How Many Keywords are Enough? Determining the Optimal Top-K for Educational Website Classification

Mohd Nazrien Zaraini, Noorrezam Yusop

Fakulti Teknologi Maklumat Dan Komunikasi, Universiti Teknikal Malaysia Melaka

DOI: https://dx.doi.org/10.47772/IJRISS.2025.906000126

Received: 30 May 2025; Accepted: 03 June 2025; Published: 03 July 2025

ABSTRACT

The classification of educational websites has become increasingly challenging, as traditional indicators such as domain extensions no longer reliably reflect a site's purpose. This study investigates the optimal number of TF-IDF-ranked keywords (K) required to balance classification accuracy and computational efficiency in a one-class setting. Using a curated dataset sourced from the DMOZ directory and verified educational websites, multiple Top-K keyword subsets (K = 10–200) were evaluated. A One-Class Support Vector Machine (SVM) was employed, with performance assessed through cross-validation and separate positive/negative test sets. Results indicate that classification accuracy peaks within the range of K = 30–100, with diminished performance beyond this range due to the inclusion of irrelevant or noisy terms. These findings offer a practical and scalable framework for content-based educational website classification, particularly for applications in low-resource environments, and challenge the default reliance on exhaustive keyword feature sets.

Keywords: ranking, web classification, tf-idf, top-k, information retrieval

INTRODUCTION

With the growing variety of domain names, it has become increasingly difficult to identify educational websites based solely on their URLs. As a result, content-based classification methods have become more important. One common approach is keyword extraction, which involves identifying significant terms from a web page's content.

TF-IDF is a widely used technique in keyword extraction and helps highlight words that are important within individual documents relative to a larger collection. While it is well established in information processing, there has been limited research on how many of these top-ranked keywords are needed to accurately classify educational websites. This study focuses on finding the optimal number of keywords (Top-K) that should be used as features to achieve reliable classification.

RELATED WORKS

Educational Website Classification

The exponential growth of online educational materials has created a crucial demand for sophisticated classification methodologies to precisely identify and categorize these resources (Chen et al., 2020). Early classification systems heavily relied on structural elements, such as domain name extensions like ".edu" or ".edu.my" which were initially reliable due to the restricted usage of these domains by accredited educational institutions (Duggal et al., 2021). However, the subsequent emerging of top-level domains and the widespread

adoption of generic TLDs (e.g., .com, .org) by educational entities have significantly eroded the effectiveness of these methods, introduced more complicated classification processes (Chen et al., 2020). Many educational resources prefer to use common TLDs such as ".org," ".com," or country-specific domains, causing more sophisticated classification techniques (Li & Peng, 2021). Furthermore, previous methodologies that depended on manually curated datasets or directories like DMOZ, while providing structured taxonomies, have proven to be labour-intensive and lack the necessary scalability to keep pace with the rapidly evolving web environment. The growing volume of online data underscores the impracticality of manual filtering and the necessity for automated tools to manage and categorize digital educational content (Álvarez et al., 2023; Jiang et al., 2021). For this, the focus has shifted towards automated approaches capable of identifying educational content based on core features and contextual analysis, leveraging advancements in machine learning and natural language processing to overcome the limitations of earlier methods.

Keyword Selection and Feature Reduction in Text Classification

Keyword selection is a fundamental step in text classification, directly influencing the accuracy and efficiency of the classification process, particularly in the context of educational website analysis (Ternikov & Александрова, 2020). The Term Frequency-Inverse Document Frequency (TF-IDF) algorithm is a term-weighting scheme that plays a vital role in ranking and filtering relevant features from web documents (Tan & Song, 2021). TF-IDF is a commonly used technique to identify important terms within a document in relation to a broader corpus. (Gutiérrez et al., 2020). It balances the frequency of a term within a document with its inverse frequency across the entire corpus, effectively highlighting words that are both frequent in a specific document and relatively rare across the broader collection (Dastani et al., 2021). The number of top-ranked keywords to include, represented by the parameter K in Top-K selection, is a critical determinant of model performance, yet its optimization is frequently ad hoc or overly tailored to specific datasets ("TF–IDF," 2010). Selecting too few keywords can lead to underfitting, causing the loss of important discriminative information and reducing the model's ability to accurately classify documents. On the other hand, including too many keywords can introduce noise, increase computational complexity, and potentially reduce the model's ability to generalize. While feature reduction techniques such as mutual information, chi-square selection, and frequency thresholds have been explored in research, the specific impact of varying the K value in TF-IDF for web classification remains an area that requires further investigation (Gasparetti et al., 2017). This is particularly relevant in resource-constrained environments, where the balance between accuracy and computational efficiency is paramount, necessitating a deeper understanding of the trade-offs involved in keyword selection. Therefore, effectively tuning the K parameter within TF-IDF is essential for optimizing text classification performance, especially in specialized domains such as educational website analysis. (Wu, 2023).

