Jurnal
Teknologi

Full Paper

# NEWS CLASSIFICATION WITH HUMAN ANNOTATORS: A CASE STUDY

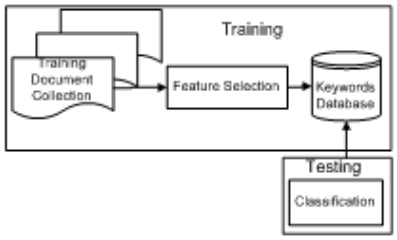
Aini Fuddoly, Jafreezal Jaafar, Norshuhani Zamin*

Department of Computer & Information Sciences, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia

**Article history**
Received
*16 January 2015*
Received in revised form
*24 March 2015*
Accepted
*15 March 2015*

*Corresponding author
norshuhani@petronas.com.my

## Graphical abstract




## Abstract

The need to classify textual documents has become an increasingly vibrant research field due to the development of online news. While most of the news in news websites are categorised manually, the task becomes more strenuous considering the tremendous surge of data updates every day. This paper addresses the question of how text classification algorithms can substitute the particular task over manual classification methods. A combined method using Bracewell's algorithm and top-n method is demonstrated and tested using Indonesian language corpus. The experiment also uses human evaluation as the benchmark. The result from the human evaluation is further investigated in order to understand how the annotators classify documents and the aspects that can be improved to enhance the method in the future. The results indicate that the method can outperform human annotators by 13% in terms of accuracy.

*Keywords*: Bracewell Algorithm, text classification, Indonesian news classification, category classification, topic identification, human annotator

## Abstrak

Keperluan untuk mengklasifikasikan dokumen teks telah menjadi semakin berkembang disebabkan oleh percambahan berita atas talian. Sedang kebanyakan berita atas talian dikategorikan oleh manusia secara manual, tanggungjawab tersebut menjadi semakin sukar apabila jumlah berita atas talian bertambah dengan pesat setiap hari. Penerbitan ini mengisarkan sebuah persoalan bagaimana algoritma klasifikasi teks dapat menggantikan tugas sukar ini jika dilakukan oleh manusia. Penggambungan dua kaedah dengan menggunakann algoritma Bracewell dan top-n di demonstrasikan dan diujikaji dengan menggunakan korpus bahasa Indonesia. Ujikaji juga turut menggunakan ujianoleh manusia sebagai penanda aras. Keputusan ujikaji oleh manusia telah di kajiselidik dengan lebih mendalam untuk memahami proses klasifikasi teks dengan lebih lanjut dan aspek-aspek yang perlu di perbaiki pada kaedah tersebut di masa hadapan. Keputusan kajian telah membuktikan bahawa kaedah yang dicadangkan dapat menandingi keputusan ujian ketepatan manusia sebanyak 13%.

*Kata kunci*: Algoritma Bracewell, klasifikasi teks, klasifikasi berita Indonesia, klasifikasi kategori, pengecaman topic, penganotasi manusia

## 1.0 INTRODUCTION

Text Classification (TC) is the assignment of label(s) to a text record that is a member of a document collection [1]. These labels are commonly known as *class*. This means that a large document collection can be divided into a set of classes and a class consequently contains a group of text documents with similar features. The application of TC has been widely used to organise online documents such as e-mails [2, 3], social networking contents [4], scientific publications, online news articles, etc. [5]

In the development, online news articles have become popular with regards to the growth of Internet. Online news articles are known for its unique characteristics, one of which is the constantly updated streams of data [6]. But the constant update does not necessarily mean a constant content of data. An article may contain a story that has never been covered by any news article stored in the classifier database, because an article inherently follows an event or discovery in the real world, which is impossible to predict. Apart from that, a class that is intuitively known by readers may not be as simple for an algorithm to deal with. For example, a news article about government, policy or election issues can be easily identified by common readers as a part of the "Politics" class or, as generally known in TC, category. But in the case of a TC algorithm, it can be possibly mistaken with "Economics" class, due to the content's words that may be mutually owned by both classes.

