Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/97/20241397

Review on Application of Chi-square Statistic in Text Classification in Recent Five Years

Chuanyu Tang

Mcmaster University, Faculty of Science, 1280 Main St W, Hamilton, ON L8S 4L8, Canada

2822807844@qq.com

Abstract. The swift expansion of online textual data has rendered text classification increasingly vital in information management. Despite the prevalent usage of the chi-square test in text classification, there has been a scarcity of thorough research regarding its specific uses in recent years. Therefore, it is vital to encapsulate the research about the use of the chi-square test in text classification throughout the last five years. This report reviews the application of the chi-square statistic in Arabic text classification, social media data analysis, and medical literature classification and analyses its effectiveness in feature selection and enhancing classification performance. By reviewing and analyzing the academic literature, this report summarizes the application of improved chi-square feature selection methods to different text data types. It explores the effectiveness of these methods in improving classification accuracy. The findings indicate that chi-square has significant advantages in text classification in different domains, especially when dealing with complex linguistic texts and user-generated content.

Keywords: Text classification, Chi-squared statistics, Feature selection, Arabic text classification, Natural language processing.

1. Introduction

The substantial increase of textual material on the web necessitates further research and advancement in text classification. Text classification pertains to the automated identification of text categories according to the content within a specified categorization framework [1]. It is essential in numerous applications, such as sentiment analysis, spam detection, and subject categorization. A chi-square test evaluates the observed proportions in a study against the expected proportions to determine if they differ significantly[2]. In the context of text classification, it helps identify and select relevant features that are significantly associated with a particular class. The chi-square distribution has become a powerful and widely used tool among the various statistical methods used for text categorization. There have been previous reviews on the use of the chi-square statistic for text categorization. At the same time, there have been fewer reviews on the development of applications of chi-square tests for text classification in recent years.

Thus, the research topic of this paper is a review on the application of chi-square statistics in text classification over the past five years. The aim is to systematically summarize and analyze the specific applications of chi-square statistics in text classification in different domains and their effectiveness, especially the role of the improved chi-square feature selection method in enhancing classification

DOI: 10.54254/2755-2721/97/20241397

performance. The research process generally involves searching for keywords such as “text classification” and “chi-square statistics” on Google Scholar and browsing a large number of papers. Ultimately, the more valuable papers are summarized, analyzed, and synthesized.

2. Classification of arabic text using improved chi-square statistic

Research on text classification across diverse languages globally has persisted for an extended period. Research on text classification in English, Chinese, and Turkish has been conducted for 8 or 9 years[3-5]. Nonetheless, because to the sophisticated nature of Arabic inflectional and derivational rules, its complex grammatical structure, and its rich morphology, the volume of related research on Arabic text classification remains constrained. Over the past five years, the demand for Arabic text classification has surged, resulting in a notable acceleration of its development and significant advancements.

Bahassine increased Arabic text categorization performance using an improved chi-square feature selection strategy (ImpCHI). Six categories were created using a dataset of 5070 Arabic documents in the study. Punctuation, stop words, and non-Arabic letters were eliminated during the preprocessing stage in order to optimize and fine-tune the dataset. The authors suggested an enhanced approach to balance the feature selection for various categories, consequently improving the classification accuracy, in order to overcome the shortcomings of the conventional chi-square method in feature selection. An analysis of the data showed that the Improved Chi-square Selection Method (ImpCHI) can greatly increase the accuracy of Arabic text categorization [6]. This offers a useful technique for classifying Arabic text.

In addition, Alshaer et al further improved the improved chi-square feature selection method (ImpCHI) for Arabic text categorization based on their previous study. In the new study, the researchers used documents containing 9055 Arabic texts from a wide range of sources that were categorized into 12 categories based on their content. The researchers designed six sets of experiments to test the effects of different preprocessing and feature selection methods on the performance of each of the six classifiers (BN, NB, NBM, RF, DT, and ANNs). The experimental results show that the classifiers using the ImpCHI feature selection method outperform the normal method in terms of accuracy, and in particular, the NB classifier performs best in all tests. In addition to this, the study also analyzed the percentage improvement in the performance of each classifier using the ImpCHI method without and with preprocessing. The results show that the performance improvement of ImpCHI is more significant in the case of no preprocessing [7]. The recent work further illustrates the efficacy of the chi-squared feature selection method (ImpCHI) for Arabic text classification. Recent study on Arabic text classification has demonstrated that the Improved Feature Selection Method (ImpCHI) substantially enhances the accuracy and performance of this classification process.

