Baltic J. Modern Computing, Vol. 9 (2021), No. 3, 290-302
https://doi.org/10.22364/bjmc.2021.9.3.05

Research on Improvement of N-grams Based Text Classification by Applying Pointwise Mutual Information Measures

Tsvetanka GEORGIEVA-TRIFONOVA

Faculty of Mathematics and Informatics, "St. Cyril and St. Methodius" University of Veliko Tarnovo, Veliko Tarnovo, Bulgaria

ORCID: 0000-0002-5997-2344

cv.georgieva@live.uni-vt.bg

Abstract. In the present paper, the text classification is examined, which is applied after extracting N-grams of words to obtain characteristics describing the text documents in the collection. The selection of the most important features in regard to the pre-defined categories is made. The built vector space model for representation of text documents is modified by pointwise mutual information (PMI) measures. The conducted experiments include computation of the accuracy and *F*-measure of text classification with different methods for feature selection, different number of selected attributes (N-grams of words) for different classifiers and different datasets. The results obtained show an improvement in the performance of the classification of short texts with unbalanced categories.

Keywords: text classification, N-grams of words, pointwise mutual information, feature selection

1. Introduction

Text classification is a task for data mining and its solution finds application in a number of areas. A special case of textual content is the so-called short text, which is gaining popularity with the rapid development of e-commerce, social networks, online communication. Short texts appear in the user reviews for online stores or products, posts, chat messages, etc. They usually consist of several sentences. Therefore, they cannot provide enough word co-occurrence to measure the similarity of the documents. In addition, the short texts are characterized by the presence of noise (due to typographical and spelling errors) and unbalanced categories.

Due to these peculiarities of the short texts, the execution of the classifiers on the vector space model (based on the occurrence frequency, tf–idf (term frequency – inverse document frequency)) does not lead to acceptable accuracy. This motivates the study of an approach to modify the vector space model, which includes:

- Extracting N-grams of words to obtain features describing the text documents in the collection;

- Performing selection of important features in regard to predefined categories;
- Modifying their weights (tf–idf) through pointwise mutual information measures of words in regard to the categories.

The rest of the paper is organized as follows. Section 2 reviews existing studies for classifying text with unbalanced categories and short texts. Section 3 describes the approach proposed in the present research. Section 4 represents, summarizes and analyzes the results of performed experiments.

2. Related work

The conducted study shows the existence of research experience (Padurariu and Breaban, 2019) to overcome the problems associated to the classification of text with unbalanced categories. Approaches such as document differentiation into smaller categories by weight calculation (Liu et al., 2009), (Naderalvojoud et al., 2015), (Sarkar and Datta, 2017) are usually followed; feature selection (Zheng et al., 2014); neural network based methods (Donicke et al., 2019).

Song et al. (2014) summarize the characteristics inherent in short texts and systematize existing methods for improving their classification. The researched approaches are applied for different domains (Li et al., 2016), (Adhi et al., 2019); different languages (Gharavi and Bijari, 2017). Most often, the methods are based on topics modeling (Zeng et al., 2018), but more specific approaches are used, such as Wikipedia concepts mapping (Wang et al., 2013); construction of taxonomic-based features (Škrlj et al., 2021).

The present research concerns the text classification, which is applied after extracting N-grams of words to obtain features describing the text documents in the collection. The most important features in regard to the predefined categories are selected. The vector model space for representing the text documents constructed in this way is modified by using pointwise mutual information measures. The values of accuracy and *F*-measure are calculated and summarized for a different number of selected attributes, which represent N-grams of words, for different classifiers and two datasets. The results obtained show an improvement in the data classification for the dataset containing short texts for the different classifiers.

3. Applying PMI measures for N-grams based text classification

3.1. Computation of PMI measures

Pointwise mutual information (Mladenic and Grobelnik, 1999) between term t_i and category C_k is defined by the following way:

$$PMI(t_i, C_k) = \log \frac{P(t_i|C_k)}{P(t_i)} = \log \frac{P(t_i, C_k)}{P(t_i)P(C_k)}$$

The defined by this way pointwise mutual information compares the joint probability of the word t_i and category C_k with the probability t_i and C_k occurring independently. If

there is an association between the occurrence of t_i and C_k, then the probability $P(t_i, C_k)$ exceeds $P(t_i)P(C_k)$ and the pointwise mutual information receives a positive value; if there is no significant relationship between both events, the pointwise mutual information has a value close to 0.

An equivalent way to compute the pointwise mutual information is:

$$PMI(t_i, C_k) = \log\left(n\frac{N(t_i,C_k)}{N(t_i)N(C_k)}\right) \quad (1)$$

where $N(t_i, C_k)$ is the number of documents that contain the term t_i and belong to a category C_k; $N(t_i)$ is the number of documents that contain the term t_i; $N(C_k)$ is the number of documents that belong to a category C_k.