The initial concept of assigning importance to words in a text by examining word frequency was introduced decades ago (Bisht, 2021). The weight of a term is typically calculated as the product of three components: a term frequency component that measures how often the term appears in the document, an inverse document frequency component that measures how rare the term is across the entire corpus, and a normalization component that adjusts for document length ("TF–IDF," 2010). The TF-IDF approach marks a significant departure from earlier methods that focused primarily on term frequency within and across documents (Nomoto, 2022). An improved method called CTF-IDF enhances the accuracy of text classification (Xu & Wu, 2014). The enhanced TF-IDF algorithm factors in the part of speech and reader's comments to improve accuracy (Guan et al., 2019). The most common formulation of TF-IDF weighting is defined by a specific equation ("TF–IDF," 2010). This equation combines local term frequency with global inverse document frequency to derive a composite weight for each term in a document (Imperial & Ong, 2021). The application of this weighting scheme has demonstrated considerable utility across various text mining tasks, enabling more accurate and effective analysis of textual data.

PROPOSED METHODOLOGY

The methodology is organized into four primary stages: data collection, pre-processing, keyword extraction and ranking, and classification with validation. A visual summary of this process is presented in Figure 1.

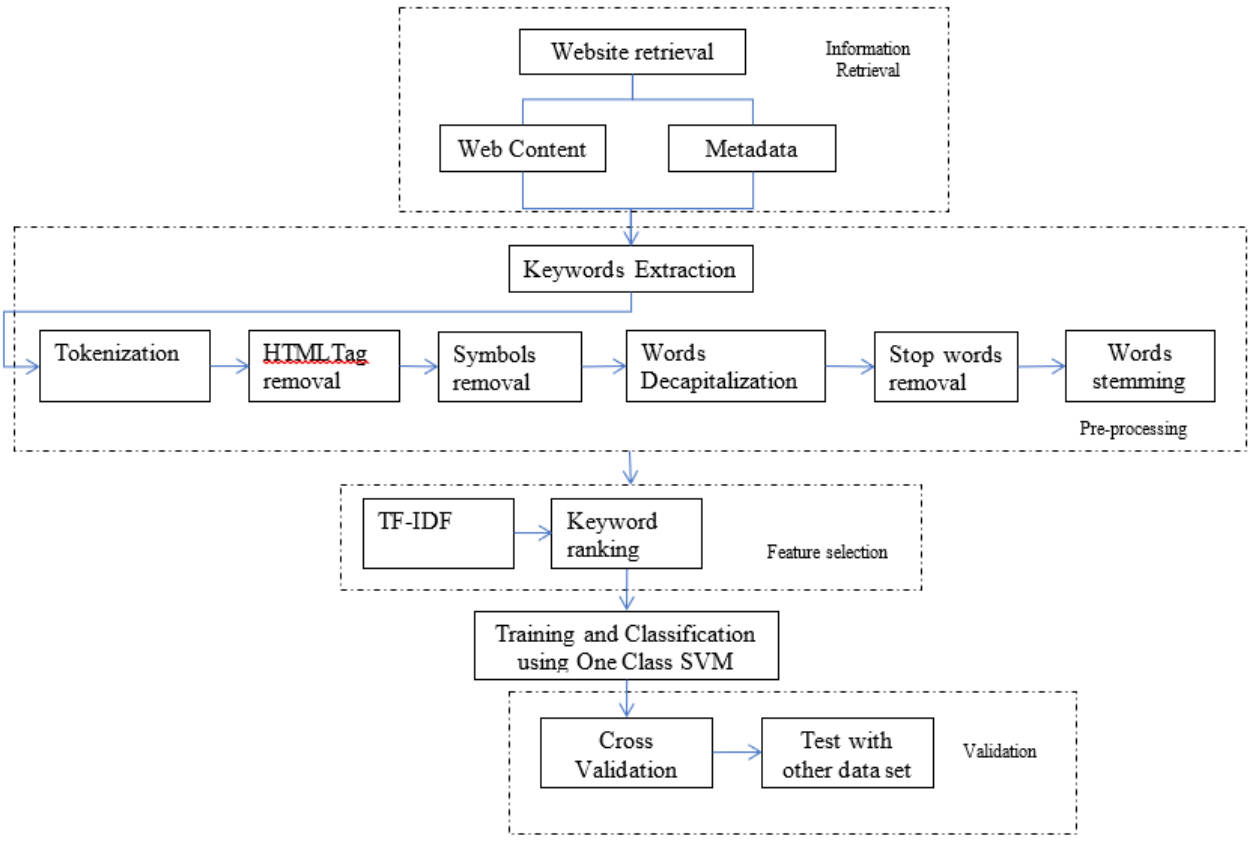

Figure 1: Methodology for extracting and ranking Top-K TF-IDF keywords for educational website classification using One-Class SVM.

Dataset Preparation

A total of 700 websites were used in this research. These were divided into training and testing datasets. For the training phase, three distinct training sets were used, each comprising 200 pre-determined educational websites randomly selected from the DMOZ directory. Although DMOZ has been archived since 2017, it remains one of the most comprehensive, structured, and openly available web directories. Its educational category provides a reliable, consistent dataset for evaluating classification strategies in a reproducible manner. The findings derived from this dataset remain relevant for content-based classification techniques, especially in low-resource or interpretable model scenarios.

- **Training Set 1**
- **Training Set 2**
- **Training Set 3**

For testing, two different sets were created:

- **Positive Test Set**: Includes 50 educational websites randomly taken from the DMOZ education category.
- **Negative Test Set**: Includes 50 websites from a mix of categories such as sports, arts, entertainment, and news, also from the DMOZ directory. This set simulates a real-world scenario where the classifier must distinguish educational content from other types.
- **Positive Test Set List**
- **Negative Test Set List**:

Data Pre-processing