3. Classification of text on social networking sites using the chi-square statistic

In recent years, with the rapid popularity of the Internet and social media, the amount of text data generated on online platforms has increased dramatically. Social media platforms have become an important channel for users to express their opinions, feelings, and attitudes. Amazon and Twitter, as one of the world's most popular online platforms, generate a large number of user comments and tweets every day. The textual data contains user comments and opinions on products, events and various topics, which are highly relevant to the study of different industries and social phenomena, and may also help to optimise different industry models, thus having a high commercial and scientific value. Therefore, how to extract useful information from these large and complex text data has become an important research topic in the field of text mining and natural language processing.

Amazon.com, as one of the largest online shopping platforms globally, boasts an extensive array of products in terms of both quantity and weight. The substantial transaction volume enhances the representation of diverse customer reviews. Falasari et al. utilized the sentiment-labeled dataset from the UCI Machine Learning repository, comprising 1,000 reviews, evenly divided into 500 positive and 500 negative evaluations. Following the use of feature weighting (TF-IDF) on the processed textual data and the acquisition of the feature weighting outcomes, the chi-square test was employed to compute the

DOI: 10.54254/2755-2721/97/20241397

chi-square value for each lexical item. Features were subsequently chosen based on the chi-square statistic. Subsequently, the dataset was partitioned into training and test subsets, Naïve Bayes classification was executed, and the classifier's accuracy was assessed using the confusion matrix. Ultimately, following the implementation of the chi-square test and TF-IDF, the accuracy improved from 82% to 83% [8]. The utilization of the chi-square test and TF-IDF for text categorization demonstrates a beneficial and efficacious effect.

In addition, in social media, many platforms represented by Twitter generate a large amount of text data posted by users on the platform every day. This is a direct manifestation of social opinion and can reflect the public's views and attitudes towards an event, policy, or product. Enterprises can understand the public's opinions in a timely manner through text classification of these text data, so as to adjust market strategies, guide public opinion and avoid public opinion crises. The experiment was divided into two main parts: one without using a specific vocabulary list (no vocabulary approach) and the other using the ISO/IEC/IEEE 24765 standard vocabulary list (vocabulary approach). In the no-vocabulary approach, a total of 20,505 features were extracted, while in the vocabulary approach, the number of features was reduced to 4,674. Then, the chi-square and mutual information values of these features were calculated to determine their relevance to the classification task. The results of the experiments showed that the features selected by the chi-square test gave the classifier an accuracy of 84% without the use of a vocabulary, compared to 77% when the features selected by mutual information were used. The chi-square test shown an accuracy of 76%, whereas the mutual information demonstrated a value of 73% when the glossary was utilized [9]. The experimental findings validate the efficacy of the chi-square test as a feature selection technique, particularly in sentiment analysis and classification tasks involving extensive text datasets. Recent research indicates that the optimization of text classification, utilizing techniques such as chi-square tests and TF-IDF, can markedly enhance the accuracy of sentiment analysis and classification in the context of extensive user comments and tweets on social media platforms.

4. Text classification using the chi-square statistic in medical research

PubMed is a biomedical literature database maintained by the U.S. National Center for Biotechnology Information, providing access to more than 37 million documents [10]. Literature in PubMed covers a wide range of biomedical fields, offering a rich source of information for researchers and clinicians worldwide. That is why automated text classification tools are particularly crucial in the face of such a vast volume of literature. P'arraga-Valle et al used 1754 preprocessed documents that underwent steps such as deactivation, word removal, and punctuation deletion to ensure the cleanliness of the data. The researchers used two feature extraction strategies: one was to extract a large number of features from the entire corpus, and the other was to limit the number of features using an ISO standard glossary. During the feature selection process, the chi-square test was used to assess the importance of lexical items to filter out features that are highly relevant to the category. The performance was evaluated by using a polynomial plain Bayesian classifier for classification with 10-fold cross-validation. The chi-square test was found to outperform the control group in terms of classification accuracy and number of features required. The chi-square test achieved a maximum accuracy of 84% without a glossary and 76% with a glossary [11]. The chi-square test is highly useful as a feature selection strategy in large-scale medical text classification tasks, enhancing the overall performance of classification models.