On the other hand, in the present research we consider $P(t_i|C_k)$ as the conditional probability of the event a term to be t_i, provided that it is contained in a document of category C_k; $P(t_i)$ – the probability that a term is t_i. Therefore

$$PMIt(t_i, C_k) = \log\left(N\frac{Nt(t_i,C_k)}{Nt(t_i)Nt(C_k)}\right) \quad (2)$$

where N is the total number of all occurrences of all terms in all documents; $Nt(t_i, C_k)$ is the number of occurrences of term t_i in documents of category C_k; $Nt(t_i)$ is the number of occurrences of term t_i in all documents; $Nt(C_k)$ is the total number of the occurrences of all terms in documents of category C_k.

In addition, we calculate a modified PMI-based measure (*m*PMIt; modified PMI; modified pairwise mutual information). For this purpose, we use the relative frequency of occurrence of the term t_i in documents of category C_k, applying equality (3).

$$mPMIt(t_i, C_k) = \frac{Nt(t_i,C_k)}{Nt(C_k)}PMIt(t_i, C_k) = \frac{Nt(t_i,C_k)}{Nt(C_k)}\log\left(N\frac{Nt(t_i,C_k)}{Nt(t_i)Nt(C_k)}\right) \quad (3)$$

Consequently

$$mPMIt(t_i, C_k) = P(t_i|C_k).\log\frac{P(t_i|C_k)}{P(t_i)}$$

We calculate in a similar way *m*PMI:

$$mPMI(t_i, C_k) = \frac{Nt(t_i,C_k)}{Nt(C_k)}\log\left(n\frac{N(t_i,C_k)}{N(t_i)N(C_k)}\right) \quad (4)$$

3.2. Applying the computed PMI measures for modifying the word weights in the vector space model

In this paper, for the purposes of text classification, we propose building a model, which consists of performing the following steps:

1. We construct the vector space model by computing tf–idf weights of N-grams of terms;
   As a result, we obtain a matrix T of type $m \times n$, where m is the number of documents, n is the number of extracted N-grams of words.

2. Applying a method for feature selecting to extract the most important features (i.e. N-grams of words) with regard to the defined categories;
   We denote the resulting matrix by *FS* of type $m \times s$, where s is the number of selected N-grams of words.
3. Computing PMI measures of selected N-grams of words with regard to the categories.
   As described in subsection 3.1, we compute *PMIt*, *mPMIt*, *PMI*, *mPMI*, which are matrices of type $s \times k$, where k is the number of categories.
4. Modifying the weights obtained in step 2.
   We compute the modified weights by multiplying the matrices *FS* and any of those obtained in step 3. The resulting matrix is of type $m \times k$ and the corresponding classifier is applied to it.

4. Experiments

In this section, the datasets, the feature selection methods, the classifiers used in the experiments, are pointed out. The results of the accuracy and the *F*-measure of text classification are represented.

4.1. Datasets

The datasets used in the experiments are subjected to pre-processing. It consists of tokenization, stop words removal, steaming. N-grams are found, where N is 5. The text documents are represented by the vector space model, and the term weights are tf–idf.

- *Reuters-21578* (Lewis, 1997);

This dataset consists of 21578 news articles in English distributed over 135 cross-cutting topics. It contains the documents that appeared in Reuters Newswire in 1987. In the experiments conducted in the present study, 10 categories are used (acq, crude, earn, grain, interest, money-fx, oilseed, ship, sugar, trade). The largest category (earn) contains 3735 documents, the smallest (oilseed) – 9. The total number of documents in them is 7363, the total number of different words (stems) is 14889. The N-grams found are filtered by selecting the occurring ones in less than 1.0% of the documents and a total of 1409 remain. The average number of words in the documents is approximately 64; of unique words – approximately 40. The average length of texts in *Reuters-21578*, expressed in number of characters (after removing characters representing the end of a word as a space, punctuation marks) is 656.

- *Customer_feedback_bg* (Georgieva-Trifonova et al., 2018).

The dataset *Customer_feedback_bg* consists of user reviews for online stores in Bulgarian. The data are extracted from otzivi.bg and pazaruvaj.com, and represent user reviews for 87 online stores. Total of 906 user reviews in free text are collected and are manually associated with the following categories: compliments, complaints, mixed, suggestions. The largest category (compliments) contains 540 documents, the smallest (suggestions) – 44. The total number of different words (stems) in the documents of the dataset *Customer_feedback_bg* is 2842. Of all N-grams are filtered those that occur in less than 0.1% of the documents, resulting in 76930. The average number of words in the documents is approximately 26; of unique words – approximately 22. The average

length of the texts in *Customer_feedback_bg*, expressed in number of characters (after removing characters representing the end of a word as a space, punctuation marks) is 253.

