Identification of topics in News Articles Using Algorithm of Porter Stemmer Enhancement and Likelihood Classifier

Alvida Mustika Rukmi\textsuperscript{1}, Devi Andriyani\textsuperscript{2}, and Imam Mukhlas\textsuperscript{3}
\textsuperscript{1,2,3}Department of Mathematics, Institut Teknologi Sepuluh November
Campus ITS, Sukolilo, Surabaya, 60111, Indonesia
E-mail: alvidamr@gmail.com

Abstract—Every piece of information contained in a story sometimes has a variety of themes and seems not specific so there is difficulty in digesting information simultaneously. This requires grouping based on the topic relevance of the news. This grouping can make it easier for readers to get the information in accordance with the topic you want to read. Each news group must have different information characteristics so that we need a special algorithm that is able to handle topic discovery and classification using training data on many Indonesian news articles. This research will apply an algorithm of Porter Stemmer Enhancement in the stemming process and Likelihood method for news classification based on categories and identification of topics. Based on the test results using 900 training data and 90 test data, obtained a fairly high accuracy, namely 95.56% for category classification and 97.78% for topic identification.

1. Introduction
Along with the rapid development of technology also increases the spread of information online as well as news or articles that we easily encounter on various sites. The collection of information certainly has a variety of discussion themes so it is not possible that all information presented can be digested simultaneously, but must be grouped based on the relevance of the topic of the news. This grouping can make it easier for readers to choose the most important information according to the topic you want to read.

Information in the news has different characteristics from the collection of other documents, namely the dynamic flow of new documents that may have information that was never present in previous documents. So to classify topics requires a special algorithm that is able to handle the discovery of topics, and classification using training data [1]. The process of identifying news topics will be pre-processed, namely training documents consisting of filtering, stopword removal, stemming and weighting processes.

In [1] research has been conducted to identify topics and categories of English language news using likelihood calculations. Whereas in [2] a similar study was carried out, namely the identification of topics and categories of Indonesian news using the likelihood calculation and the use of the Confix Stripping Stemmer algorithm for stemming. Although the running time needed to identify the topic is quite long, the precision value produced is quite high at 97.26%.
The identification of Indonesian news topics in this paper uses the Porter Stemmer enhancement algorithm modified by Putu BagusSusstra Wiguna and Bimo Sunarfr Hüntono [5]. This method is applied to the stemming process of identifying news topics and compared to previous research. The accuracy of identifying news topics can facilitate the selection of news information based on topics for users.

2. Basic Theory

2.1. Corpus

Corpus is a collection of structured texts. More specifically, it is the news text that is downloaded from online news sites that is stored in a particular text file format and has a category linkage between its corpus. Corpus that will be used in this paper is corpus with the existence of .news to make it easier for the program to recognize the corpus when it is being processed [2]. If the corpus is still stored in the .txt extension, then there is a high possibility that in a folder there are other unrelated files that use the same extension, so that it will be processed and disrupt the application process.

The corpus as a whole is the result of downloading news on the site by removing the html or php attributes on the page that are structured following the format described in Figure 3. The training data corpus does not write the categories that have been set by the compass website because there are several categories whose names are changed to more general names so that they are different from the category names written on www.kompas.com. All news is stored in folders according to category names. The significant difference between training data and test data is the topic attribute. The test data which also had corpus formation did not have topic attributes because it was assumed that those attributes would be the result of identification in the process of identifying topics carried out by the program.

![Figure 1. Flow chart of stemmer porter enhancement](image)

2.2. Porter Stemmer Enhancement
The words forming process from basic words can’t be always completed by morphology 1st-level (word formation by adding affixes to basic words). Examples morphological 1st-level are "mem" + "baca" become "membaca", "men" + "cari" become "mencari". The addition of affixes to the basic words to form new words by changing the phonemes of the basic words cannot be completed with morphology 1st-level. It need morphology 2nd-level, examples : words "memutar" derived from the basic word "putar" which gets the affix "men-".