5. Conclusion

This review presents an overview of the applications of the chi-square statistic in text classification over the past five years, with a particular emphasis on its utilization in Arabic text, social media data, and medical research. The chi-square statistic has demonstrated significant advantages in text classification tasks, effectively improving the performance of classifiers in linguistically complex Arabic texts, social media data with a large amount of user-generated content, and the classification of medical literature. In the future, researchers may integrate other statistics alongside the chi-square statistic to enhance the precision of text classification and address the evolving requirements of this field. The limitations of the

DOI: 10.54254/2755-2721/97/20241397

research in this paper include a narrow coverage of application domains and languages, a lack of in-depth comparison with other feature selection methods, insufficient experimental validation, a lack of in-depth exploration of the mechanism for improving the algorithm, and limited innovative suggestions. Future research can focus on expanding the application of chi-square statistics in multiple domains and languages, comparing the performance differences with other feature selection methods, optimizing the algorithm to cope with high-dimensional and dynamic data environments, and improving the interpretability and practicability of the model, so as to enhance the wide application and innovation potential of chi-square statistics in text classification.

References

[1] Zhai, Y., Song, W., Liu, X., Liu, L., & Zhao, X. (2018, November). A chi-square statistics based feature selection method in text classification. In 2018 IEEE 9th International conference on software engineering and service science (ICSESS) (pp. 160-163). IEEE.

[2] Onchiri, S. (2013). Conceptual model on application of chi-square test in education and social sciences. Educational Research and Reviews, 8(15), 1231.

[3] Barigou, F. (2016). Improving K-nearest neighbor efficiency for text categorization. Neural Network World, 26(1), 45.

[4] Chen, Y. W., Zhou, Q., Luo, W., & Du, J. X. (2016). Classification of Chinese texts based on recognition of semantic topics. Cognitive Computation, 8(1), 114-124.

[5] Kilimci, Z. H., Akyokus, S., & Omurca, S. I. (2016, August). The effectiveness of homogenous ensemble classifiers for Turkish and English texts. In 2016 International Symposium on INnovations in Intelligent SysTems and Applications (INISTA) (pp. 1-7). IEEE.

[6] Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. Journal of King Saud University-Computer and Information Sciences, 32(2), 225-231.

[7] Alshaer, H. N., Otair, M. A., Abualigah, L., Alshinwan, M., & Khasawneh, A. M. (2021). Feature selection method using improved CHI Square on Arabic text classifiers: analysis and application. Multimedia Tools and Applications, 80, 10373-10390.

[8] Falasari, A., & Muslim, M. A. (2022). Optimize naïve bayes classifier using chi square and term frequency inverse document frequency for amazon review sentiment analysis. Journal of Soft Computing Exploration, 3(1), 31-36.

[9] Paudel, S., Prasad, P. W. C., Alsadoon, A., Islam, M. R., & Elchouemi, A. (2019). Feature selection approach for Twitter sentiment analysis and text classification based on Chi-Square and Naïve Bayes. In International Conference on Applications and Techniques in Cyber Security and Intelligence ATCI 2018: Applications and Techniques in Cyber Security and Intelligence (pp. 281-298). Springer International Publishing.

[10] U.S. National Library of Medicine. (n.d.). PubMed. National Center for Biotechnology Information. https://pubmed.ncbi.nlm.nih.gov/

[11] Párraga-Valle, J., García-Bermúdez, R., Rojas, F., Torres-Morán, C., & Simón-Cuevas, A. (2020, April). Evaluating mutual information and chi-square metrics in text features selection process: A study case applied to the text classification in PubMed. In International Work-Conference on Bioinformatics and Biomedical Engineering (pp. 636-646). Cham: Springer International Publishing.