Text Classification by Enhancing Weights of Terms based on their Positional Appearances

Anagha R Kulkarni
CCOEW,
Pune, India

Vrinda Tokekar
IET, DAVV,
Indore, India.

Parag Kulkarni
EKlat Research,
Pune, India

ABSTRACT
Huge store of hidden information in text documents is available. Extracting accurate, useful information from this store is very important. Multinomial Naïve Bayes classification algorithm is effective in processing text and extracting accurate information.

A new approach of assigning weights to terms based on their positional appearance is proposed. The effectiveness of this approach is demonstrated for two standard text datasets Reuters-21578 and 20-newsgroups. This proposed approach improves average F-measure by 1.0% for Reuters-21578 and by 2% for 20-newsgroups at least.

General Terms
Information Processing, Machine Learning.

Keywords
Term Weighting, Document Classification, Multinomial Naïve Bayes Classification Algorithm.

1. INTRODUCTION
Nowadays, most of the information in all types of organizations is stored in digital form, mostly in text format. To gain useful information from this huge amount of hidden information, one must process this information intelligently. If richer information is processed, richer knowledge is gained.

A document contains multiple terms. Each term does not contribute equally in defining the meaning of the document. Some terms contribute more whereas some contribute less. It was observed that the terms which are relevant to the class of document occur at the top or in the initial part of the document. They appear again in the later parts of the document. A new term weighting scheme is introduced by assigning more weight to the terms which lie in the initial part of the document.

Naïve Bayes (NB) classification algorithm is a popular technique of classification [1]. It is simple, effective and accurate in processing text documents. Multinomial Naïve Bayes (MNB) classification algorithm, a variant of NB, considers the frequency of occurrence of terms for classification [2, 3]. This paper uses MNB for classification of documents.

Many schemes are available for term weighting. Term frequency (TF) and TF with inverse document frequency (TF-IDF) are well known. Recent work on term weighting replaces IDF by term relevance ratio [4]. It uses class probability estimations on positive and negative classes. M. Mendoza et al introduce a new term weighting scheme [5]. Their method is based on BM25. They consider TF and IDF for modification. This paper focuses on TF.

The paper is organized as follows. Section 2 proposes a new method of classification of documents. Relevant experimental setup and results are discussed in section 3. Conclusion is presented in section 4.

2. PROPOSED METHOD FOR CLASSIFICATION OF TEXT DOCUMENTS
A text document has terms. Some terms are very common and some are non common or useful terms. The documents have to be pre-processed to remove common terms, also called stop words, (commonly occurring words like with, for, the, etc) and get stems of non common terms [6]. Every stemmed term occurs in a document certain number of times which is called its frequency of occurrence or TF. Thus document d can be represented as:

$$d = \{t_1 : w_1, t_2 : w_2, \cdots, t_m : w_m\} \quad (1)$$

where t_i is i^{th} term in d, w_i is term frequency of t_i in d and m is total number of terms in d [2].

It was observed that the terms that occur at the beginning of the document occur more frequently than other terms. They contribute more in giving meaning to the document. Weights wt of each term t based on the positional appearance of the term in the document were calculated, as follows:

$$wt = \sum_{i=1}^{n} x_i \times p_i \quad (2)$$

where a document is assumed to be divided into n equal parts and x_i is frequency of t in i^{th} part of the document and p_i is weight assigned to t in ith part such that p1 > p2 > …> pn.

Using eq. 2 weights are assigned to terms in training database. Training database is used to build a classification model based on MNB. Using this model, classes of documents in test database were predicted. Higher posterior probability indicates class of the document.

The algorithm is as follows:

TC(Document Classes $\{c_i\}$, Documents in Training database {TR}, Documents in Test Database {T}, Partitions n) {

Step 1: Pre-processing of documents: For all documents in {TR} and {T}

1. Remove stop words
2. Find stems of remaining terms $\{t_j\}$

Step 2: Preparing Term-document matrix: For each term t in $\{t_j\}$

Find in each document d from {TR} if t is present

1. If present, divide d into n partitions
2. For each part p_i of d
   a. Count number of times x_i, t occurs in p_i^{th} part of d
3. Calculate weight wt of t using eq. 2
4. $wt = \sum_{i=1}^{n} x_i \times p_i$
5. Prepare term-document matrix for all documents in {TR} and assign a class c_i to each document manually

Step 3: Prepare a classification model using Multinomial Naïve Bayes classification algorithm using {TR}