Both datasets have unbalanced categories, but differ in language, number of categories, length of texts.

4.2. Methods for feature selection

The methods for feature selection used in the experiments are:

- Relief algorithm (Kira and Rendell, 1992);
- Chi-squared feature selection (Yang and Pedersen, 1997), (Forman, 2003);
- Information gain feature selection (Yang and Pedersen, 1997), (Forman, 2003);
- Gini index feature selection (Shang et al., 2007).

4.3. Classifiers

The following classifiers are applied in the conducted experiments:

- K-nearest neighbors (K-NN) (Lu and Bai, 2010);
  In the experiments, the measure cosine similarity is used as a measure of the distance between the instances in the datasets.
- Decision tree (DT) (Mitchell, 1996);
  Gain ratio is applied as a measure to define the criterion for selecting a splitting attribute. It represents the ratio of information gain to intrinsic information.
- H2O's Deep Learning (Candel and Parmar, 2015);
  The rule-based classifiers RIPPER (JRip) (William and Cohen, 1995), Ridor (Gaines and Compton, 1995), PART (Frank and Witten, 1998).

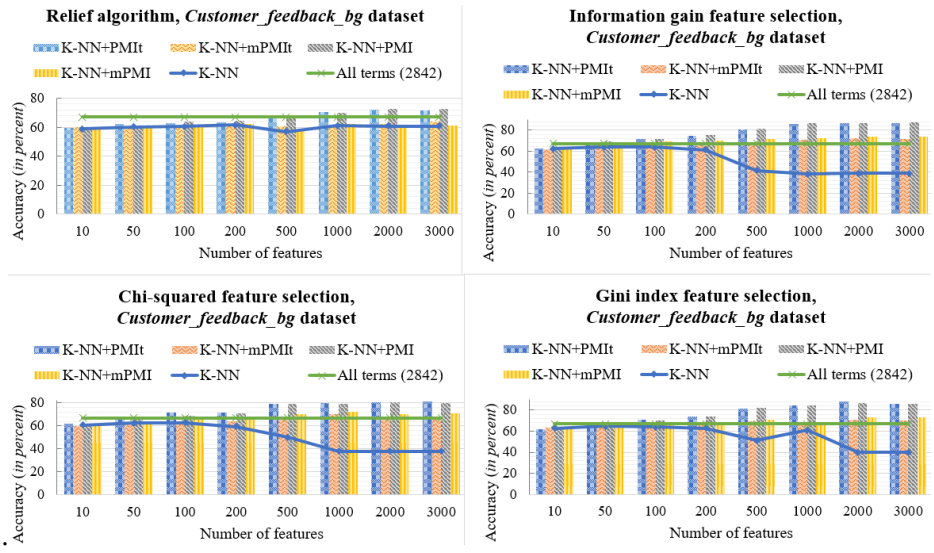

Figure 1: The accuracy of K-NN algorithm for *Customer_feedback_bg* dataset

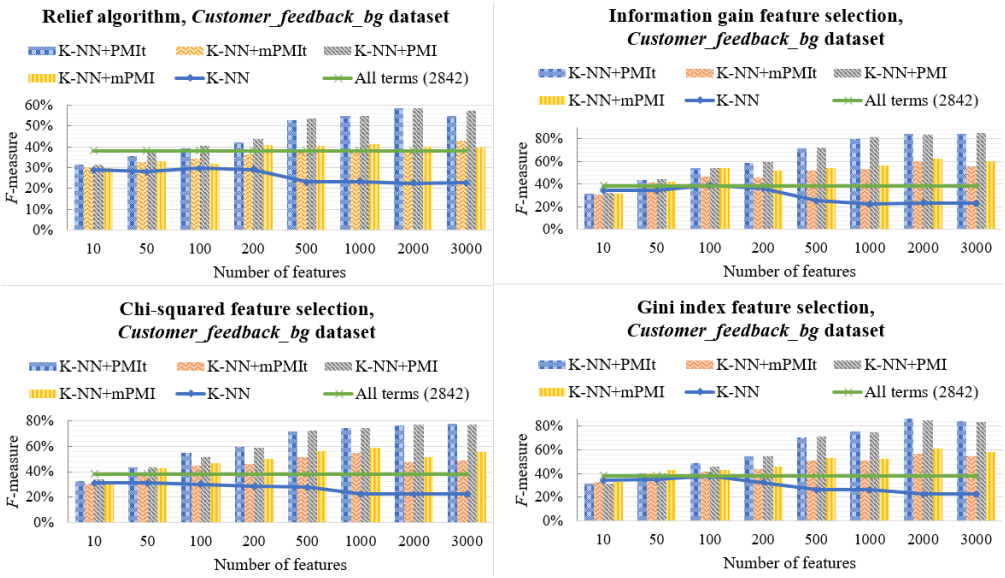

Figure 2: *F*-measure of K-NN algorithm for *Customer_feedback_bg* dataset

Figure 3: The accuracy of DT algorithm for *Customer_feedback_bg* dataset

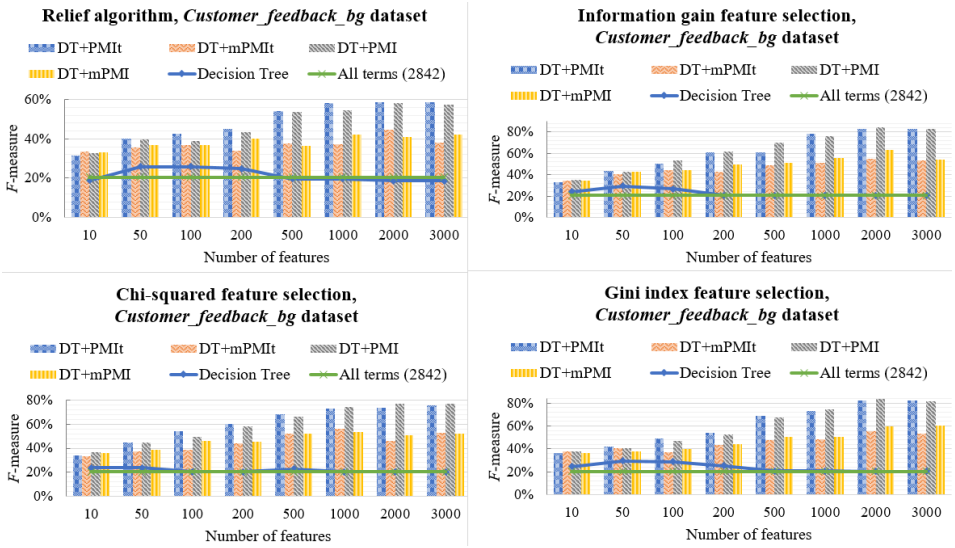