In general, the process of stemming is divided into 5 parts, namely: eliminating the first prefix ("meng", "peng-", "mem", "pem-", "meny-", "peny-", "men-", "pen- "and others), removes the second prefix ("BER ", "PER ", "TER ", " SE ", " PEL ", etc.), removes particles (" -AH ", " -Lah ", " -tah ", " -tah "), eliminates the personal pronouns (" - my ", " your ", " his "), removes the suffix ("-kan", "-an", "-I", "-ism", "-ization", "-onal"). The general stemming process carried out in this study can be seen in Figure 4. [3]

In this research, use word database consisting of 7 tables as the dictionary of exclusion words for each stemming process. The tables used are dsr_pilik tables, dsr_partikel tables, dsr_prefiks1 tables, dsr_prefiks1_sufiks1 tables, dsr_prefiks2 tables, dsr_prefiks21 tables, dsr_sufiks tables. Example dsr_prefiks1 table is used to store basic words that have the first prefix phoneme. This table is useful in the process, so that the first phoneme of the word is not removed because it is part of the basic word. The second table is the dsr_prefiks1_sufiks1 table. This table is used to store words that must be processed 2 levels of morphology for basic words beginning with the letters "k" and "p". And so on for other tables.

2.3. Vector Space Model

Vector Space Model (VSM) is a mathematical model used in the information retrieval system to determine relevant between document and information. This model will calculate the degree of similarity among document stored in the system with the query given by the user. This model was first introduced by Salton (1989) [7]. Vector space models are one of the most commonly used approaches to represent digital text models. Each document \( d_i \) will be represented as a vector [4].

\[
d_j = (w_{1j}, w_{2j}, ..., w_{lj})
\]

where \( w_{ij} \) is the term weight in the relevant document \( j \).

2.4. TF-IDF Method

Baeza-Yates and Ribeiro-Neto (1999), states that weighting (tf · idf) consists of two factors:

2.4.1 tf (term frequency). Frequency of occurrence of the term \( k_i \) in a document \( d_j \), is compared to the frequency of the term \( k_i \) which often appears in that document. We can formula obtained:

\[
tf_{ij} = \frac{freq_{ij}}{\max(freq_{ij})}
\]

2.4.2. idf (inverse document frequency). Frequency of occurance of term \( k_i \) in all documents is idf. The use of the IDF factor based on the terms that appear on each document does not provide a specific characteristic to determine the relevant document from the irrelevant one. If the total number of documents in the system is stated as \( N \) value and the number of documents that have the term \( k_i \) is stated as \( df_i \), then the idf value can be stated by:

\[
idf_i = \log_2 \left( \frac{N}{df_i} \right)
\]

The weight of each term can be represented by the inverse frequency of the document (TF-IDF):

\[
w_{ij} = tf_{ij} \log_2 \left( \frac{N}{df_i} \right)
\]

where \( w_{ij} \) is the weight of the term-i on the document-j. \( idf \) is the frequency of the term-i on document-j. \( N \) is the number of documents processed and \( idf \) is the number of documents that contain term-i [5]

2.5. Likelihood

The likelihood calculation for a category is explained in Equation(5).
\[Likelihood(c_i|A=\{k_1,k_2,\ldots,k_n\}) = -\sum_{j=1}^{n} P(k_i|c_j) \log P(k_i|c_j)\]  
\[\text{Threshold} = \frac{\sum_{i=1}^{|A|} l_i}{|A|} + \sqrt{\frac{\sum_{i=1}^{|A|} (l_i - \frac{\sum_{i=1}^{|A|} l_i}{|A|})^2}{|A|}}\]

2.6. Algorithm of Topic Identification

Topic classification algorithm can be divided into two processes, namely classification and dynamic threshold. This algorithm calculates the similarity between previously published topic keywords and test article keywords. After that, the value that has the highest similarity is determined for the article as a conditionally determined topic. [1]

To compare keyword vectors with vector topics, they are transformed into the same vector-space. Example:

![Figure 2. Vector Transformation](image)

After that the similarity value is calculated using the following Equation:

\[CosSim(t_c, A) = \frac{t_c A}{|t_c||A|}\]  

where \(t_c\) is a vector of topic-i, A is the test article A, \(|t_c|\) and \(|A|\) each is the length of vector of topic-i and the length of the article vector A.