Among other fields, TC has been dwelt widely in the past few decades. Astoundingly, there has been a small number of TC algorithms dedicated for Indonesian language; and even less for those that evaluate its performance with human annotators, despite that annotations can yield a significant improvement on a classifier's performance.

This paper presents the application of a TC algorithm based on Bracewell's algorithm for Indonesian news collection. The algorithm essentially divides the news classification into two levels: category and topic [6]. Accordingly, the algorithm is chosen because of its ability in detecting both the category and topic of an article, identifying the category among sparse members and new topics that have never been studied by the classifier. The work presented in this paper also delves on the evaluation with human annotators [7]. Annotations provided by annotators have recently been used in many research works as part of the training phase [8]. Zaidan [17] referred to this as "annotator rationale", where the results from the human annotator is employed as a set of examples during training. However, this technique is most preferable when the training set information itself is not sufficient. In this work, the annotations serve as the benchmark in the testing phase, as to replace a TC algorithm comparison due to the limited number of publicly available algorithms focused on Indonesian corpus. Moreover, discussions are conducted to further analyse how the annotators treat the corpus and on which part can the algorithm advance

## 2.0 RELATED WORK

This section shall address algorithms in the field of TC with regards to two major discussions: several prominent algorithms that have been extensively used in many different corpora and algorithms which apply the use of human annotations.

One approach that has been well-known in the TC area is Naive Bayes [9]. NB classifiers are often referred to as a generative classifier [10] because it creates a probabilistic model that actualizes the assumptions of how example data are generated. The NB classifier has remained superior in the field of TC and has been continuously expanded through many different languages, one of which, being the Arabic language [11]. Noaman *et al*. [11] proposed the application of the NB classifier on 300 volumes of data and acheived high Micro-Average within many categories. However, the work also suggested that there were unstable results over different categories.

*K*-means clustering [12] is an approach that falls under the Unsupervised Learning method. *K*-means clustering works by first selecting an initial set of cluster centroids *k*. The similarities among data points in the collection with the cluster centroids are then computed. Afterwards, the documents with the closest similarity with a centroid are assigned as the member of the centroid's cluster. *K* is recalculated until the global function criterion is minimised or maximised. The downside of this method is the criteria of setting the optimum number of *k* [13]. Many extensive studies have been carried out to improve the performance of *k*-means, such as the bisecting *k*-means algorithm [14], Fuzzy *c*-means [15], *k*-medoid [16], and so forth.

Bracewell's algorithm was proposed in 2009 [6] and was tailored to improve the constant updating problems for online news domain classification, especially for English and Japanese corpora. Bracewell addressed the issue by proposing a method with two properties [6]:

1. The elimination of used training data.
2. Easy update.

These properties set the algorithm apart from the previously explained algorithms such as Naïve Bayes [9] or *k*-means clustering in which the need to easily update the algorithm or to remove the training data is not embodied.

The significance of human annotators in a classification process is explored intensively in the work proposed by Zaidan [17]. The research work suggests that the role of human annotators can be enriched as opposed to simply marking a label to a

collection of documents. In [17], human annotators were also asked to construct a support text containing information that eventually helps them come to their decision. In the process, these informations should be highlighted by the annotators as to indicate a leastwise *why* they chose a particular label. The information, more commonly known as rationale, was then improved by Zaidan [18] to extend the supportive information by directly modelling the annotators' rationales. The same notion is applied to a video classification applied in [19]. In [19], the human annotators were asked to browse through a collection of videos, and mark labels on them based on their judgments. These judgments are applied as the ground truth of the classification process.

An example of a comparison of classification between a text classifier and human classification is presented in [20], applied in the problem of multi label emotions classification in a sentence. The notion that underlies this paper is that a system needs to resemble humanbehaviour as closely as possible. The paper further discussed the correlation between the performance of the machine and the human annotation. Experiment showed that the machine classification could emulate the human annotators' approach.