Figure 4: *F*-measure of DT algorithm for *Customer_feedback_bg* dataset

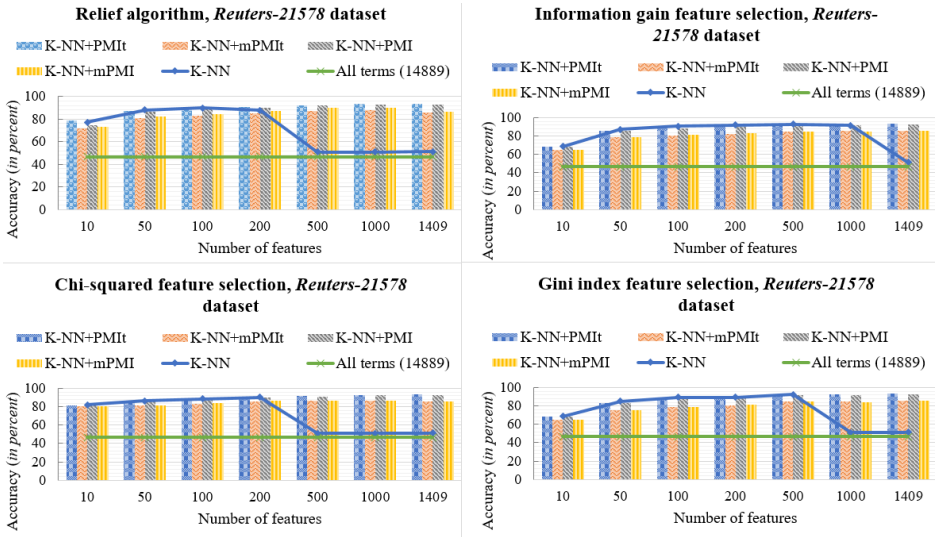

Figure 5: The accuracy of K-NN algorithm for *Reuters-21578* dataset

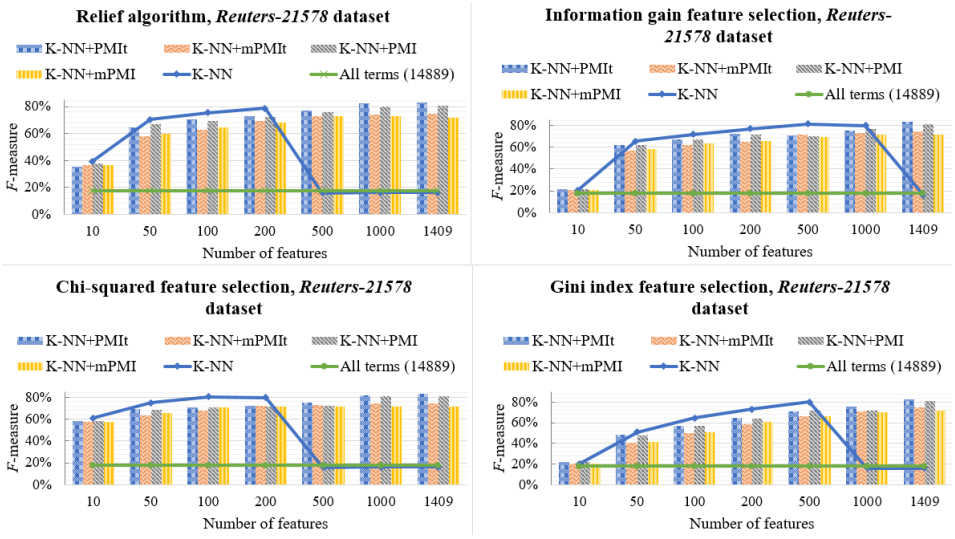

Figure 6: *F*-measure of K-NN algorithm for *Reuters-21578* dataset

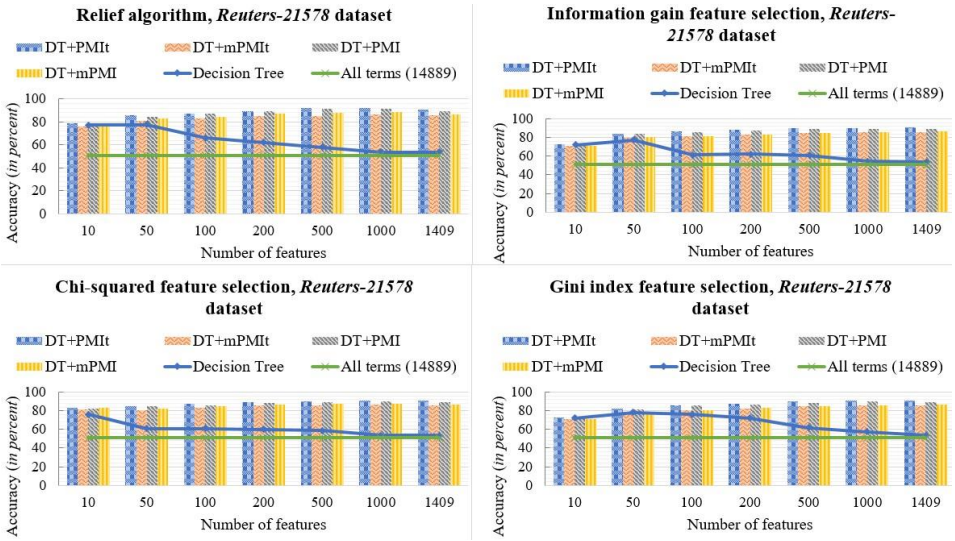

Figure 7: The accuracy of DT algorithm for *Reuters-21578* dataset

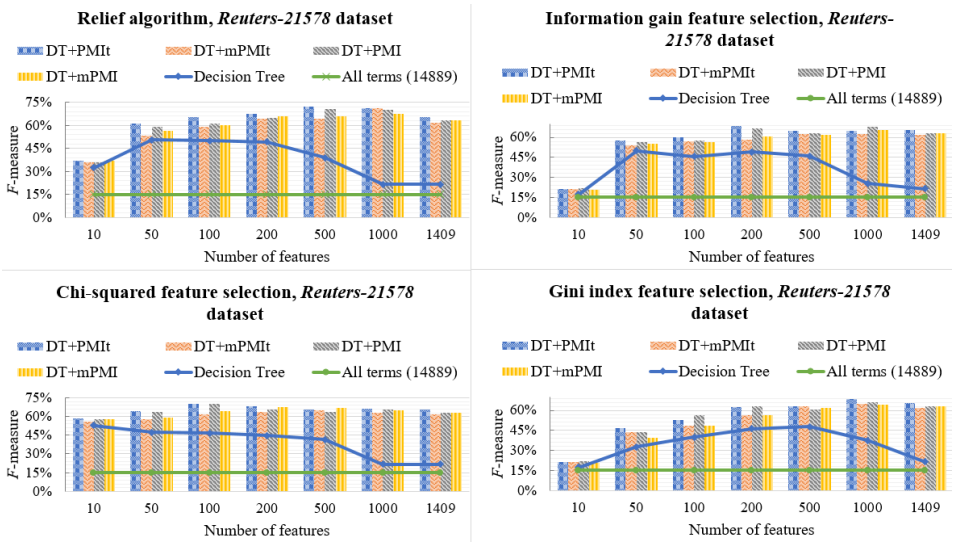

Figure 8: *F*-measure of DT algorithm for *Reuters-21578* dataset

4.4. Results

Figures 1-4 show the results from the accuracy and *Macro F*-measure about *Customer_feedback_bg* dataset for the different feature selection methods and PMI measures for K-NN and DT classifiers. Figures 5-8 show similar results for *Reuters-21578* dataset.