There are seven activities of data pre-processing proposed for this research, which are data extraction,

tokenization, HTML tag removal, symbol removal, word decapitalization, stop words removal and words stemming.

Data Extraction

The first step in pre-processing data is the data extraction process. By using the Data Extraction operator in Rapid Miner, it extracts the textual content of a given HTML document and returns the extracted text blocks as documents. Only text blocks consisting of a given number of words are extracted to prevent single words (e.g. In navigation bars) to be kept.

Tokenization

Tokenization is a process of chopping character sequence or sentence into words, phrases, symbols, or other meaningful elements called tokens (Tamang & Bora, 2024). The objective of doing tokenization process is to explore the words in a sentence. Textual data is only a textual interpretation or block of characters at the beginning. In information retrieval require the words of the data set. Therefore, a parser used to process the tokenization of the documents. This may be trivial as the text is already stored in machine-readable formats. The main use of tokenization is the identification of meaningful keywords. Another problem are abbreviations and acronyms which need to be transformed into a standard form The list of tokens will become input for further process such as classifications. Figure 2 below shows the example of the tokenization process output.

Figure 2: Example of Tokenization Process

But there are still some problems that have been left, for example, the removal of punctuation marks as well as other characters like brackets, hyphens and such. The symbol removal process will help to solve this problem.

HTML Tag Removal

HTML or Hypertext Markup Language is very common in web pages. It is the standard markup language used to create web pages. It is written in the form of **HTML** components comprising of **tags** enclosed in angle brackets, for example <html>. Most of these tags do not have any meaning and can be removed.

Symbols Removal

Text pre-processing often includes removing punctuation, special characters, and converting text to lowercase (Ding et al., 2023; Gangar et al., 2021). Punctuation can be defined as the utilization of spacing, conventional signs, and certain typographical devices to help in the process of understanding and correct reading, both silently and aloud, of handwritten and printed texts (Brown, 2013). The example of punctuation marks includes apostrophe, bracket, colon, semi-colon and others.

Word Decapitalization

Decapitalization is a process to change the words in lower case. This is also an important process as words extracted from websites come in lower- and upper-case form. Words that are in lower case and upper case will be counted as two different elements, although it is the same words as the learners and classifiers are case sensitive. By doing the decapitalization process, all the words can be changed to lower case and will be counted as a single element.