Step 4: Evaluate the model for all documents in {T}

1. Class for each document in {T} is predicted

}

3. RESULTS AND DISCUSSION

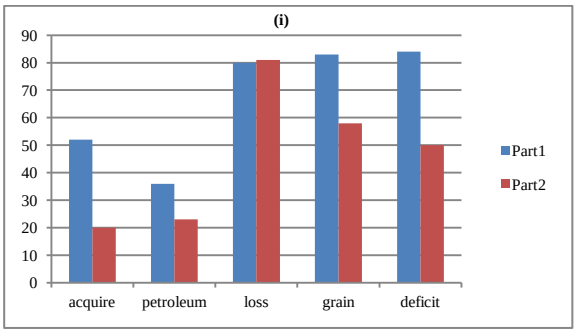
After pre-processing, a term-document matrix was prepared. Using eq. 2, weight of each term from the matrix was calculated. MNB was used to prepare a classification model using documents in training database. The model was evaluated using documents in test database.

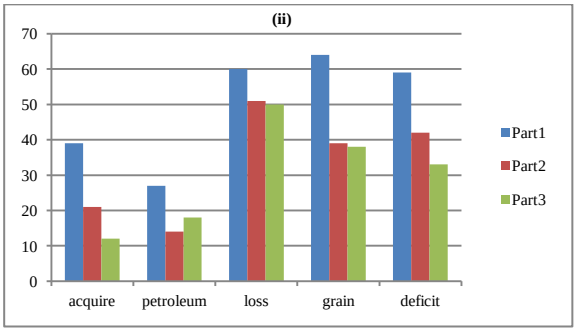
Two standard text datasets were used for testing – Reuters-21578 and 20-newsgroups. Experiments were conducted using WEKA's MNB [7].

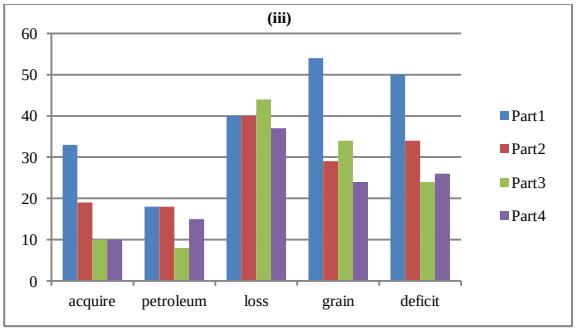
Reuters-21578 corpus contains five different categories out of which TOPICS category was selected. It has 135 sub-categories. This paper focuses on 5 sub-categories.

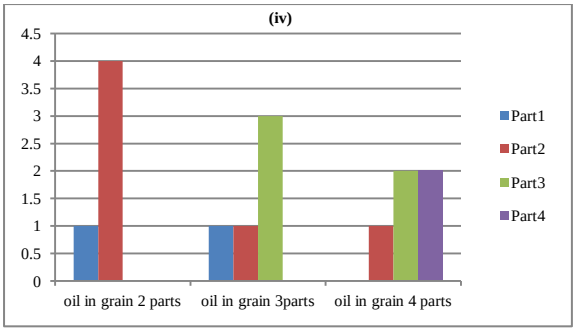
Experiments were carried out by dividing every document into 2, 3 and 4 parts. As per eq. 2 more weights were given to the terms that occurred in earlier parts of the document than to the terms that occurred in later parts. The classification results were improved. It was observed that a term relevant to the class starts occurring from initial part of the document. It occurs more number of times in the initial part, generally, than in subsequent parts. But a term less relevant or irrelevant to a class appears somewhere towards later parts of the document. Therefore, enhancing weights of terms that appear in the initial part improves classification results.

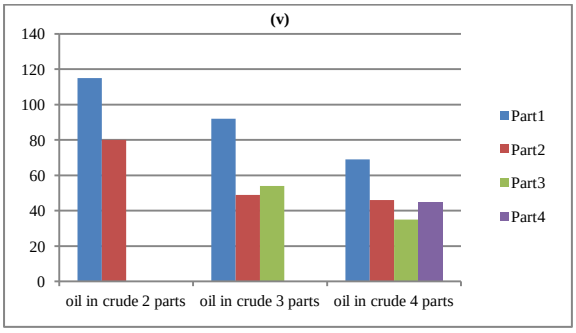
Fig 1 shows various graphs. (i), (ii) and (iii) show distribution of representative terms 'acquire', 'petroleum', 'loss', 'grain' and 'deficit' from 'acq', 'crude', 'earn', 'grain' and 'trade' respectively. These terms are relevant to respective classes. It was observed that except for term 'loss' rest of the terms appeared more number of times in first part than in subsequent parts. 'loss' appeared equal number of times in first and second part when documents were divided into 2 parts. When documents were divided into 4 parts, 'loss' occurred more number of times in third part than in first and second part. (iv) shows graph of 'oil'. 'oil' is related to 'grain' in context of 'palm oil' or 'linseed oil'. 'oil' is more relevant to 'crude'. It was observed that 'oil' occurred more number of times in later parts of documents relevant to 'grain' but same term occurred more number of times in first part than in other parts in documents relevant to 'crude' (v). F-measure for documents before and after enhancing the weights is shown in (vi). F-measure has not increased for 'crude' and 'trade'. Although number of true positives have increased by 1.3% for 'crude' and remained same for 'trade', numbers of false positives have also increased in both the cases. In case of 'crude', false positives have increased by 2.7%. In case of 'trade' false positives have increased by 1.3%. Thus there is no change in F-measure for these two classes.











