Classification of News on “Radar” Tarakan Online Using K-Nearest Neighbor Method with N-Gram Features

E D Bintar1*, Gunawan2 and A Indriani1

1 STMIK PPKIA Tarakanita Rahmawati, Tarakan, Kalimantan Utara, Indonesia
2 Sekolah Tinggi Teknik Surabaya, Surabaya

*Email: evidianti@ppkia.ac.id

Abstract. Classification or text categorization is one of the most common themes in analysing complex data. The classification in this research aims to define the class of an object that has not been known to the class. The classification process will be performed in the process of learning and testing against a known class dataset object. As an online news information, Radar Tarakan classifies news to ease readers to find the desired news and to ease the work of the administrators to do the automatic classification. The website of Radar Tarakan daily news (www.radartarakan.co.id) publishes approximately 480 local news everyday. Local news is the news in the surrounding areas of Tarakan in North Borneo. The website contains the classification of eight local news including the dimensions of the area surrounding the town of Tarakan. This research uses 720 news to build the system of news classification automatically.

A text classification method used in this research is the K-Nearest Neighbor (KNN) with N-grams. The accuracy of matching classes generated by the K-Nearest Neighbor with N-Gram is 85.65% (out of 180 test data) and that of the sub class is 70.35%. The classification time is around 2.7 second per news.

1. Introduction

The reliability of Internet technologies was instrumental in fulfilling the information needs. The people currently getting smarter to choose online media to fulfill information needs every time, because it can be accessed anytime and anywhere. Information/news can be seen through the website of the newspaper online. The daily newspaper Radar Tarakan have a Website (www.radartarakan.co.id).

The website contains of news that have a several classifications, such as the economic news headlines, entertainment, sports and news regarding areas around Tarakan. Before the news was publish on its website, the first such news is classified in accordance with its classification. The process of the news is still done with the selection process of classification in the form of news, so the chances of selection news category error may occur. From the above condition that occurs, then it takes a discipline of text mining to overcome it.

The algorithms that are used in this classification is the K-Nearest Neighbor. Steps to be taken in the research starting from later entering information in the form of news from the website (www.radartarakan.co.id) and then do the preprocessing [1] with the groups of news, after that the preprocessing performed then an N-gram document of the next news feature n-gram into training and materials will eventually do the calculation with K-Nearest Neighbor method for the determination of the proper classification with the news that will be published on the website www.radartarakan.co.id. The end result of this research is the classification of local news and news that has the exact number of which are classified.
2. K-Nearest Neighbor

K-Nearest Neighbor algorithm (K-NN) classify the document test based on k nearest Neighbors [2]. Examples of training can be considered as vectors in a multidimensional feature space. This space is partitioned into a region with locations and labels of the training samples. Usually Euclidean distance or cosine similarity is used. During this stage, the data classification training (classes that need to be identified) is also represented as a vector in space features. Distance or similarity of test vectors for all vector training calculated and k nearest training samples are selected. There are several ways to classify a test vector for certain classes. Classic K-NN algorithm determines the class with a majority of the electors of the K-Nearest Neighbor. This research proposes to use a simple algorithm KNN without weighting features for text categorization. This paper proposes to use the method of selection of the features so finding the relevant features in the task of learning interactions (based on interdependence). KNN algorithm can be reduce a lot of number of features that was not be uses, so it is important for the size of a very large document such as the world wide web [3]. The ease of this algorithm makes more efficient computation time but can also be easily used by users instead of the expert. Conceptually, the example of document x called instance that was represented as a vector with a length of | F | so the size of the vocabulary are symbolised as follows:

\[ < w_1, w_2(x)(x), w_3(x), ...; w_F(x) > \]

Where the \( W_j(x) \) is the weight of the term j. Weights calculated from the obtained frequencies. KNN algorithm, distance was uses for the basis weighting for the contribution of each neighbor k in the process of determining the class. Next defined confidence document d to class c as:

\[
\text{Confidence (c, d)} = \frac{\sum_{k \in k'} \text{Sim}(k_i',d)}{\sum_{k \in k} \text{Sim}(k_i,d)}
\]

Where the sim is the value returned by the functions in common used to compare with its Neighbors, that is, for each neighbor in the set of neighbors K (size k) belongs to a certain class c, then summed up the similarities to and document d divided by the sum of all the similarities of neighbors k relating to the document d. To compare the document d with instance i for example, selected functions very simple CosSim using term frequency weighting approach:

\[
\text{CosSim (i, d)} = \frac{c}{\sqrt{A*B}}
\]

Where C is the number of terms is the same between the document d with instance i, A is the number of term frequent on instance i (test data) and B is the number of frequent terms in document d (training data). So the highest score got the ranking for the determination of class.

3. N-Gram

An N-gram is a sequence of sub item N that already given from sequence. N-gram can be understood as long as window that moves as much as N of the text, the contents of this window is N-grams. N-gram level can be either character or word level, level statement. N-gram method is very widely used in natural language processing statistics. It depends on the language and works well in the case of the text that has the noise [4]. N-gram can be described in two differences. The first is to take a token by each character in a sentence. The second is from every word. For example, it is assumed that there is a sentence “A special utility is available to import images and insert them us links in the questions,” then it can be extracted two different logic for N-Grans as follows:

- Logic first:
  - Trigram : \{asp, spe, pec, eci, cia, ial,…\}
  - 4gram : \{aspe, spec, peci, ecia, cial,ialu, ….\}
  - 5gram : \{aspec, speci, pecia, ecial, cialu, ialut, …\}
• The second Logic
  - Trigram: \{ A special utility, special utility is, \ldots \}
  - 4gram: \{ A special utility is, special utility is available, \ldots \}
  - 5gram: \{ \{ A special utility, special utility is available, \ldots \}

In Figure 1 system architecture, the main process is described in detail in research on radar online local news classification method using the K-Nearest Neighbor (K-NN) with N-grams, system architecture consists of 3 stages, namely the process of data acquisition, data preprocessing and stage classification with detail calculation KNN with N-grams.