Topics that have the greatest similarity will be tested using dynamic threshold values. This threshold value compares the initial topic value specified with the new topic value that has been determined.

\[\text{NewTSim}(t_c, A) = \frac{(0.05 x |t_c| x \text{mean}(A) - \text{StdDev}(A)) x \text{mean}(t_c)}{(|A| x \text{mean}(A))^2 x (|t_c| x \text{mean}(t_c))^2)}\]

form NewTSim and \(t_c\) is the initial topic has been determined which get the max CosSim calculation result, Mean (A) is the average of document A, StdDev (A) is the standard deviation of the document vector A, and the mean (\(t_c\)) is the average of the initial topic that has been determined.

The next step is dynamic thresholding, which compares the value of NewStim with the initial topic value that has been determined as follows (9):

- \(\text{CosSim}(t_c, A) > 0.1\) \& \(\text{CosSim}(t_c, A) > \text{NewTSim}(t_c, A)\)
- \(\text{NumTopics} > 10\) \& \(\text{CosSim}(t_c, A) > (2 x \text{StdDev(AllTopicsSims)} + \text{Mean(AllTopicsSims)})\)

With CosSim (\(t_c, A\)) is the biggest Cosine Similarity calculation result obtained from Equation (7) and is assumed to be the initial topic specified. NumTopic is the total number of topics that have been known beforehand, StdDev (AllTopicsSims) and Mean (AllTopicsSims) respectively are standard deviations and averages of all similarity topics. [1]
2.7. Trial Evaluation

The evaluation of trials often uses the formula precision, recall, F-Measure and accuracy. The understanding of some of the methods above is:

- **Precision** is the level of accuracy between the information requested by the user and the answers given by the system which are formulated as follows:
  \[
  \text{Precision}(P) = \frac{TP}{TP + FP}
  \]

- **Recall** is the level of success of the system in finding back an information that is formulated as follows:
  \[
  \text{Recall}(R) = \frac{TP}{TP + FN}
  \]

- **F-Measure** is the harmonic mean of precision and recall which is formulated as follows:
  \[
  \text{FMeasure}(F) = \frac{2 * P * R}{P + R}
  \]

- **Accuracy** is defined as the level of closeness between the predicted value and the actual value which is formulated as follows:
  \[
  \text{Accuracy}(A) = \frac{TP + TN}{TP + FP + FN + TN}
  \]

Table 1. Precision, Recall, F-measure, Accuracy between Actual and Prediction Value

| Actual Value | Prediction Value |
|--------------|------------------|
| TRUE         | FALSE            |
| TP(True Positive) | FP(False Positive) |
| Correct result                        | Unexpected result         |
| FALSE        | TRUE             |
| FN(False Negative)                   | TN(True Negative)         |
| Missing result                        | Correct absence of result |

3. Result and Discussion

3.1. Collection of news documents and Corpus

This application input data in the form of corpus English news documents with the extension. News. The .news extension is used to facilitate the retrieval of files both during the preproces or application learning process. Corpus has the date format, source code, title and contents of the news document. Corpus will be taken through the site www.kompas.com. Corpus used in this paper uses data that has been used in [2] namely 932 training data and 10 testing data for each category.

```
Tanggal_berita   <Day, DD Month YYYY>
Topik_berita    <News Topic>
ID_Sumber       <ID News Source>
Judul_berita    <News Title>
Isi_berita      <News Content>
```

Figure 3. Corpus Format
3.2. Document Text Training

In this stage, Preprocess will be carried out, which is the design phase of functions that can be applied in the application. Among them is the training document must be represented in vector form which includes Case folding, Filtering, Stoplist Removal, Stemming, and Weighting. [6]

- Case Folding: All letters in each word in a document are converted to lowercase letters
- Filtering: eliminating punctuation
- Stoplist Removal: removal of characters that have a high frequency, because it is considered not an important word. These words include: prepositions, conjunctions, and others. The words included in the stoplist are called stopwords and have been stored in a database.
- Application of Porter's Improvement Algorithm in Stemming Process
- Weighting: weighting of each term that has been stemmed through the TF-IDF method.
3.3 . **Classification of category**

![Classification of category process](image)

**Figure 5.** Classification of category process

3.4. **Topic Identification**

Process in topic identification consist of :
- CosSim Counting
- CosSim Max Selection
- Treshold Counting
- Topic Selection

4. **Trial and Evaluation of Classification and Identification Results**

At this stage will be tested on the program. Namely testing Corpus testing data that has been stored previously with a total of 10 data for each category. The results of the classification of each category will be calculated the value of precision, recall, f-measure, accuracy. As for the identification of topics using 90 data testing and accuracy values will be calculated based on the number of identified correctly divided by the total test data.