Stop Words Removal

Classification often starts by looking at documents and finding the most important keywords in those documents.

The words that appear frequently in a document are not necessarily the important keywords. It is exactly the other way round. The most frequent words will most surely be the common words such as "the" or "and," which help build ideas and sentences but do not carry any significance meaning themselves (Fan et al., 2017). Therefore, the several hundred most common words in English (called stop words) are often removed from documents before any attempt to classify them.

Stop words are a list of common words that sometimes have a very little value in the process of machine learning and classification. A method to determine stop words is to set the terms of collection frequency (the total number of times each term appears in the document collection). MySql has provided a complete list of stop words which can be used in pre-processing data which make the process of finding stop words become much easier.

Word Stemming

Stemming is a process of converting words to its root words. The most common stemming algorithm for English words is the Porter Stemmer. It is being developed by Martin Porter at the University of Cambridge in 1980. Sometimes it just chops off the ends of words in the hope of achieving this goal correctly most of the time. But it can be very helpful to get a better classification result.

TF-IDF Keyword Extraction and Ranking

After pre-processing, Term Frequency-Inverse Document Frequency (TF-IDF) was applied to compute the importance of each term within a document relative to the entire corpus. The terms were then ranked based on their TF-IDF scores, with higher values indicating more informative keywords. To explore the impact of keyword set size, six distinct Top-K keyword sets were created, where K = 10, 20, 50, 100, 150, and 200. These sets represented varying levels of feature granularity and were used in the subsequent classification phase.

Training and Classification

This research focusses on classifying a single class website where the training data only consist of positive data and no negative data. Hence, one class classification method (**One-Class Support Vector Machine (SVM)**) must be used instead of conventional multi-class classification method. The One-Class SVM learns to recognize patterns typical of educational content and identifies deviations as non-educational.

Experimental Setup

Three separate training datasets (Training Data Set 1, 2, and 3) were used to evaluate the performance of web content classification.

Different validation techniques were applied:

- **Training Data Set 1**: Cross-validation (10 folds).
- **Training Data Set 2**: Positive Test Set and Negative Test Set evaluations.
- **Training Data Set 3**: Cross-validation and Positive Test Set evaluations.

Each dataset was tested using varying numbers of keywords (top 10, top 20, ..., top 200, and all extracted keywords).

- **Positive Test Set**: To assess how well the model identifies new educational websites.
- **Negative Test Set**: To evaluate its ability to reject non-educational content and avoid false positives.

The performance of the classification model was measured using **accuracy**, defined as the proportion of total correct predictions (true positives and true negatives) over all predictions made.

This combination of cross-validation and external testing ensures both robustness and generalizability of the classification results.

RESULT AND DISCUSSION

Data Pre-Processing

There are seven activities of data pre-processing being done in this research, which are data extraction, tokenization, HTML tag removal, symbol removal, word decapitalization, stop words removal and words stemming.

After tokenization, HTML tag removal and stop words removal process, there are more than 3000 words being extracted from the content and metadata of training and testing web sites. The words are also being changed to lowercase so that the same word will be counted as one data.

Stop words are a list of common words that sometimes have a very little value in the process of machine learning and classification. Some examples of stop words that were removed in this phase is shown in Table 4.1 below. The full list of stop words can be retrieved from MySQL websites (MySQL, 2015)

Table 4.1: Example of stop words removed

a's	able	about	above	according
accordingly	across	actually	after	afterwards
again	against	ain't	all	allow
allows	almost	alone	along	already
also	although	always	am	among
amongst	an	and	another	any
anybody	anyhow	anyone	anything	anyway
anyways	anywhere	apart	appear	appreciate
appropriate	are	aren't	around	as
aside	ask	asking	associated	at
available	away	awfully	be	became

The purpose of doing word stemming process by using Potter Stemmer is to change a word into its root word. There are five steps in Potter Stemmer, and within each step, rules are applied until one of them passes the conditions (Potter, 1980). If a rule is accepted, the suffix is removed accordingly, and the next step is performed. The resulting stem at the end of the fifth step is returned. APPENDIX A shows in detail all the steps available in the Potter Stemmer algorithm.

Although most of the words can change to its root words, but there are still several word changes into a different word instead of its root word. Table 4.2 below shows some examples of words that are not being changed into its root words.

