.845</td>
</tr>
<tr>
<td>engineering</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>eps</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>generation</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>graph</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>human</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
</tr>
<tr>
<td>interface</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
</tr>
<tr>
<td>intersection</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>machine</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
</tr>
<tr>
<td>management</td>
<td>0.000</td>
<td>0.100</td>
<td>0.000</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
</tr>
<tr>
<td>minors</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>opinion</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
<td>0.100</td>
</tr>
<tr>
<td>ordered</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>random</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>response</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>system</td>
<td>0.000</td>
<td>0.368</td>
<td>0.736</td>
<td>0.368</td>
<td>0.368</td>
<td>0.368</td>
<td>0.368</td>
</tr>
<tr>
<td>testing</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>time</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>user</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>user interface</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**D. Optimize Weight by PSO**

This stage is also referred to as the optimization which will be selected words that have become token and the most represent the document. The particles in the PSO will seek out and determine which tokens are the best and also best describe the contents of the document. Selection of tokens will reduce the document dimension. However, the content contained by
the document is maintained because the selected tokens strongly represent the document [30].

E. Classification with NBC

After all the steps have been passed, it will produce data that are really ready to be classified. By using the Naïve Bayes Classifier algorithm, this stage will perform statistical calculations. Where will generate the probable value of a document enter into a certain classification. It is expected that the previous process can facilitate classification while improving its accuracy [2], [5], [31], [32].

F. Measurement Values

The final stage is to show the measurement results of the classification phase. The result is a table (confusion matrix) or a graph that shows the accuracy value. At this stage the results will be known.

III. DISCUSSION AND RESULT

Research produces outputs that can be analyzed to obtain useful information. The following elaboration of the results of research that has been done using RapidMiner.

A. Document Collection

This design refers to the proposed-research method in the previous chapter. The first design will contain the document collection stage, which will take from the OA-STM-Corpus Elsevier repository.

B. Preprocessing

Design at this stage will be sequenced in accordance with the diagram in the research method. Starting from Transform Cases, Tokenization, Filter Token by Length, Stopwords Filter, and Stemming by Porter.

C. Optimize Weight by PSO

Design at this stage will use operator optimize weight by PSO from RapidMiner. In it there are subprocesses for classification and measurement.

D. Classification

At this stage, the X-Validation operator will be used to determine the final value of the measurement stage.
Fig. 9. Classification

E. Measurement

The final stage in the research method design is the measurement, which will produce accuracy value.

Fig. 10. Measurement

The end result of this study will compare the classification process without PSO and vice versa. The result of classification without using optimization obtained an accuracy rate of 82.73%, with matrix as follows.

### TABLE 5

<table>
<thead>
<tr>
<th></th>
<th>Agriculture</th>
<th>Astronomy</th>
<th>Biology</th>
<th>Chemistry</th>
<th>Computer_Science</th>
<th>Earth_Science</th>
<th>Engineering</th>
<th>Mathematics</th>
<th>Medicine</th>
</tr>
</thead>
<tbody>
<tr>
<td>pred.</td>
<td>100.0%</td>
<td>100.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100.0%</td>
<td>100.0%</td>
<td>0</td>
<td>100.0%</td>
</tr>
<tr>
<td>true</td>
<td>100.0%</td>
<td>100.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100.0%</td>
<td>100.0%</td>
<td>0</td>
<td>100.0%</td>
</tr>
<tr>
<td>class recall</td>
<td>90.91%</td>
<td>90.91%</td>
<td>100.0%</td>
<td>81.82%</td>
<td>81.82%</td>
<td>90.91%</td>
<td>90.91%</td>
<td>81.82%</td>
<td>90.91%</td>
</tr>
</tbody>
</table>

While the results of the classification of NBC are optimized with PSO produces an accuracy of 87.27%. This shows an increase of 4.54%, with the matrix as follows.

### TABLE 6

<table>
<thead>
<tr>
<th></th>
<th>Agriculture</th>
<th>Astronomy</th>
<th>Biology</th>
<th>Chemistry</th>
<th>Computer_Science</th>
<th>Earth_Science</th>
<th>Engineering</th>
<th>Mathematics</th>
<th>Medicine</th>
</tr>
</thead>
<tbody>
<tr>
<td>pred.</td>
<td>100.0%</td>
<td>100.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100.0%</td>
<td>100.0%</td>
<td>0</td>
<td>100.0%</td>
</tr>
<tr>
<td>true</td>
<td>100.0%</td>
<td>100.0%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100.0%</td>
<td>100.0%</td>
<td>0</td>
<td>100.0%</td>
</tr>
<tr>
<td>class recall</td>
<td>90.91%</td>
<td>90.91%</td>
<td>100.0%</td>
<td>81.82%</td>
<td>81.82%</td>
<td>90.91%</td>
<td>90.91%</td>
<td>81.82%</td>
<td>90.91%</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

This study was conducted to find out the accuracy of STM by using NBC and PSO. A significant result has been obtained from the classification process by applying NBC without PSO and vice versa. That result shows that PSO is able to improve the accuracy of classification, although the used data set is different than the one in other previous studies which used news document.

There are still room for improvement, which will be conducted in the future project, such as a classification based on Self-Training and LDA Topic Models[33].

REFERENCES


