Research on N-grams feature selection methods for text classification

Tsvetanka Georgieva-Trifonova¹, Mahmut Duraku¹
¹“St. Cyril and St. Methodius” University of Veliko Tarnovo, Veliko Tarnovo, Bulgaria

e-mail: cv.georgieva@live.uni-vt.bg

Abstract. Text classification requires previously extraction of features describing the text documents in the collection. Usually these features are based on the occurrence frequency of words, N-grams of words in documents, i.e. the vector space model for document representation is built. Feature selection allows to reduce redundancy in high-dimensional representation of text data, which can significantly improve text classification performance. In the present paper, research on feature selection methods is performed in terms of the accuracy and F-measure of text classification with different number of selected attributes (N-grams of words) for different classifiers and different datasets. The obtained results can be used to apply further pre-processing steps, which include modifying the vector space model in order to achieve its improvement in terms of the text classification.

1. Introduction
The classification of text documents consists in associating each document with a category that is among a set of predefined categories. There is a growing interest in this text mining task, due to the accumulation of more and more text data in digital format. Methods for solving it are useful for many applications, such as information retrieval, web page classification, email filtering, content management systems, sentiment analysis.

In most text classification methods, the presentation of documents is based on the vector space model (VSM), which is widely used in text mining due to its simplicity and acceptable performance of algorithms. In this model, the textual content is considered as a "bag of words" and each document is represented as a vector of features, where each feature corresponds to a unique word (term) of the documents. Term weight is calculated to express the importance of each feature in the document and the collection of documents. Because VSM ignores word combinations and word order, its disadvantage is the loss of semantic and syntactic information. Various approaches are proposed to overcome this problem, with some researchers trying to construct more complex features derived from Natural Language Processing (NLP) such as phrases, syntactic relationships [1]; frequent term set [2]; bigrams [3], N-grams of words [4, 5, 6], etc.

Another problem that arises is the high dimensionality of the presentation of text data. To solve it, it is necessary to apply feature selection to remove redundant and unrelated features. In the present paper, methods for feature selection are examined in terms of accuracy and F-measure of text classification for different number of selected attributes (N-grams of words) for different classifiers and different datasets. The obtained results are discussed and some assumptions are made about their future usage in order to improve the performance of text classification.
2. Related work

This section reviews the existing experience in the usage of N-grams for the text document presentation in text classification and the application of filtering methods for feature selection.

Feature selection methods calculate a feature weight (*feature score*) for each characteristic, which is used to assess the relevance of the characteristic to the target concept. Filtering approaches evaluate the importance of the features based only on their inherent properties through statistical calculations. These methods are distinguished by the fact that they are fast and with reasonable calculation costs. Therefore, they are widely used in high dimensional data, such as text data.

In [7], the automatic identification of authorship is considered as a task for classifying texts. N-grams are used as features because they successfully represent text for stylistic purposes. Character N-grams are extracted from a subset of the Reuters dataset and an approach for their selection is proposed, which is compared with Information gain feature selection.

In [8], chi-squared feature selection is applied, where unigrams and bigrams are extracted as features. Experiments are performed with Naive Bayes classifier for a specific dataset.

In [9], an approach for feature selection for N-grams based classification of text in Chinese language is presented. A comparative analysis is performed after applying SVM (Support Vector Machine) and Naive Bayes classifiers.

[10] presents preliminary results from a study on N-grams based feature selection for the purposes of sentiment analysis of texts on twitter related to commercial products. The weights are dictionary extracted and Naive Bayes classifier is applied.

There are some recent studies [11, 12, 13], which focus on specific types of texts or specific measures for selecting features.

The present research concerns the feature selection in order to improve the text classification. Frequently used filtering methods for feature selection are applied and the values of accuracy and *F*-measure are calculated for a different number of selected attributes, which represent N-grams of words, for different classifiers and two datasets. The obtained results are summarized and guidelines for future development of the research are given.

3. Research on methods for feature selection from word N-grams for text classification

The methods for feature selection used in the experiments are:

- Relief algorithm [14];
- Chi-squared feature selection [15, 16];
- Information gain feature selection [15, 16];
- Gini index feature selection [17].

3.1. Datasets

The datasets used in the experiments are described below. They are subjected to pre-processing, which consists of tokenization, removal of stop words, stemming. N-grams are found where N is 5. Text documents are represented by the vector space model, where the word weights are *tf–idf* (*term frequency – inverse document frequency*).