This paper applies the use of Bracewell's Algorithm due to its overall austerity and efficiency, in the sense that it allows the removal of the training data once it is used, thus making it lightweight; and its simplicity which eases the algorithm's re-training once a new data is classified. In terms of human annotators, annotations are utilised as a method that serves as a benchmark to the Bracewell's algorithm, as opposed to using it as an extension or added information to the training data, because of the limited number of Indonesian TC algorithms available for public use. Moreover, there has also been a small number of works for Indonesian language in which a group of human annotators are involved. This, nevertheless, still implies that the test procedure has to equalise between the human behaviour and the classifier's approach.

## 3.0 PROPOSED METHOD

The Bracewell's algorithm dwells with two different classifications types, categorization and topic identification for news documents. Despite the different formulas employed, the rest of the workflow is largely similar. Category is described as a more general hierarchy of the classification, which does not necessarily reflect the main content or event in the article. It simply defines what kind of story it tells. Topic, however, is described as a more specific classification, which shows the theme or event of the article.

The first step of this algorithm is the training stage. In the training stage, keywords model is generated and kept in the database. The model consists of information such as the frequency of the keywords, the document frequency in the collection of the keywords, and the total number of documents in the category. The method for extraction of keywords is based on [21]. The workflow of the algorithm is illustrated in Figure 1.

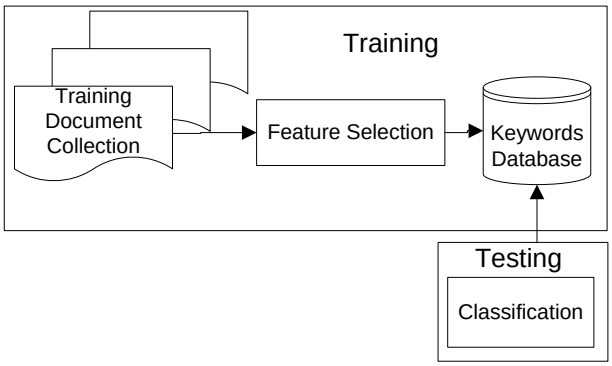



**Figure 1** Bracewell's algorithm workflow

Classification for category is done by calculating the likelihood [6] between documents and the training keyword collection using Equation 1.

$$\text{Likelihood}\ (c_j|\ A=\{k_1, k_2, \dots, k_n\}) = -\ \sum_{i=1}^{n} P(k_i|c_j)\log\left(P(k_i|c_j)\right) \quad (1)$$

A likelihood calculation between the document and the keywords is only possible when keywords from the document have been selected as well. The method of selecting keywords from the document is the same with the method used in the training phase. After the likelihood calculation is yielded for each category, the best category chosen for the article is the one that exceeds the given threshold of the mean of all the likelihood of the category plus one standard deviation.

The method divides the topic identification step into two sub-steps as described in [6], according to the requirements the step has: the first is the topic identification, where the article will be given a previously seen topic, and the second sub-step is the topic discovery where the assigned topic will be further analysed to detect whether the article requires a new topic to be created.

Using this method, keywords in the database are retrieved and transformed into vectors, as well as the keywords from the article input. To create the corresponding vector, both topics and articles are standardised into the same Vector Space Model [22].

Once both vectors are normalized, the calculation of similarity between two vectors is done. In this step, the cosine similarity is used. Similar to what had been

applied in [23], the cosine similarity formula used in Bracewell's algorithm treats the articles and topics as vectors. Equation 2 demonstrates the calculation for cosine similarity [6].

$$\text{CosSim}(t_i, A) = \frac{t_i \cdot A}{|t_i||A|} \quad (2)$$

Equation (2) calculates the cosine between *A*, which denotes article *A* and topic $t_i$, where $t_i$ denotes the $i^{th}$ topic. Thus, $|A|$ and $|t_i|$ represent the vector length of *A* and $t_i$, respectively. Once a topic has been selected, it is then denoted by $t_c$ in the thresholding phase. Thresholding is carried out by first calculating $\text{NewTSim}(t_c, A)$. This component is used to compute the cosine similarity between the article and the potentially new topic, as shown in Equation 3.