Table 4.2: Example of words that are not changed into root words

Original words	Stem
Educational, education, educating, educator, educate	educ
University, universities	univers
Course, courses	cours

Student, students, studies, study	studi
Resource, resources	resourc
Online	onlin
Communicate, communication, communicates, communicating	commun
Day	dai
Manage, manages, manager, management	manag
Service, services	servic
Certify, certificate, certification	certif
Science	scienc

This happened because Porter Stemmer focuses on removing suffixes and prefixes (Khan & Yadav, 2019; Luo et al., 2021), and might not always identify the true root of a word. Despite its limitations, it remains a significant initial step in many Information Retrieval processes (Khan & Yadav, 2019).

Keywords Ranking

After the information about content and metadata extracted from training websites, it is found that the attributes for content, 3199 are much higher compared to the attributes of metadata data which is only 4950. If all the attributes were chosen for classification the result will become bias as the difference between both attributes are very high, more than six times fold. Therefore, keyword ranking was used based on the TF-IDF score to determine same number of attributes for both web content experiment and metadata experiment.

TF-IDF uses term frequency and times the inverse document frequency of each attribute. The Table 4.3 below shows twenty of the attributes with its TF-IDF score.

Table 4.3: k20 keyword example based on average TF-IDF

Rank	Keyword	Average_TF-IDF
1	colleg	0.033919529
2	languag	0.031320286
3	homeschool	0.029999825
4	learn	0.025583805
5	teacher	0.023995656
6	school	0.023971804
7	book	0.022452303
8	read	0.021418466
9	student	0.021374708
10	program	0.020375315
11	univers	0.020250898
12	intern	0.019222332
13	onlin	0.018737295
14	educ	0.018705899
15	disabl	0.018553035

16	curriculum	0.018108217
17	studi	0.01763062
18	test	0.017522349
19	lesson	0.017362117
20	cours	0.017133632

Training Data Set 1

Training Data Set 1 consists of 200 predetermined educational websites taken from DMOZ library. Several tests were conducted using several datasets that contain all keywords, top 10 keywords, top 20 keywords, top 30 keywords, top 40 keywords, top 50 keywords, top 100 keywords, top 150 keywords and the top 200 keywords..

Each test was validated using a cross validation technique by using 10 numbers of validation. Table 4.4 below shows the results of each test.

Table 4.4: Classification accuracy

Keywords	Classification Acuracy
k-10	89.07%
k-20	89.07%
k-30	90.05%
k-40	89.55%
k-50	88.05%
k-100	89.07%
k-150	88.52%
k-200	88.00%
All	38.71%

Based on the result shown in Table 4.4 the top 30 keywords have the best accuracy of classification based on web content with 90.05%.

Training Data Set 2

For Training Data Set 2, the experiment being done again, but using different training and testing data sets and a different method of validation. Another set of 200 pre-determined educational websites being used as training data. For the test data, 2 sets of data being used; all positive set name Positive Test Set and set with various categories of website name Negative Test Set. Table 4.5 shows the result for all positive and Table 4.6 below shows the result for various categories.

Table 4.5: Result for Positive Test Set

Keywords	Classification Acuracy
k-10	90.00%
k-20	90.00%
k-30	88.00%
k-40	90.00%

k-50	90.00%
k-100	92.00%
k-150	86.00%
k-200	84.00%
All	86.00%

For Positive Test Set, the best classification is by using top 100 keywords (k-100) as the attributes for classification based on web content which is 92% and the worst result came from top 200 for content-based classification with the value of 84%.

Table 4.6: Result for Negative Test Data

Keywords	Classification Acuracy
k-10	70.00%
k-20	72.00%
k-30	82.00%
k-40	72.00%
k-50	74.00%
k-100	70.00%
k-150	66.00%
k-200	64.00%
All	58.00%

For Negative Test Set, the best classification is by using top 30 keywords as the attributes for classification based on web content which is 82%. The worst result came from all keywords for content-based classification with the value of 58%.

The best attribute for classification comes from **the range k-30 keywords** only while classification with all keywords gives the worst classification accuracy. This because data from all keywords may consist of words that are not related with educational terms. Table 4.7 below shows the example of keywords that are ranging from 1000-1020 which are not quite related to education.