When applying DT classifier, there are an improvement in the accuracy and *Macro F*-measure for both datasets. In the execution of the K-NN classifier, in the majority of cases an improvement is observed when the number of selected features is at least 500 of the considered datasets.

All detailed results for the listed classifiers are available and freely available from (Georgieva-Trifonova, 2021). The appendix of this paper includes the diagram presentations of the results from the accuracy and *Macro F*-measure for H2O's Deep Learning algorithm for the considered datasets.

For the other classifiers (H2O's Deep Learning, JRip, Ridor, PART), there is no improvement in the classification after PMI-based weight modification when the dataset *Reuters-21578* is used. But on the other hand, the modification of the weights gives a significant improvement of the classification of the short texts contained in the dataset *Customer_feedback_bg* for all applied classifiers. In addition, for the short texts in *Customer_feedback_bg* dataset comparable values of the accuracy and *Macro F*-measure for *Reuters-21578* dataset are achieved, obtained without weight modification.

5. Conclusion

In the present paper, PMI-based modification of the weights of N-grams of words for text classification purposes is examined. It is applied after selecting the most important features in regard to the pre-defined categories. The results of the experiments show an improvement in the accuracy and the *Macro F*-measure for the dataset containing short texts.

Appendix

Figures 9 and 10 illustrate the results from the accuracy and *Macro F*-measure about *Customer_feedback_bg* dataset for the different feature selection methods and PMI measures for H2O's Deep Learning classifier. Figures 11 and 12 depict similar results for *Reuters-21578* dataset.