$$\text{NewTSim}(t_c, A) = \frac{(0.05 \times |t_c|) \times \left(\text{Mean}(A) - \text{StdDev.}(A)\right) \times \text{Mean}(t_c)}{\left(|A| \times \left(\text{Mean}(A)\right)^2\right) \times \left(|t_c| \times \left(\text{Mean}(t_c)\right)^2\right)} \quad (3)$$

When the value of $\text{NewTSim}(t_c, A)$ has been obtained, the Cosine Similarity is then compared with two thresholds as demonstrated in Equation 4 and 5.

$$CosSim(t_c, A) > 0.1 \wedge CosSim(t_c, A) > NewTSim(t_c, A) \quad (4)$$

$$NumTopics > 10 \wedge CosSim(t_c, A) > (2 \times StdDev(AllTopicSims) + Mean(AllTopicSims)) \quad (5)$$

If the topic satisfies both thresholds, then the classifier does not have to reassign the document with a new topic. Once a suitable topic has been selected, this topic is then assigned to the document. If the topic selected is a new topic, then this topic is going to be stored in the topic database. But if it is a pre-known topic in the database, then the particular topic is updated by adding 1 more document that falls under its class.

During experiment, the procedure of which the annotators are instructed to follow is designed to mimic the classification procedure performed by the classifier. The annotators are first provided with a description of each category and then given the task of classifying a collection of news articles. This concept is similar with the one applied in [24], where basically the annotation process are also designed to follow the targeted system's procedures. The experimental setup of the human annotation is further described in the next section.

## 4.0 EXPERIMENTAL RESULTS

The objective of this experiment is to measure whether the classification has satisfied the end-user needs in terms of news classification, hence the discussion with human annotators. Accordingly, the experiment model is built to achieve this goal. As mentioned in the literature review, the number of publicly available methods for Indonesian classification is very low, especially methods which can categorise news articles into two levels of classification. Thus, the benchmark for this experiment is set from human evaluation. Human evaluation is not only helpful in measuring the classifier's performance, but also in understanding the manual classification for which TC is aimed to replace.

### 4.1 Experimental Setup

The human evaluation test setup is as follows:

1. **Category Classification**

   Environment: Offline.
   Objective: To compare the performance of the combined method with the performance of the manual classification done by human annotators.
   Procedure: The annotators are provided with testing data and briefed with the testing procedure. They are also briefed about the descriptors for each category. They are provided the most convenient time to classify the testing samples. While the annotators are evaluating the data, classification using the algorithm is conducted. Once the annotators are finished with the evaluation, the results are compared.

2. **Topic Identification**

   Environment: Offline.
   Objective: To compare the performance of the topic identification method using the best threshold with the manual topic identification performed by human annotators.
   Procedure: The procedure for this model is the same with the classification procedure, and the annotators are asked to perform topic identification along with category classification.

3. **Human Annotators Discussion**

   Objective: To retrieve and analyse the feedback from human annotators regarding the classification method.
   Procedure: The annotators are asked with questions about the experiment. The answers are recorded and analysed.

In order to measure the classification performance based on human's evaluation, the human annotators are selected to act as the readers of online news. These readers consist of multiple layers of society with average knowledge in the language. Thus, the selected human annotators in this case are a group of well-educated people (university graduates and master degree holders) and are able to perceive and interpret the content of a news

article, but are neither language experts nor well-trained in linguistic study—in order to prioritise content over writing styles. The annotators are also not particularly selected from a press company either because these "common" people are seen as the actual reader of the news. It is assumed that they can easily place themselves as the end-user of the news website without the influence of professional relation to a certain news website.

There are five annotators in this experiment and all are of Indonesian nationality, with various regional origins in Indonesia, with ages ranging between 20 and 35 years old. 45 data have been used from the year 2012. These 45 articles consist of five articles in each category. The testing dataset can be seen in Table 1.