Table 4.7: Keywords range from 1000 - 1020

Rank	Keyword
1000	strand
1001	summerhai
1002	uganda
1003	upstag
1004	continun
1005	educt
1006	estat
1007	saftei
1008	camp

1009	cultur
1010	enrich
1011	fantast
1012	lyceum
1013	present
1014	red
1015	ribbon
1016	counsellor
1017	cypru
1018	dubai
1019	franc
1020	germani

Training Data Set 3

To ensure the reliability of the results, the web content classification experiment was repeated using an alternative dataset, Training Data Set 3. Two validation approaches were employed: cross-validation and evaluation using a Positive Test Set.

Table 4.8: Result obtained through cross-validation.

Keywords	Classification Acuracy
k-10	87.50%
k-20	89.50%
k-30	87.50%
k-40	86.00%
k-50	84.50%
k-100	83.50%
k-150	81.50%
k-200	81.00%
All	40.50%

Table 4.9: Result based on the Positive Test Set evaluation.

Keywords	Classification Acuracy
k-10	86%
k-20	94%
k-30	94%
k-40	98%
k-50	98%
k-100	98%
k-150	98%

k-200	98%
All	98%

For the Positive Test Set evaluation, the classification accuracy showed a steady increase with the number of keywords used. Accuracy improved from 86% with the top 10 keywords to 98% when 40 or more keywords were used, maintaining a consistent 98% accuracy even when all extracted keywords were included.

In contrast, the cross-validation results exhibited a different pattern. The highest accuracy of 89.5% was achieved when using the top 20 keywords. Beyond this point, the accuracy progressively declined as more keywords were included, reaching only 40.5% when all extracted keywords were used. This indicates that, under cross-validation, using a limited number of carefully selected keywords yields better classification performance, whereas using all keywords introduces noise and significantly degrades accuracy.

DISCUSSION OF KEY TRENDS

Optimal Keywords Threshold

Analysis across multiple datasets revealed that selecting between **30 and 100 top-ranked keywords** consistently yielded the highest classification accuracy. This range appears to capture the core thematic content of educational websites while minimizing the inclusion of irrelevant or low-signal terms. Beyond this threshold, performance began to degrade, likely due to the introduction of noisy or overly specific keywords (e.g., “lyceum,” “dubai”), as shown in Table 4.7. These findings support the principle that more features do not always translate to better performance, particularly in text classification settings.

Impact of Validation Method

The effectiveness of keyword quantity was also influenced by the choice of validation method. In **cross-validation experiments**, smaller keyword sets (K = 20–30) performed best, suggesting that leaner models are more generalizable when trained and tested across randomly partitioned folds. In contrast, evaluation using **separate test sets** favoured larger keyword sets (K = 40–100), particularly when the test data shared structural similarity with the training data. However, this also raises concerns about potential overfitting, especially when the test set is curated or limited in scope.