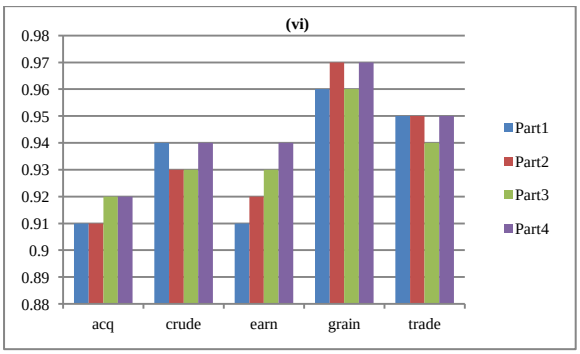
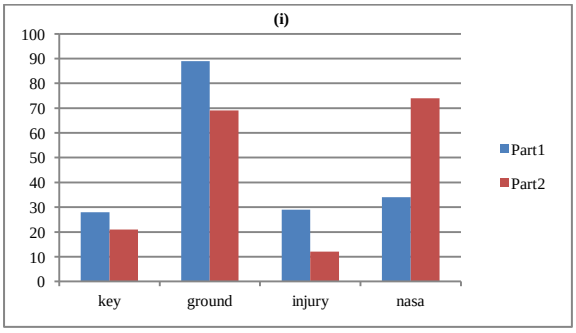
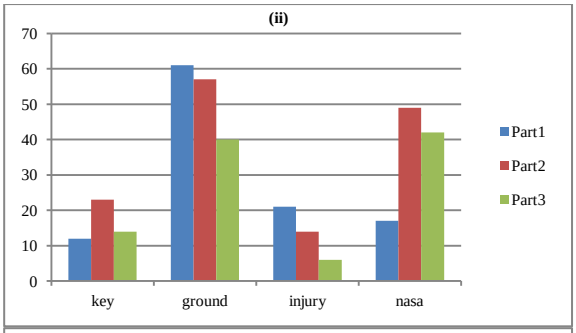
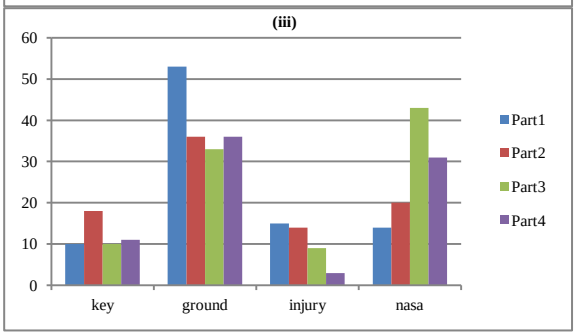


Fig.1: Experiments on Reuters-21578 (i), (ii) and (iii) distribution of terms 'acquire', 'petroleum', 'loss', 'grain' and 'deficit' from 'acq', 'crude', 'earn', 'grain' and 'trade' respectively when documents were divided into 2, 3 and 4 parts. (iv) and (v) distribution of 'oil' in 'grain' and 'crude' respectively. (vi) F-measure when documents were not divided, divided into 2, 3 and 4 parts.

Experiments were carried out for 20-newgroups. It contains documents from twenty newsgroups. All sub-categories under science were considered.