**Table 1** Test dataset for human annotations

| Categories | Number of Articles |
|---|---|
| *Nasional* (National) | 5 |
| *Regional* (Regional) | 5 |
| *Internasional* (International) | 5 |
| *Metropolitan (*Metropolitan) | 5 |
| *Ekonomi* (Economy) | 5 |
| *Olahraga* (Sports) | 5 |
| *Sains dan Teknologi* (Science and Technology) | 5 |
| *Edukasi* (Education) | 5 |
| *Pariwisata* (Tourism) | 5 |
| **Total** | **45** |

In some of the categories, the articles are sparsely distributed to test the classifier on a sparsely distributed set of articles. The testing entries given to the human annotators are articles within the May – June 2013 period. The period of time is selected to test the classifier whether it can accurately detect if an article does not have any match with the study materials.

## 4.2 Results and Discussions

The human annotators have been asked to go through 45 articles and decide the category that best fits each of the articles. To adopt the system's process, the annotators are verbally informed with the common keywords that frequently occur in the articles in the category. The keywords informed are subsets of the keywords selection method. There is no other option apart from the provided categories. This procedure is conducted to simulate the learning mechanism of the system's method.

Another adaptation of the system—where no other category apart from the ones in the primitive list may be assigned to the article—is the non-existent option apart from the set of categories. If an annotator feels that there is no suitable category that can represent the article, they may choose one category closest to their judgment.

The results of this experiment show that the system's performance is generally equal to the results of human evaluation. 80% matches means that there is only one article misclassified by the human annotator and the system. After a discussion with the human annotators and a study on the classification documents, it has been found that there had been one article that was supposedly belong to the Science & Technology class but was classified as Regional due to the ambiguous content. The overall performance of the classifier illustrated in Figure 2 was relatively high with an average result of 89% match with the ground truth.

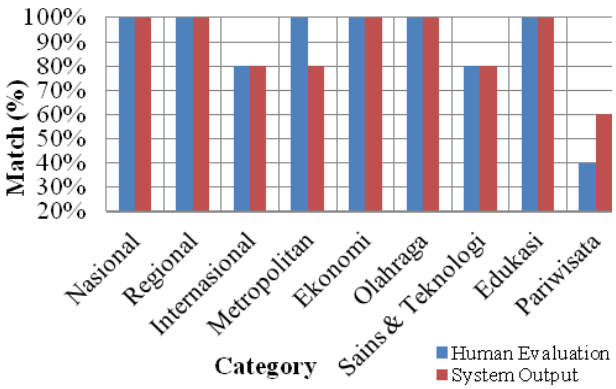



**Figure 2** Category classification compared with human evaluation

The topic identification experiments are done in the same fashion as the category classification. The human annotators are asked to classify the articles into a choice of topics. Like the category classification tests, the topic identification test on human evaluation experiments are set to simulate the algorithm's workflow. The human annotators are asked to assign each article with a topic which they think is the truest representative of the content of the article. However, if they think that there is no topic in the list that matches the article's content, they are provided with the option to state that there is no matching topic.

In every five articles of the set of categories, the number of topics might vary between two to five topics. Five topics indicate that all five articles are from different topics, while two means that a maximum of four articles would have the same topic. One topic range is not used because it is assumed that it cannot help to investigate whether the algorithm can identify sparsely distributed dataset. The number of topics is shown in Table 2.

**Table 2** Topic identification dataset for human evaluation testing

| Categories | Topics |
|---|---|
| *Ekonomi* (Economy) | 3 |
| *Edukasi* (Education) | 5 |
| *Internasional* (International) | 4 |
| *Metropolitan* (Metropolitan) | 5 |
| *Nasional* (National) | 3 |
| *Olahraga* (Sports) | 5 |
| *Regional* (Regional) | 5 |
| *Pariwisata* (Tourism) | 2 |
| *Sains dan Teknologi* (Science and Technology) | 4 |
| **Total** | **36** |

The articles selected for this experiment cover both a previously seen topic and a new topic. The purpose of this is to test the classifier performance on a new, never-been-seen topic, and whether it can detect if an article requires a new topic to be made. The system was proven able to outperform human evaluation with significant difference. In categories such as Regional, Economy, Sports, Science & Technology and Education, the algorithm has outperformed the human annotators by at least 20%.