- **Reuters-21578** [18];
  This dataset consists of 21578 news articles in English distributed among 135 intersecting topics. It contains the documents that appeared in Reuters Newswire in 1987. Usually, in the studies related to text classification, the ten or twelve categories with the most documents are selected. In the experiments conducted in the present study, 10 categories are used (acq, crude, earn, grain, interest, money-fx, oilseed, ship, sugar, trade). The total number of documents in them is 7363, the total number of different words (stems) is 14889. The found N-grams are filtered, by selecting the ones appeared in less than 1.0% of the documents and leaving a total of 1409. The average number of words in the documents is approximately 64; of unique words – approximately 40.
- **Customer_feedback_bg** [19].
The dataset Customer_feedback_bg consists of user reviews for online stores in Bulgarian language. The data are extracted from otzivi.bg and pazaruvaj.com, and represent user reviews for 87 online stores. This dataset consists of 906 user reviews in free text that are manually associated with the following categories: compliments, complaints, mixed, suggestions. The total number of different words (stems) in the documents of the dataset Customer_feedback_bg is 2842. Of all found N-grams, those appeared in less than 0.1% of the documents are filtered, leaving 76930. The average number of words in the documents is approximately 26; of unique words – approximately 22.

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

3.2. Classifiers

The following classifiers are applied in the conducted experiments:
- K-NN [20];
- Decision tree [21];
  Gain ratio is used 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 [22];
- The rule-based classifiers RIPPER (JRip) [23], Ridor [24], PART [25].

The implementation of the selected algorithms for text mining are performed by using RapidMiner (https://rapidminer.com).

3.3. Results

Appropriate measures to assess the validity of text classification are computed. Accuracy is defined by the ratio of the number of correctly classified documents to the total number of documents. The $F$-measure is defined as the mean harmonic value between the precision $P$ and the recall $R$:

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

The $F$-measure is usually useful for uneven distribution of categories in the dataset. After finding the values of the $F$-measure by category, the Macro $F$-measure is calculated as their average value.

Figures 1, 2, 3 and 4 show the results of the accuracy and Macro $F$-measure of the different classifiers for the dataset Reuters-21578, after applying the respective feature selection methods.

**Figure 1.** Accuracy and $F$-measure of different classifiers on Reuters-21578 dataset after feature selection by applying Relief algorithm.

On the graphs, *All terms* means the accuracy and the Macro $F$-measure by applying the respective classifier, when the features are all words after pre-processing (tokenization, removal of stop words, stemming), but without finding N-grams and selecting features. This illustrates for which Number of features values, the improvements are obtained in the accuracy and Macro $F$-measure as a result of feature selection (*All terms* are used as a baseline for comparison).
As a result of applying the feature selection methods for the dataset Reuters-21578, an improvement of the considered measures for the classifiers is observed:

- **K-NN;**
  - The improvement of the accuracy is for all Number of features values in all applied feature selection methods. The Macro $F$-measure archives better results when 10, 50, 100, 200 features are selected for all methods. When applying the Relief algorithm and Chi-squared feature selection, the best results of the Macro $F$-measure are observed for Number of features = 200; when applying Information gain and Gini index feature selection – for Number of features = 500.

- **Decision tree;**
There is an improvement in the accuracy and *Macro F*-measure for all feature selection methods and all values of *Number of features*.

- H2O’s Deep Learning.

Improvements to the measures occur when the number of selected features is at least 100.

Figures 5, 6, 7 and 8 show the results of the accuracy and *Macro F*-measure of the different classifiers for the dataset *Customer_feedback_bg*, after applying the respective methods for feature selection.

**Figure 5.** Accuracy and *F*-measure of different classifiers on *Customer_feedback_bg* dataset after feature selection by applying Relief algorithm.

**Figure 6.** Accuracy and *F*-measure of different classifiers on *Customer_feedback_bg* dataset after Chi-squared feature selection.

**Figure 7.** Accuracy and *F*-measure of different classifiers on *Customer_feedback_bg* dataset after Information gain feature selection.
As a result of applying the features selection methods for the dataset Customer_feedback_bg, an improvement of the considered measures for the classifiers is observed:

- **K-NN:**
  - The value of the Macro $F$-measure after applying Information gain feature selection for Number of features = 100 exceeds that without applying a feature selection method.
- **Decision tree:**
  - The Macro $F$-measure gets better results with the Relief algorithm when Number of features is 50, 100, 200. For the other feature selection methods, both measures show improvement for all values of Number of features.
  - **H2O’s Deep Learning:**
    - Better values of both measures are observed for all feature selection methods for Number of features at least 50.
  - **Rip:**
    - There is a preservation or slight improvement of the results of the measures when applying all feature selection methods.
  - **Ridor:**
    - The Macro $F$-measure shows a slight improvement for Number of features = 500 when applying Information gain feature selection.
  - **PART:**
    - Both measures give better values when applying Information gain and Gini index feature selection for Number of features at least 100.

A more detailed study of the results of the Macro $F$-measure and the values from which it is derived (i.e. the $F$-measures by category) can be drawn as follows, which applies to both datasets despite significant differences between them. There is a lack of improvement in the computed measures in the smallest categories (i.e. the categories with which at least text documents are associated). An approach to address this shortcoming is proposed in [26]. A model is described that enriches the vector space model by additionally extracting the pointwise mutual information (PMI) of the words with respect to the categories. Applying it to selected features among the extracted N-grams, and not to all words, is a useful direction for further research. It could answer the following research question: If, after removing redundant features, the model is enriched by PMI in terms of categories, whether the performance of text classification be improved? If the answer proves to be positive, the statistical significance of the results must be established using appropriate statistical methods [27].