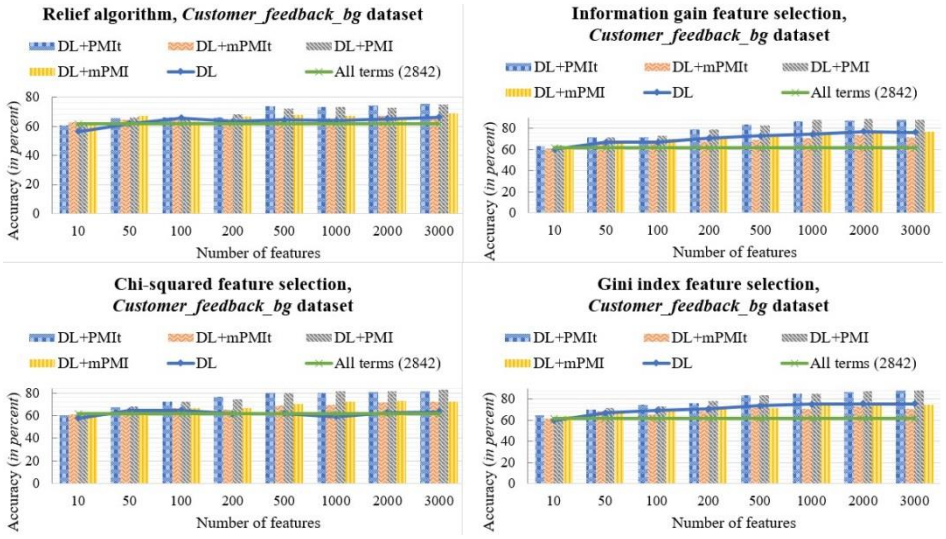

Figure 9: The accuracy of DL algorithm for *Customer_feedback_bg* dataset

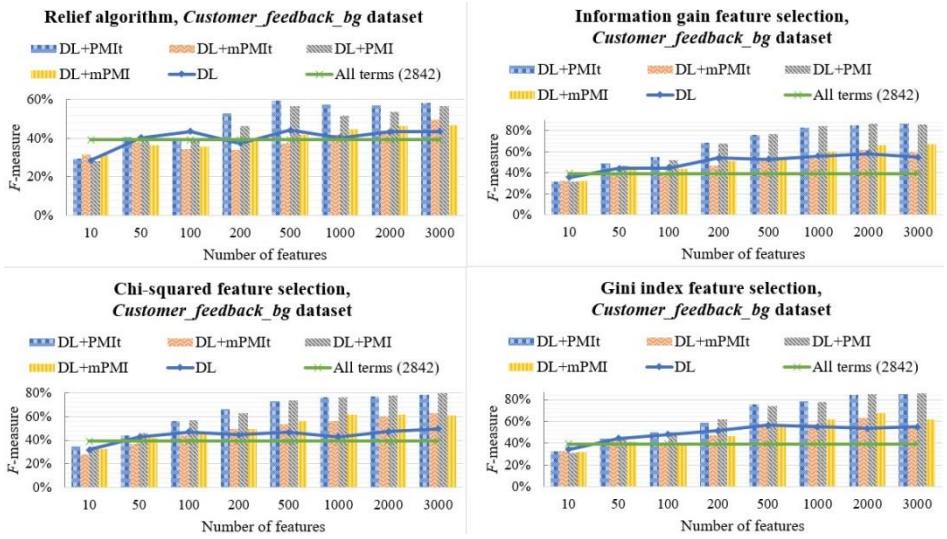

Figure 10: *F*-measure of DL algorithm for *Customer_feedback_bg* dataset

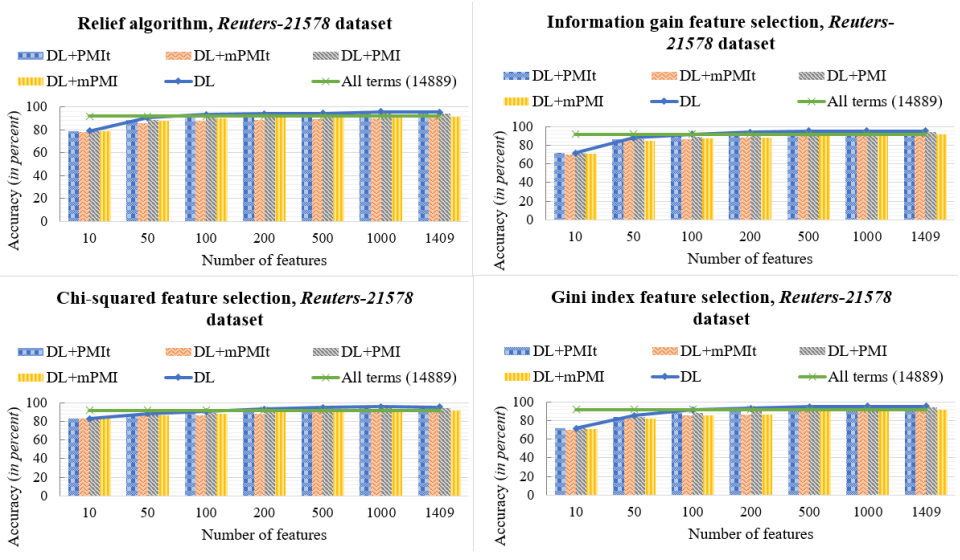

Figure 11: The accuracy of DL algorithm for *Reuters-21578* dataset

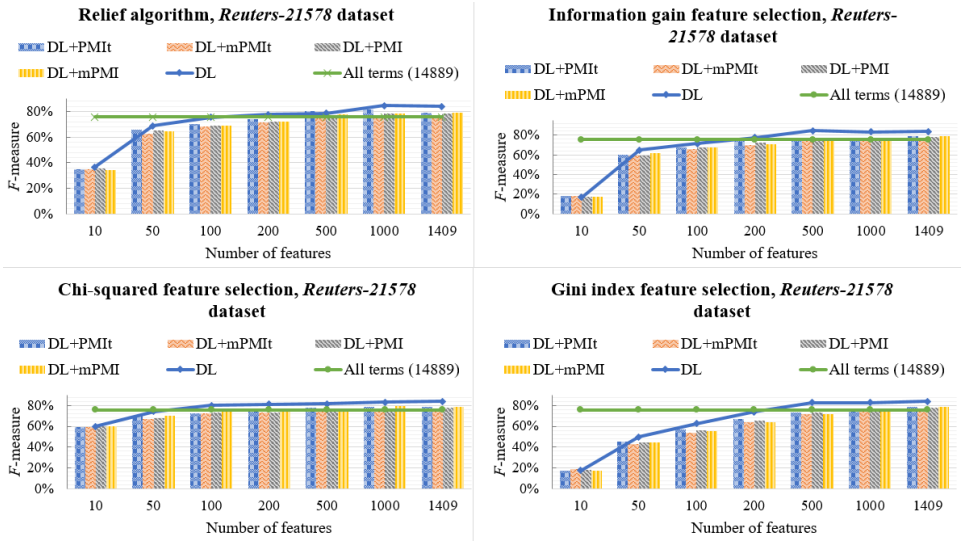

Figure 12: *F*-measure of DL algorithm for *Reuters-21578* dataset

References

Adhi, B. P., Saskiah, D., Widodo, W. (2019). A Systematic Literature Review of Short Text Classification on Twitter, *KnE Social Sciences* **3**(12), 625-635.