While the classifier has performed well on a never-been-seen corpus, the human annotators have found difficulties in determining whether the article should be assigned a new topic or not. This is one of the underlying reasons on why the algorithm has its advantages to human annotators. The result of the experiment is shown in Figure 3.

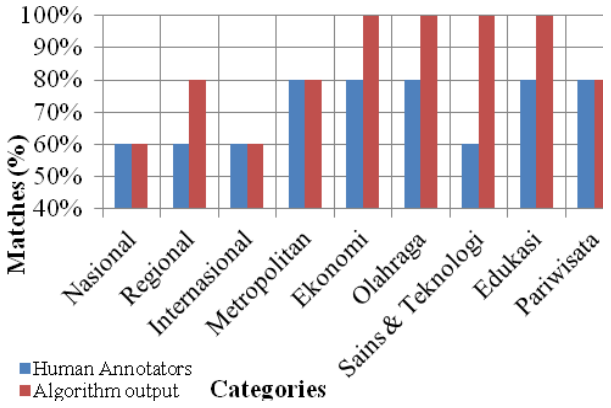



**Figure 3** Topic identification results against human evaluation

The topics in this category have produced higher results as compared to the National and International category. The possible reason for one of the declining topic category like the *Internasional* topic is because it has been frequently confused with the *Pariwisata* topic. This is likely because some of the articles with the new topics are articles about tourism abroad (outside of Indonesia), leading to the decreased results. Nevertheless, the algorithm's performance in topic identification experiments against the human evaluation as benchmark remains satisfying. Figure shows that the proposed method can outperform the human evaluation results by more than 10%, with a total match of 84%.

As part of the experiment, the human annotators' responses are explained and discussed based on the phase and the issues that the annotators felt as a major challenge during the experiment. The objective of the discussion is more towards analyzing the technique that the annotators used in manual classification, and the issues that they discovered during the experiment, and not further comparing the results from those techniques with the system's results. The discussions are grouped into several clusters as follows:

**1. Category Classification**

a. Categories List

In the category classification phase, the human annotators were asked to choose from a list of nine different categories. All of the human annotators agreed that this list of categories is adequate to accommodate the variety of news given during the experiment in specific terms, and news in the real world in general. When asked if there was news that could not be assigned to any of the categories, none of the annotators responded with an affirmative response. It can be concluded from this feedback that the overall categories list is rigid enough to be used for formal news classification.

**2. Topic Identification**

a. Number of Topics

After the articles were categorised, the annotators were asked to choose a topic from the topics list and assign it to the articles. Three out of five annotators agreed that the number of topics were excessive. They stated that the excessive number of topics has become a factor perplexing them in the process of topic identification. Although the ground truth of the testing topics was only 36, but the number of available topics to choose for the annotators were 559. From the system's side, this is not seen as a problem since the method allows the classifier to work within a large database.

b. Similar Topics

The second subject of discussion with the annotators was the number of similar topics. During the identification, all five annotators found that there were groups of topics which look very similar to each other. These topics have also become an element that slows down their topic identification process. One

example is the topic about "*Negara Islam Indonesia*" or NII (Islamic State Indonesia). Many of the articles in this news were published in 2011, when the name of a political figure was involved. When the news broke, KOMPAS published many articles covering this story. The news was divided into several topics such as "*Gerakan NII*" (The movement of NII), "*Sepak Terjang NII*" (The lunge of NII), "*Korban NII*" (Victims of NII), and so forth. During topic identification, the annotators admitted to have found difficulties in differentiating between topics. In total, there are five topics covering NII which include: *Gerakan NII*, *Korban NII*, *Penyebaran NII* (The spreading of NII), *Sepak Terjang NII*, *Cuci otak NII* (NII's brainwash), *Testimoni Korban NII* (NII's victims testimonial). The human evaluation results using the NII example are described in Table 3.