![Figure 8. Accuracy and F-measure of different classifiers on Customer_feedback_bg dataset after Gini index feature selection.](image-url)
4. Conclusion
The present paper is devoted to the application of feature selection methods for text classification. For this purpose, the presentation of text documents is based on extracted N-grams of words. Feature selection methods are executed and the change in the classification performance in terms of accuracy and F-measure is tracked with a different number of selected attributes for different classifiers and specific datasets. The obtained results are illustrated and their usefulness for their future usage in order to improve the performance of text classification is justified.

References
[1] Moschitti A and Basili R 2004 Complex linguistic features for text classification: A comprehensive study *Lecture Notes in Computer Science* 2997 pp 181-196
[2] Yuan M, Ouyang Y X and Xiong Z A 2013 Text Categorization Method using Extended Vector Space Model by Frequent Term Sets *Journal of Information Science and Engineering* 29 pp 99-114
[3] Bekkerman R and Allan J 2004 Using Bigrams in Text Categorization *Technical Report IR-408* (UMass Amherst: Center of Intelligent Information Retrieval)
[4] Farhoodi M, Yari A and Sayah A 2011 N-gram based text classification for Persian newspaper corpus, *The 7th International Conference on Digital Content, Multimedia Technology and its Applications* pp 55-59
[5] Zhang X and Wu B 2015 Short Text Classification based on feature extension using The N-Gram model *In Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery* pp 710-716
[6] Graovac J, Kovacevic J and Pavlovic-Lažetic G 2015 Language Independent n-Gram-Based Text Categorization with Weighting Factors: A Case Study *Journal of Information and Data Management* 6 1 pp 4–17
[7] Houvardas J and Stamatatos E 2006 N-Gram Feature Selection for Authorship Identification *Artificial Intelligence: Methodology, Systems, and Applications* Lecture Notes in Computer Science vol 4183, ed J Euzenat and J Domingue (Berlin: Springer)
[8] Meena M J, Chandran K R, Brinda J M and Sindhu P R 2010 Enhancing Feature Selection Using Statistical Data with Unigrams and Bigrams *International Journal of Computer Applications* 1 1 1 pp 7-11
[9] Wei Z, Miao D, Chauchat J H, Zhao R and Li W 2012 N-grams based feature selection and text representation for Chinese Text Classification *International Journal of Computational Intelligence Systems* 2 4 pp 365-374
[10] Cabanlit M A and Espinosa K J 2014 Optimizing N-gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons *The 5th International Conference on Information, Intelligence, Systems and Applications* pp 94-97
[11] Vadehra A, Grossman M R and Cormack G V 2017 Impact of Feature Selection on Micro-Text Classification, available at: https://arxiv.org/abs/1708.08123
[12] Shedbale S and Shaw K 2017 N-Gram and KLD Based Efficient Feature Selection Approach for Text Categorization *International Journal of Advance Engineering and Research Development* 4 6 pp 660-668
[13] Ahuja R, Chug A, Kohli S, Gupta S and Ahuja P 2019 The Impact of Features Extraction on the Sentiment Analysis *Procedia Computer Science* 152 pp 341-348
[14] Kira K and 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
[15] Yang Y and 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
[16] Forman G 2003 An extensive empirical study of feature selection metrics for text classification *Journal of Machine Learning Research* 3 pp 1289-1305
[17] Shang W, Huang H, Zhu H, Lin Y, Qu Y and Wang Z 2007 A novel feature selection algorithm
for text categorization *Expert Systems with Applications* **33** 1 pp 1-5

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

[19] Georgieva-Trifonova T, Stefanova M and 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)

[20] Lu F and 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

[21] Mitchell T M 1996 *Machine learning* (New York: McGraw Hill)

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

[23] William W and Cohen 1995 Fast effective rule induction *In Proceedings of the Twelfth International Conference on Machine Learning* pp 115-123

[24] Gaines B R and Compton P 1995 Induction of ripple-down rules applied to modeling large databases *Journal of Intelligent Information Systems* **5** 3 pp 211–228

[25] Frank E and Witten I H 1998 Generating accurate rule sets without global optimization *In Proceedings of the Fifteenth International Conference on Machine Learning* pp 144-151

[26] Georgieva-Trifonova T, Stefanova M and Kalchev S 2018 Customer Feedback Text Analysis for Online Stores Reviews in Bulgarian *IAENG International Journal of Computer Science* **45** 4 pp 560-568

[27] Trifonov T and Tsonkova V 2009 *Statistics in economics and management* (Veliko Tarnovo: Astarta) (in Bulgarian)