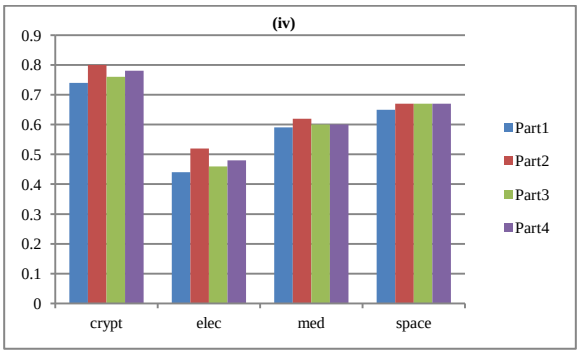


Fig.2: Experiments on 20-newsgroups (i), (ii) and (iii) distribution of terms 'key', 'ground', 'injury' and 'nasa' from 'crypt', 'elec', 'med' and 'space' respectively when documents were divided into 2,3 and 4 parts. (iv) F-measure when documents were not divided, divided into 2, 3 and 4 parts.

Graphs were plotted for 20-newsgroups. Fig. 2 shows graphs for 20-newsgroups dataset. 'key', 'ground', 'injury' and 'nasa' represent 'crypt', 'elec', 'med' and 'space' classes respectively ((i), (ii) and (iii)). Similar observations were made for all terms as were done for Reuters-21578 except 'nasa'. 'nasa' occurred less frequently in initial parts of documents which were relevant to 'space'. It was observed that classification results were improved after enhancing the weights of terms. F-measure for documents has been plotted in (iv). F-measure has increased when documents were divided into 2, 3 and 4 parts.

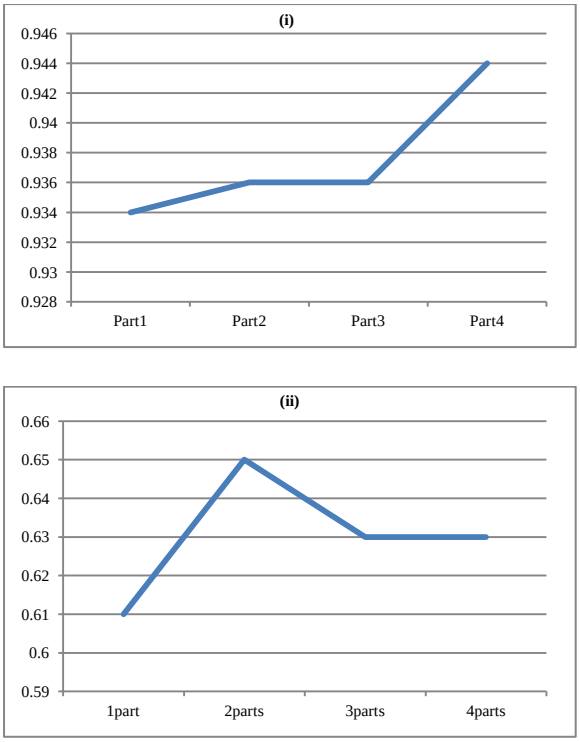


Fig.3: Average F-measures for (i) Reuters-21578 and (ii) 20-newsgroups datasets

Fig. 3 shows graphs for both datasets when documents were not divided and divided into 2, 3 and 4 parts. In both cases, improvement was seen in average F-measure. Reuters-21578 is known to be a "simple" dataset, so only a few terms are enough to classify the documents correctly [8]. As opposed to this, 20-newsgroups is not a "simple" dataset and every single term is required for correct classification [8]. So, when relevant terms have enhanced weights, average F-measure increases.

4. CONCLUSION

Consideration of positional distribution of terms in a document has helped in improving classification. Average F-measure was improved by 1.0% for Reuters-21578 and by 2% for 20-newsgroups datasets at least. This technique could be used by search engines while looking for text documents especially when more than two keywords are submitted as query keys to fetch the relevant results.

Further work can be done in improving classification by reducing number of false positives. Partitioning can be applied on documents depending upon their sizes.

5. REFERENCES

[1] David D. Lewis. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. Proc. of the 10th European Conference on Machine Learning 1998, pp. 4-15.

[2] A. McCallum and K. Nigam. A comparison of event models for naïve Bayes text classification. Proc of AAAI, 1998.

[3] JDM Rennie, L. Shih, J. Teevan and D. R. Karger. Tackling the poor Assumption of Naïve Bayes Text Classifiers. Proc of the twelfth Intl Conf on Machine Learning (ICML) 2003.

[4] Y. Ko. A study of term weighting schemes using class information for text classification. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval 2012, pp. 1029–1030.

[5] M. Mendoza. A new term-weighting scheme for naïve Bayes text categorization. International Journal of Web Information Systems, Vol. 8 Issue 1 2012, pp. 55 – 72.

[6] C. J. van Rijsbergen. Information Retrieval. Butterworth, 1990.

[7] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1, 2009.

[8] R. Bekkerman and J. Allan. Using bigrams in text categorization. Department of Computer Science, University of Massachusetts, Amherst 1003 (2004).