**Table 3** Human evaluation results on nii topic

| Articles | Annotators | | | | | Classification |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| 1 | 0 | 0 | 0 | 1 | 1 | FALSE |
| 2 | 0 | 0 | 0 | 0 | 0 | FALSE |
| 3 | 1 | 0 | 1 | 1 | 0 | TRUE |
| 4 | 0 | 0 | 0 | 0 | 0 | FALSE |
| 5 | 1 | 0 | 1 | 0 | 0 | FALSE |
| 6 | 0 | 0 | 0 | 0 | 0 | FALSE |
| 7 | 1 | 1 | 0 | 0 | 0 | FALSE |
| 8 | 0 | 0 | 0 | 0 | 0 | FALSE |
| 9 | 1 | 0 | 1 | 1 | 1 | TRUE |
| 10 | 1 | 1 | 1 | 1 | 1 | TRUE |
| 11 | 0 | 0 | 0 | 0 | 0 | FALSE |
| 12 | 0 | 0 | 0 | 0 | 0 | FALSE |
| 13 | 0 | 0 | 0 | 0 | 0 | FALSE |
| 14 | 1 | 1 | 0 | 0 | 1 | TRUE |
| 15 | 0 | 0 | 0 | 0 | 0 | FALSE |
| **Sum** | 40% | 20% | 27% | 27% | 33% | 27% |

As seen in Table 3, none of the human annotator could correctly assign the right topic to the article. Most of the annotators had chosen the topic "*Gerakan NII*" over the correct topic from KOMPAS which is "*Sepak Terjang NII*". This behavior is also seen in the rest of the documents with similar topics. In the case of similar topics, only very few annotators could correctly assign the topic for the article.

c. New Topics

One of the aforementioned advantages of the algorithm is the ability to detect a new, never-existed topic. The method of detecting a new topic has also been a subject of discussion with the annotators. Most of the annotators agreed that the technique they used to detect whether a new topic should be assigned is by first setting in mind that the topic exists somewhere inside of the list. They would rather repeat the process of reading the complete list of the topics and only decide if no topic matches the article, than directly decide that a new topic has to be made. This causes a tendency to conform to the existing topic as opposed to state that there is no matching topic in the list. As the articles become trickier, the results of this mindset also reduced the performance quite substantially. The result for this case is described in Table 4.

**Table 4** Annotator's results on new topic identification

| Annotator | Result | Ground Truth |
|---|---|---|
| **1** | *Gerakan NII* | |
| **2** | *Gerakan NII* | |
| **3** | *Penyebaran NII* | *Sepak Terjang NII* |
| **4** | *Terorisme* | |
| **5** | *Gerakan NII* | |

From the 15 news articles of which primitive topics do not exist in the database, the annotators could only classify correctly four of them. These 15 articles were distributed among different categories, and were put in a random order. The best rate of correct topic identification that the annotator could produce was 40%, or six articles out of a total of 15. From this discussion, it is evident why the annotators scored significantly lower than the classifier's performance in the human evaluation test. The new topic discussion also shows that the algorithm can advance in terms of identifying new topics as its procedure enables to immediately detect if no topic in the database is a match to the article by using the threshold.

## 3. Classification Time

The annotators were asked about the duration of time that they needed to finish classifying and identifying an article. All of the annotators agreed that it took approximately 5 – 10 minutes for them to finish an article. For the complete classification, some of the annotators managed to complete them in one day, and the rest in 2 – 3 days. However, the overall classification depends massively as well, on the annotators' schedules and personal matters. Thus, the classification time for the complete list is not counted as relevant. However, the classification time for one article can be considered as a benchmark for the algorithm's time. 5 – 10 minutes is considered as very long, compared to the system's performance which required only 2 – 3 seconds per article. An annotator also stated that an article's length and difficulty was a determining factor on how long the classification for that article can take. This shows that the classifier is able to perform the classification on a timely manner, with high accuracy.