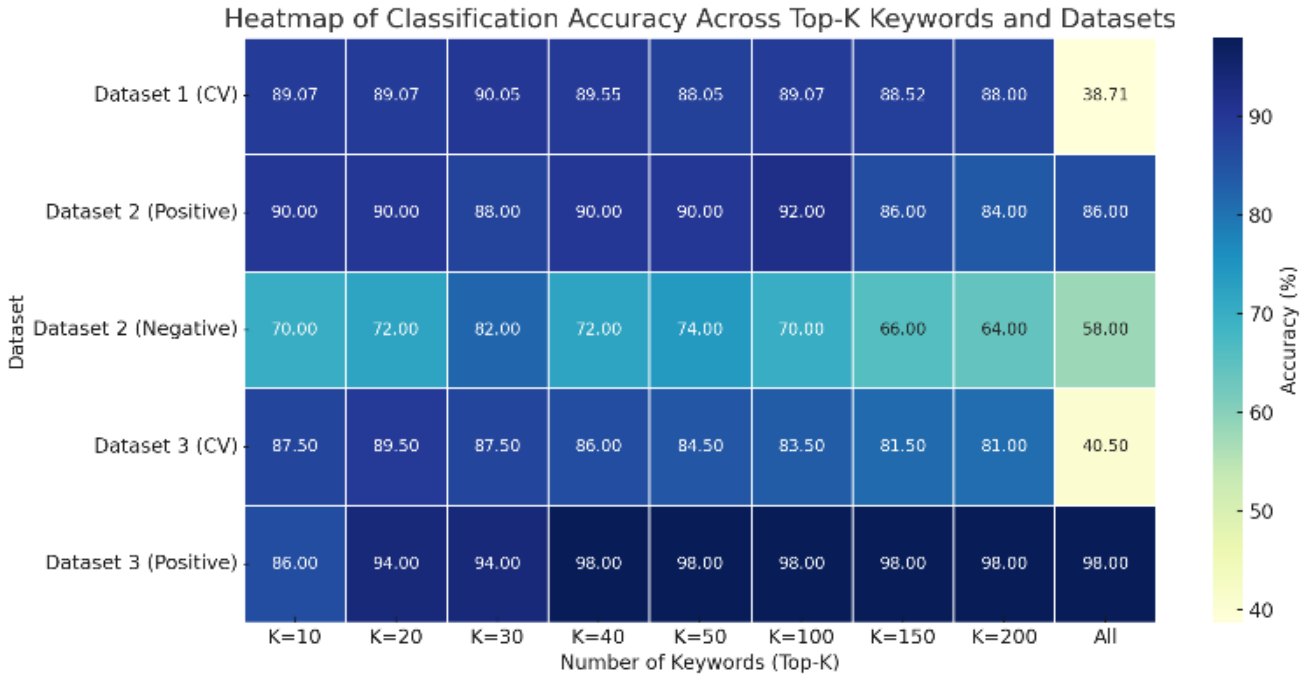

Figure 3 Heatmap Accuracy vs Number of Keywords for Each Dataset

Practical Implication

From a practical standpoint, selecting the **top 30–50 keywords** provide an optimal balance between model accuracy and computational efficiency. This range is especially suitable for real-time or resource-constrained applications, such as automated content filtering, educational search systems, or lightweight classifiers deployed in embedded environments.

CONCLUSION

This study presents a systematic evaluation of Top-K keyword selection for the classification of educational websites using a one-class learning approach. By leveraging TF-IDF to rank keyword relevance and employing a One-Class SVM model trained exclusively on positive instances, we examined how varying the number of keywords (K) impacts classification accuracy.

Our results demonstrate that selecting between **30 and 100 keywords** consistently yields high classification performance, with peak accuracy achieved around **K = 40–50**. Beyond this range, the inclusion of lower-ranked, noisy keywords introduces feature redundancy and degrades accuracy. These findings challenge the common practice of using all available keywords and underscore the importance of feature curation, particularly in resource-constrained or interpretable web classification scenarios.

The proposed approach offers a lightweight, interpretable, and effective framework for content-based classification especially valuable for educational institutions, digital libraries, and filtering systems that lack labelled negative examples. Moreover, the methodology can be extended to other domains where positive-only training data is available and computational simplicity is essential.

Future research may explore the integration of semantic keyword representations, multilingual stemmer and the possibility enhancing the Porter Stemmer capabilities to get a better and understandable root words.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support provided by University Technical Malaysia Melaka (UTeM) and the Faculty Technology Maklumat & Komunikasi (FTMK).

REFERENCES

1. Álvarez, D. P., Orozco, A. L. S., García-Miguel, J. P., & Villalba, L. J. G. (2023). Learning Strategies for Sensitive Content Detection. Electronics, 12(11), 2496. https://doi.org/10.3390/electronics12112496
2. Bisht, R. K. (2021). A Comparative Evaluation of Different Keyword Extraction Techniques. International Journal of Information Retrieval Research, 12(1), 1. https://doi.org/10.4018/ijirr.289573
3. Chen, Y., Zheng, R., Zhou, A., Liao, S., & Liu, L. (2020). Automatic Detection of Pornographic and Gambling Websites Based on Visual and Textual Content Using a Decision Mechanism. Sensors, 20(14), 3989. https://doi.org/10.3390/s20143989
4. Dastani, M., Chelak, A. M., Ziaei, S., & Delghandi, F. (2021). Identifying Emerging Trends in Scientific Texts Using TF-IDF Algorithm: A Case Study of Medical Librarianship and Information Articles. Health Technology Assessment in Action. https://doi.org/10.18502/htaa.v4i2.6231
5. Ding, K., Niu, Y., & Choo, W. C. (2023). The evolution of Airbnb research: A systematic literature review using structural topic modeling [Review of The evolution of Airbnb research: A systematic literature review using structural topic modeling]. Heliyon, 9(6). Elsevier BV. https://doi.org/10.1016/j.heliyon.2023.e17090
6. Duggal, K., Singh, P., & Gupta, L. R. (2021). Intrinsic and Extrinsic Motivation for Online Teaching in COVID-19: Applications, Issues, and Solution. In Studies in systems, decision and control (p. 327). Springer International Publishing. https://doi.org/10.1007/978-3-030-60039-6_17
7. Fan, A., Doshi-Velez, F., & Miratrix, L. (2017). Prior matters: simple and general methods for evaluating and improving topic quality in topic modeling. arXiv (Cornell University).