Data in the form of corpus Indonesian online news obtained from www.kompas.com. News is downloaded based on predetermined categories. Primitive categories in trials are useful for evaluating the results of classifications. Between a category with other categories have the same number of test documents. Specifications for the number of documents for each category can be seen in Table 2.
### Table 2. Specifications for the Number of Documents Each Category

| Category          | Sum of Document |
|-------------------|-----------------|
| Nasional          | 10              |
| Regional          | 10              |
| Internasional     | 10              |
| Metropolitan      | 10              |
| Bisnis dan Ekonomi| 10              |
| Olahraga          | 10              |
| Sains dan Teknologi| 10             |
| Edukasi           | 10              |
| Pariwisata        | 10              |
| **Total**         | **90**          |

### Table 3. Accuracy of category classification with 10 keywords

| No  | Category          | Precision | Recall   | Measure | Accuracy   |
|-----|-------------------|-----------|----------|---------|------------|
| 1   | Internasional     | 1.00000000 | 0.80000000 | 0.88888889 | 1.00000000 |
| 2   | Nasional          | 0.88888889 | 0.88888889 | 0.88888889 | 1.00000000 |
| 3   | Regional          | 0.77777778 | 0.77777778 | 0.77777778 | 0.70000000 |
| 4   | Metropolitan      | 0.87500000 | 0.77777778 | 0.82352941 | 0.80000000 |
| 5   | Bisnis Ekonomi    | 0.88888889 | 0.88888889 | 0.88888889 | 1.00000000 |
| 6   | Olahraga          | 0.88888889 | 0.88888889 | 0.88888889 | 0.90000000 |
| 7   | Pariwisata        | 0.88888889 | 0.88888889 | 0.88888889 | 1.00000000 |
| 8   | Sains Teknologi   | 0.88888889 | 0.88888889 | 0.88888889 | 1.00000000 |
| 9   | Edukasi           | 0.88888889 | 0.88888889 | 0.88888889 | 0.90000000 |

**Accuracy Mean** 0.922222222

### Table 4. Accuracy of category classification with 25 keywords

| No  | Category          | Precision | Recall   | Measure | Accuracy   |
|-----|-------------------|-----------|----------|---------|------------|
| 1   | Internasional     | 0.88888889 | 0.80000000 | 0.842105263 | 0.90000000 |
| 2   | Nasional          | 0.88888889 | 0.88888889 | 0.88888889 | 0.90000000 |
| 3   | Regional          | 0.88888889 | 0.88888889 | 0.88888889 | 0.90000000 |
| 4   | Metropolitan      | 1.00000000 | 0.88888889 | 0.941176471 | 1.00000000 |
| 5   | Bisnis Ekonomi    | 1.00000000 | 0.88888889 | 0.941176471 | 1.00000000 |
| 6   | Olahraga          | 0.88888889 | 0.88888889 | 0.88888889 | 0.90000000 |
| 7   | Pariwisata        | 1.00000000 | 0.88888889 | 0.941176471 | 1.00000000 |
| 8   | Sains Teknologi   | 1.00000000 | 0.88888889 | 0.941176471 | 1.00000000 |
| 9   | Edukasi           | 1.00000000 | 0.88888889 | 0.941176471 | 1.00000000 |

**Accuracy Mean** 0.955555556

### Table 5. Accuracy of category classification with 30 keywords

| No  | Category          | Precision | Recall   | Measure | Accuracy   |
|-----|-------------------|-----------|----------|---------|------------|
| 1   | Internasional     | 1.00000000 | 0.80000000 | 0.88888889 | 1.00000000 |
| 2   | Nasional          | 0.88888889 | 0.88888889 | 0.88888889 | 0.90000000 |
| 3   | Regional          | 0.88888889 | 0.88888889 | 0.88888889 | 0.90000000 |
| 4   | Metropolitan      | 1.00000000 | 0.88888889 | 0.941176471 | 1.00000000 |
| 5   | Bisnis Ekonomi    | 1.00000000 | 0.88888889 | 0.941176471 | 1.00000000 |
| 6   | Olahraga          | 0.88888889 | 0.88888889 | 0.88888889 | 0.90000000 |
| 7   | Pariwisata        | 1.00000000 | 0.88888889 | 0.941176471 | 1.00000000 |
| 8   | Sains Teknologi   | 1.00000000 | 0.88888889 | 0.941176471 | 1.00000000 |
| 9   | Edukasi           | 1.00000000 | 0.88888889 | 0.941176471 | 1.00000000 |