## 5.0 CONCLUSION

This paper addresses the issues of constant updates in Text Classification with the help of human annotators. The algorithm used is Bracewell's algorithm applied on Indonesian news corpus. Instead of utilised as an additional information to the training data, annotations are used as benchmarks in the testing phase to measure the classifier's performance. Human annotators are asked to do the classification in the most similar way that it can simulate the classifier's procedure. The overall results showed that the classifier can outperform the annotation-based classification on both category classification and topic identification. This paper also presents an in-depth study of the human annotators' behaviour towards the collection of data. It can be concluded that Bracewell's algorithm is suitable for Indonesian news classification and it can take over the time-consuming task of manual classification.

## References


[1] F. Sebastiani. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR)*. 34: 1-47.

[2] M. Chang and C. K. Poon. 2009. Using Phrases as Features in Email Classification. *Journal of Systems and Software*. 82: 1036-1045.

[3] S. Kiritchenko and S. Matwin. 2011. Email Classification with Co-training. In *Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research*. 301-312.

[4] X. Wang, *et al.* 2011. Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach. In *Proceedings of the 20th ACM international conference on Information and Knowledge Management*. 1031-1040.

[5] H. Chen and D. Zimbra. 2010. AI and Opinion Mining. *Intelligent Systems, IEEE*. 25: 74-80.

[6] D. B. Bracewell, *et al.* 2009. Category Classification and Topic Discovery of Japanese and English News Articles. *Electronic Notes in Theoretical Computer Science*. 225: 51-65.

[7] F. Rodrigues, *et al.* 2013. Learning from Multiple Annotators: Distinguishing Good from Random Labelers. *Pattern Recognition Letters*. 34: 1428-1436.

[8] A. Yessenalina, *et al.* 2010. Automatically Generating Annotator Rationales to Improve Sentiment Classification. In *Proceedings of the ACL 2010 Conference Short Papers*. 336-341.

[9] A. McCallum and K. Nigam. 1998. A Comparison of Event Models for Naive Bayes Text Classification. In *AAAI-98 Workshop on Learning for Text Categorization*. 41-48.

[10] C. C. Aggarwal and C. Zhai. 2012. A Survey of Text Classification Algorithms. In *Mining Text Data*. ed: Springer, 163-222.

[11] H. M. Noaman, *et al.* 2010. Naive Bayes Classifier based Arabic Document Categorization. In *Informatics and Systems (INFOS), 2010 The 7th International Conference on*. 1-5.

[12] J. MacQueen. 1967. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 14.

[13] K. Wagstaff. *et al.* 2001. Constrained k-means Clustering with Background Knowledge. In *ICML*. 577-584.

[14] M. Steinbach, *et al.* 2000. A Comparison of Document Clustering Techniques. In *KDD Workshop on Text Mining*. 525-526.

[15] J. C. Dunn. 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters.

[16] L. Kaufman and P. J. Rousseeuw. 2009. *Finding Groups In Data: An Introduction to Cluster Analysis*. 344: John Wiley & Sons.

[17] O. Zaidan, *et al.* 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *HLT-NAACL*. 260-267.

[18] O. F. Zaidan and J. Eisner. 2008. Modeling Annotators: A Generative Approach to Learning from Annotator Rationales. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 31-40.

[19] A. Srivastava and M. Sahami. 2010. *Text Mining: Classification, Clustering, and Applications*. CRC Press.

[20] P. K. Bhowmick, *et al.* 2010. Classifying Emotion in News Sentences: When Machine Classification Meets Human Classification. *International Journal on Computer Science and Engineering*. 2: 98-108.

[21] D. B. Bracewell, *et al.* 2005. Multilingual single document keyword extraction for information retrieval. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*. 517-522.

[22] G. Salton, *et al.* 1975. A Vector Space Model for Automatic Indexing. *Commun. ACM*. 18: 613-620.

[23] B. Bigi, *et al.* 2001. A Comparative Study of Topic Identification on Newspaper and E-mail. In *String Processing and Information Retrieval-SPIRE*. Villers-l`es-Nancy.

[24] D. Higgins. 2007. Reliability of Human Annotation of Semantic Roles in Noisy Text. In *Semantic Computing, 2007. ICSC 2007. International Conference on*. 501-508.