https://doi.org/10.48550/arXiv.1701.0322
8. Gangar, K., Ruparel, H., & Lele, S. (2021). Hindi to English: Transformer-Based Neural Machine Translation. In Lecture notes in electrical engineering (p. 337). Springer Science Business Media. https://doi.org/10.1007/978-981-33-4909-4_25
9. Gasparetti, F., Medio, C. D., Limongelli, C., Sciarrone, F., & Temperini, M. (2017). Prerequisites between learning objects: Automatic extraction based on a machine learning approach. Telematics and Informatics, 35(3), 595. https://doi.org/10.1016/j.tele.2017.05.007
10. Guan, X., Li, Y., & Gong, H. (2019). Improved TF-IDF for We Media Article Keywords Extraction. Journal of Physics Conference Series, 1302(3), 32003. https://doi.org/10.1088/1742-6596/1302/3/032003
11. Gutiérrez, L. F. G., Abri, F., Armstrong, M., Namin, A. S., & Jones, K. S. (2020). Phishing Detection through Email Embeddings. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2012.14488
12. Imperial, J. M., & Ong, E. (2021). Under the Microscope: Interpreting Readability Assessment Models for Filipino. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2110.00157
13. Jiang, Z., Bo, G., He, Y., Han, Y., Doyle, P., & Zhu, Q. (2021). Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports. Mathematical Problems in Engineering, 2021, 1. https://doi.org/10.1155/2021/6619088
14. Khan, N., & Yadav, Prof. S. R. (2019). Analysis of Text Classification Algorithms: A Review [Review of Analysis of Text Classification Algorithms: A Review]. International Journal of Trend in Scientific Research and Development, 579. Rekha Patel. https://doi.org/10.31142/ijtsrd21448
15. Li, W., & Peng, Y. (2021). An Intelligent Recommendation Strategy for Online Courses Based on Collaborative Filtering Algorithm for Educational Platforms. Frontiers in Educational Research, 4(16). https://doi.org/10.25236/fer.2021.041605
16. Luo, Y., Liang, P., Wang, C., Shahin, M., & Zhan, J. (2021). Characteristics and Challenges of Low-Code Development: The Practitioners' Perspective.
17. MySQL, 2015. MySQL :: MySQL 5.1 Reference Manual :: 12.9.4 Full-Text Stopwords. [online] Available at: https://dev.mysql.com/doc/refman/5.1/en/fulltext-stopwords.html [Accessed 24 July. 2025].
18. Porter, M. (1980). An algorithm for suffix stripping. Program Electronic Library and Information Systems, 14(3), 130. https://doi.org/10.1108/eb046814
19. Porter Stemming Algorithm. (2002). https://tartarus.org/martin/PorterStemmer/index-old.html
20. Nomoto, T. (2022). Keyword Extraction: A Modern Perspective. SN Computer Science, 4(1). https://doi.org/10.1007/s42979-022-01481-7
21. Tamang, S., & Bora, D. J. (2024). Evaluating Tokenizer Performance of Large Language Models Across Official Indian Languages. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2411.12240
22. Tan, M., & Song, Y. (2021). Research on the Topic Mining of Learners' Interest Based on the Mongolian MOOC Platform Course Discussion Text. 52, 1563. https://doi.org/10.1145/3456887.3459721
23. Ternikov, A., & Александрова, Е. (2020). Demand for skills on the labor market in the IT sector. Business Informatics, 14(2), 64. https://doi.org/10.17323/2587-814x.2020.2.64.83
24. TF–IDF. (2010). In Encyclopedia of Machine Learning (p. 986). https://doi.org/10.1007/978-0-387-30164-8_832
25. Wu, S. (2023). JobHam-place with smart recommend job options and candidate filtering options. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2303.17930
26. Xu, D., & Wu, S. B. (2014). An Improved TFIDF Algorithm in Text Classification. Applied Mechanics and Materials, 2258. https://doi.org/10.4028/www.scientific.net/amm.651-653.2258