**Accuracy Mean** 0.955555556
Table 6. Accuracy of category classification with 35 keywords

| No | Category      | Precision | Recall   | Measure            | Accuracy            |
|----|---------------|-----------|----------|--------------------|---------------------|
| 1  | Internasional | 1.000000000 | 0.800000000 | 0.888888889 | 1.000000000 |
| 2  | Nasional      | 0.888888889 | 0.888888889 | 0.888888889 | 0.900000000 |
| 3  | Regional      | 1.000000000 | 0.888888889 | 0.941176471 | 1.000000000 |
| 4  | Metropolitan  | 0.888888889 | 0.888888889 | 0.888888889 | 0.900000000 |
| 5  | Bisnis Ekonomi| 1.000000000 | 0.888888889 | 0.941176471 | 1.000000000 |
| 6  | Olahraga      | 0.888888889 | 0.888888889 | 0.888888889 | 0.900000000 |
| 7  | Pariwisata    | 1.000000000 | 0.888888889 | 0.941176471 | 1.000000000 |
| 8  | Sains Teknologi| 0.888888889 | 0.888888889 | 0.888888889 | 0.900000000 |
| 9  | Edukasi       | 1.000000000 | 0.888888889 | 0.941176471 | 1.000000000 |

Accuracy Mean: 0.955555556

Table 7. Accuracy of category classification with 40 keywords

| No | Category      | Precision | Recall   | Measure            | Accuracy            |
|----|---------------|-----------|----------|--------------------|---------------------|
| 1  | Internasional | 0.888888889 | 0.800000000 | 0.84205263 | 0.900000000 |
| 2  | Nasional      | 0.888888889 | 0.888888889 | 0.888888889 | 0.900000000 |
| 3  | Regional      | 0.888888889 | 0.888888889 | 0.888888889 | 0.900000000 |
| 4  | Metropolitan  | 1.000000000 | 0.888888889 | 0.941176471 | 1.000000000 |
| 5  | Bisnis Ekonomi| 1.000000000 | 0.888888889 | 0.941176471 | 1.000000000 |
| 6  | Olahraga      | 0.888888889 | 0.888888889 | 0.888888889 | 0.900000000 |
| 7  | Pariwisata    | 1.000000000 | 0.888888889 | 0.941176471 | 1.000000000 |
| 8  | Sains Teknologi| 0.888888889 | 0.888888889 | 0.941176471 | 1.000000000 |
| 9  | Edukasi       | 1.000000000 | 0.888888889 | 0.941176471 | 1.000000000 |

Accuracy Mean: 0.955555556

Table 3-7 show a Precision, Recall, F-Measure, and Accuracy calculation for each category. The calculation uses the selection of the number of keywords from 10, 35, 40.

Table 8. Accuracy of Category Classification

| No | Category      | Accuracy 5 keywords | Accuracy 10 keywords | Accuracy 15 keywords | Accuracy 20 keywords | Accuracy 25 keywords |
|----|---------------|----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| 1  | Internasional | 100.00%              | 100.00%               | 90.00%                | 80.00%                | 100.00%            |
| 2  | Nasional      | 90.00%               | 100.00%               | 100.00%               | 100.00%               | 100.00%            |
| 3  | Regional      | 80.00%               | 70.00%                | 90.00%                | 90.00%                | 90.00%             |
| 4  | Metropolitan  | 80.00%               | 80.00%                | 80.00%                | 100.00%               | 100.00%            |
| 5  | Bisnis Ekonomi| 90.00%               | 100.00%               | 100.00%               | 100.00%               | 100.00%            |
| 6  | Olahraga      | 90.00%               | 90.00%                | 90.00%                | 90.00%                | 90.00%             |
| 7  | Pariwisata    | 100.00%              | 100.00%               | 100.00%               | 100.00%               | 100.00%            |
| 8  | Sains Teknologi| 100.00%              | 100.00%               | 100.00%               | 100.00%               | 100.00%            |
| 9  | Edukasi       | 90.00%               | 90.00%                | 90.00%                | 90.00%                | 80.00%             |