Candel, A., Parmar, V. (2015). *Deep learning with H2O*, H2O.ai, Inc.

Donicke, T., Lux, F., Damaschk, M. (2019). Multiclass Text Classification on Unbalanced, Sparse and Noisy Data, *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pp. 58–65.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification, *Journal of Machine Learning Research* **3**, pp. 1289-1305.

Frank, E., Witten, I. H. (1998). Generating accurate rule sets without global optimization *In Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 144-151.

Gaines, B. R., Compton, P. (1995). Induction of ripple-down rules applied to modeling large databases, *Journal of Intelligent Information Systems* **5**(3), 211–228.

Georgieva-Trifonova. T., Stefanova. M., Kalchev, S. (2018). Dataset for: customer feedback text analysis for online stores reviews in Bulgarian, available at: https://doi.org/10.7910/DVN/TXIK9P, Harvard Dataverse.

Georgieva-Trifonova, T. (2021), Dataset for: research on improvement of N-grams based text classification by applying pointwise mutual information measures, available at: https://doi.org/10.7910/DVN/NTL2HT, Harvard Dataverse.

Gharavi, E., Bijari, K. (2017). Short text classification using deep representation: A case study of Spanish tweets in Coset Shared Task, *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages*, pp. 28-35.

Kira, K., Rendell, L. (1992). The feature selection problem: traditional methods and a new algorithm, *Proceedings of the tenth national conference on Artificial intelligence*, pp. 129-134.

Lewis, D. D. (1997). Reuters-21578 text categorization test collection, available at: https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.

Liu, Y., Loh, H., Sun, A. (2009). Imbalanced text classification: A term weighting approach, *Expert Systems with Applications* **36**(1), 690-701.

Li, Y., Tripathi, A., Srinivasan, A. (2016). Challenges in Short Text Classification: The Case of Online Auction Disclosure, *Proceedings of the Mediterranean Conference on Information Systems*, pp. 1-13.

Lu, F., Bai, Q. (2010). A Refined weighted k-nearest neighbours algorithm for text categorization, *In Proceedings of International Conference on Intelligent Systems and Knowledge Engendering*, IEEE, pp. 326-330.

Mitchell, T. M. (1996). *Machine learning*, New York: McGraw Hill.

Mladenic, D., Grobelnik, M. (1999). Feature Selection for Unbalanced Class Distribution and Naïve Bayes, *In Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pp. 258-267.

Naderalvojoud, B., Sezer, E. A., Ucan, A. (2015). Imbalanced Text Categorization Based on Positive and Negative Term Weighting Approach, *Lecture Notes in Computer Science*, vol. 9302. Springer, pp. 363–371.

Padurariu, C., Breaban, M. E. (2019). Dealing with Data Imbalance in Text Classification, *Procedia Computer Science* **159**, 736-745.

Sarkar, A., Datta, D. (2017). A Frequency Based Approach to Multi-Class Text Classification, *International Journal of Information Technology and Computer Science* **9**(5), 15-22.

Shang. W., Huang ,H., Zhu, H., Lin, Y., Qu, Y., Wang, Z. (2007). A novel feature selection algorithm for text categorization, *Expert Systems with Applications* **33**(1), 1-5.

Škrlj, B., Martinc, M., Kralj, J., Lavrač, N., Pollak, S. (2021). tax2vec: Constructing Interpretable Features from Taxonomies for Short Text Classification, *Computer Speech & Language* **65**(2021), 1-21.

Song, G., Ye, Y., Du, X., Huang, X., Bie, S. (2014). Short Text Classification: A Survey, *Journal of multimedia* **9**(5), 635-543.

Wang, X., Chen, R., Jia, Y., Zhou, B. (2013). Short Text Classification Using Wikipedia Concept Based Document Representation, *International Conference on Information Technology and Applications*, Chengdu, pp. 471-474.

William. W., Cohen (1995). Fast effective rule induction *In Proceedings of the Twelfth International Conference on Machine Learning*, pp. 115-123.

Yang, Y., Pedersen, J. O. (1997). A comparative study on feature selection in text categorization, *In Proceedings of the 14th International Conference on Machine Learning*, pp. 412-420.

Zeng, J., Li, J., Song, Y., Gao, C., Lyu, M. R., King, I. (2018). Topic Memory Networks for Short Text Classification, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 3120–3131.

Zheng, Z., Wu, X., Srihari, R. (2014). Feature Selection for Text Categorization on Imbalanced Data, *ACM SIGKDD Explorations Newsletter* **6**(1), 80-89.

Received June 20, 2021, revised August 6, 2021, accepted August 11, 2021