![System Architecture Diagram]

Figure 1. System architecture

In the early stages of the classification of local news radar Tarakan, data acquisition is performed by using a crawler to get documents from the news website radar Tarakan, after it's done pre-processing the data consists of stage tokenizing, Filtering/stop word removal and stemming then conducted the process of division of training data and test data for N-gram feature formed, further from the N-gram feature done by the method of classification of K-Nearest Neighbor. Training data used to get classified documents while the test data used to test in the determination of the classification with the K-Nearest Neighbor method and achieved the expected accuracy.
Before doing the experiment, prepared 720 local news documents comprising the class dimension of the Tarakan to Kaltara with all the sub class of economy, politics, sports, education, development, crime and sub classes etc. In Table 1 and 2 can be seen the level of accuracy of the classification of local news with Unigram and N-Gram

| Data Training | Data Testing | Accuracy of classes | Accuration of sub classes |
|---------------|--------------|---------------------|---------------------------|
| 540           | 180          | 86.27%              | 64.71%                    |
| 264           | 74           | 84.93%              | 64.38%                    |
| 168           | 178          | 83.89%              | 65.10%                    |
| 117           | 54           | 77.78%              | 72.22%                    |
| Average       |              | 83.22%              | 67.34%                    |

| Data Training | Data Testing | Accuracy of classes | Accuration of sub classes |
|---------------|--------------|---------------------|---------------------------|
| 540           | 180          | 88.24%              | 70.63%                    |
| 264           | 74           | 86.30%              | 71.49%                    |
| 168           | 178          | 86.58%              | 68.44%                    |
| 117           | 54           | 81.48%              | 72.22%                    |
| Average       |              | 85.65%              | 70.35%                    |

After a trial of all features with a wide variety of comparative testing and the training dataset a different conclusion that the classification of local news with N-Gram increase accuracy of 2.43% and 3.01% for the class to the accuracy of the sub class with an average processing time of classification features of unigram 1 304.36 seconds, while the time it takes to process a classification feature N-Grams on average by 2, 371.58 seconds. Analysis of the writer can convey is can the occurrence classification accuracy trials have been conducted that training composition data distribution comparisons that balanced against a trainer all data classes and sub classes join in influencing whether or not high accuracy of classification of local news Radar online Tarakan.

In previous research, namely research that presents the results of classification of text documents of Arabic-speaking produces the value 87.62% precision and recall 63.41% experienced an increased precision of 10% of the unigram feature use. On research the author did an increase in accuracy precision value of 11%. The author can convey analysis is the use of the term N-Gram can provide increased precision and accuracy values because the system gets a choice selection of features against the dataset news, further distribution of the amount of data a balanced composition of the training classes and sub classes are very helpful to the process of classification has high accuracy.

The trial has been conducted on the decline the value of accuracy and precision and recall the sub class. Note on process tests the value of accuracy class can reach 84.84% while the value of accuracy class sub of 71,87%. Analysis of the author’s response to the significant difference between the accuracy of the classes and sub classes is due at the sub class there is a sub class of others. Sub class of others that many hold datasets with different keywords.

4. Conclusions

From the discussion that already studied in this experiment, the conclusion can be drawn as follows:

1. Using text classification methods, namely the method of K-Nearest Neighbor (KNN) with N-Gram, in terms of the suitability of different types of classes and sub classes that are generated by the KNN against the determination of the types of classes that are conducted by the information (IT) Radar Tarakan belongs either. Percentage the accuracy of the match type of class against class 180 test data of 83.22% and sub class of 67.34%.
2. Features of the N-Gram has accuracy better than/bag unigram features of words/single term, with increased accuracy of ± 3%. Test data for as many as 180 features unigram/single term class type of match accuracy 85.65% and sub class of 70.35%.

3. KNN Method has high accuracy if the existing system has many categories (classes), because in the process of testing the results of prediction classes that do not fit with the manual class average is a sub class of others. The author's analysis is because a sub class of others hold many data sets that are diverse, making it difficult to find a suitable pattern for test data (testing).

5. Acknowledgment
   This research was supported by STMIK PPKIA Tarakanita Rahmawati Tarakan in North Borneo, Radar Tarakan newspaper for our research object and Sekolah Tinggi Teknik Surabaya (STTS). We are thankful to our colleagues who provided expertise that greatly assisted the research.

6. References
   [1] Byeongkyu Ko, Chang Choi, Pankoo Kim, Dongjin Choi, Junho Choi, 2012, Document Classification Through Building Specified N-Gram, Dept. of computer engineering Chosun University Gwangju, South Korea.
   [2] Jiawei Han and Micheline Kamber, Data Mining : Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006.
   [3] Raymond J. Mooney. 2006, CS, Machine Learning Text Categorization, University Of Texas at Austin.
   [4] Riyad Al-Shalabi dan Rasha Obeidat, 2008, Improving KNN Arabic Text Classification with N-Grams Based Document Indexing, Jordan University For Science and Technology.