Accuracy Mean: 0.955555556

The calculation result data in Table 8 is a summary of the Accuracy calculation for each category by selecting the number of keywords from 5-25. Actually these calculations use the selection of the number of keywords of 5, 10, 15, 20, 25, 30, 35, 40. Obtained the maximum value is to use 25 keywords. Because the number of 30-40 keywords produces an average accuracy that is equal to 25.
Figure 6. Accuracy Mean of Category Classification

And, accuracy mean for topic identification, shown below:

Figure 7. Accuracy Mean of Topic Identification

Accuracy value for topic identification is 0.9778. The maximum accuracy value of topic identification resulted when selection keywords is 20.

If we look at [2], the resulting accuracy values for category classification and topic identification are 93.84% and 97.26%, respectively. Whereas in this paper, a higher accuracy value is obtained, namely 95.56% for category classification and 97.78% for topic identification.

In addition to the results above, in Figure 6 we can see that an increase in the average accuracy of 2.22% in the number of keywords by 10, this is because there is a change in the number of documents that are identified correctly in 4 categories, namely a 10% decrease in the category of “Regional” and a 10% increase on category of “Nasional”, “Bisnis Ekonomi” and “Pariwisata”. While the number of keywords 10 to the number of keywords 15 to 25, an increase in the average accuracy is constant that is equal to 1.11%, it is because there is a change in the number of documents that are identified correctly in 2 categories, namely a decrease of 10% accuracy in the category of “Internasional” and a 10% increase in accuracy in the category of “Metropolitan” or a 20% increase in the category of “Internasional “but a 10% decrease in the category of “Edukasi”. Besides that, in Figure 7 there was a constant increase of 1.11% from the keyword selection of 5-20, while from the selection of the number of keywords 20-40
there was no increase, but the maximum results obtained for the selection of the number of keywords by 20.

5. Conclusion

- The program has been completed using the Porter Stemmer and Likelihood Improvement Algorithm and was tested capable of carrying out the process of category classification and topic identification in Indonesian news articles
- Based on the trial results, the category classification process gets optimal results when using the number of keywords as many as 25, while for the identification of topics obtained the maximum results with the number of keywords as much as 20.
- The accuracy value for category classification is obtained at 95.56%, while for topic identification is 97.78%. Both values appear better than the accuracy values generated in previous studies.

References

[1] Bracewell D, Jiajun Yan, Fuji Ren dan Shingo Kuroiwa. 2009. “Category Classification and Topic Discovery of Japanese and English News Article,” Electronic Notes in Theoretical Computer Science 225(2009) 51-65.
[2] Fuddoly, Aini Rachmania Kusumaagama, Agus Zainal Arifin. 2011. “Klasifikasi Kategori dan Identifikasi Topik pada Artikel Berita Bahasa Indonesia,” ITS.Surabaya
[3] Karaa, Wahiba Ben Abdessalem, “A New Stemmer to Improve Information Retrieval,” International Journal of Network Security & Its Applications (IJNSA), Vol.5, No.4, July 2013
[4] DR. E. Garcia, 2006. The Classic Vector Space Model, <URL: http://www.miisilita.com/term-vector/term-vector-3.html>
[5] Wiguna, Putu Bagus Susastra, Bimo Sunarfri Hantono. "Peningkatan Algoritma Porter Stemmer Bahasa Indonesia berdasarkan Metode Morfologi dengan Mengaplikasikan 2 Tingkat Morfologi dan Aturan Kombinasi Awalan dan Akhiran," JNTETI, Vol.2, No.2, 2013
[6] Agusta Ledy, "Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia", Konferensi Nasional Sistem dan Informatika 2009; Bali, November 14, 2009
[7] Nadirman Firnas, 2006. Sistem Temu-Kembali Informasi Dengan Metode Vector Space Model Pada Pencarian File Dokumen Berbasis Teks, <URL: http://kabulkurniawan.web.ugm.ac.id/wp-content/uploads/SKRIPTI.pdf>