International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

# COPY RIGHT

ELSEVIER
SSRN



IJIEMR Transactions, online available on 23rd Nov 2019. Link :http://www.ijiemr.org/downloads.php?vol=Volume-08&issue=ISSUE-11

Title A MACHINE LEARNING DOCUMENTS FOR AUTOMATIC HIGHLIGHTING

Volume 08, Issue 11, Pages: 169–173.

Paper Authors

**G.JYOTHI, DUGGEMPUDI . PRIYANKA**

St. Mary's Women's Engineering, Budampadu, Guntur-522017,AP,India


USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

# A MACHINE LEARNING DOCUMENTS FOR AUTOMATIC HIGHLIGHTING

[1]G.JYOTHI, [2]DUGGEMPUDI . PRIYANKA

[1]Assistant Professor, Dept of CSE, St. Mary's Women's Engineering, Budampadu, Guntur-522017,AP,India.

[2]Dept of CSE, St.Mary's Women's Engineering, Budampadu, Guntur-522017,AP,India.

[1]jyothi1218@gmail.com, [2]duggempudipriyanka123@gmail.com

**Abstract:** Electronic textual documents are among the most popular teaching content accessible through e-learning platforms. Teachers or learners with different levels of knowledge can access the platform and highlight portions of textual content which are deemed as particularly relevant. The highlighted documents can be shared with the learning community in support of oral lessons or individual learning. However, highlights are often incomplete or unsuitable for learners with different levels of knowledge. This paper addresses the problem of predicting new highlights of partly highlighted electronic learning documents. With the goal of enriching teaching content with additional features, text classification techniques are exploited to automatically analyze portions of documents enriched with manual highlights made by users with different levels of knowledge and to generate ad hoc prediction models. Then, the generated models are applied to the remaining content to suggest highlights. To improve the quality of the learning experience, learners may explore highlights generated by models tailored to different levels of knowledge. We tested the prediction system on real and benchmark documents highlighted by domain experts and we compared the performance of various classifiers in generating highlights. The achieved results demonstrated the high accuracy of the predictions and the applicability of the proposed approach to real teaching documents.

## 1. INTRODUCTION

Automatic Text Summarization is the process of reducing document text to highlight the essence of it and retain only the major points of the original document. There are two different approaches to create these summaries –extractive text summarization and abstractive text summarization. Extractive summaries are created by choosing sentences from the text that are necessary to capture the meaning of it while ignoring those sentences that can be removed without losing the meaning of the given text. It does not generate any new sentences. Abstractive summaries, on the other hand, seek to understand the meaning of the sentences and uses Natural Language Processing techniques to generate new sentences that capture this meaning in a shorter text. The project highlighted in this paper uses only extractive techniques since it focuses on text highlighting rather than summary generation. Text highlighting

chooses those sentences verbatim from the text that provide an overview of the entire passage. Additionally, extractive summarization techniques have shown better results than most abstractive summarization techniques since this technique does not modify the intent of the sentence, especially when the usage of a particular figure of speech is not common across languages or the way a specific language is spoken in different regions. Most abstractive systems use extractive techniques as well to improve the accuracy of the output.Work has been carried out in the field of automatic text summarization from as early as the 1950s. Early works mainly used features such as word and phrase frequency to identify salient sentences for extractive summarization. Since then, various models such as Naive Bayesian classification, neural networks etc. have been developed to generate these extractive summaries[3]. SupportVector Machines have also been proposed to identify and extract important sentences [6]. This project explores several machine learning models and their performance in extractive text summarization. We also propose a new method of generating extractive summarization datasets from human generated summaries based on the work ofNallapatti et. Al[1].ConvNets have also been used to perform text summarization [5]. Two different CNNs are constructed in this project and their accuracies compared. To evaluate the accuracy of the generated summaries, several metrics have been proposed. These include human evaluation of generated summaries and metrics that calculate the deviation of a generated summary from the standard human-created summary [4]. This project uses the ROUGE metric for evaluation which uses n-gram and longest common subsequence statistics to compare the similarity between the standard summaries and the generated summaries.2.

In this paper we address the issue of automatically generating document highlights. Highlights are graphical signs that are usually exploited to mark part of the textual content. For example, the most significant parts of the text can be underlined, colored, or circled. The importance of

text highlights in learning activities has been confirmed by previous studies on educational psychology (e.g. [3]) and visual document analysis (e.g. [4]). The highlighted documents can be easily shared between teachers and learners through e-learning platforms [2]. However, the manual generation of text highlights is time-consuming, i.e., it cannot be applied to very large document collections without a significant human effort, and prone to errors for learners who have limited knowledge on the document subject. Automating the process of text highlighting requires generating advanced analytical models able to (i) capture the underlying correlations between textual contents and (ii) scale towards large document collections. The contribution of this paper is twofold: (1) It proposes to use text classification techniques to automate the process of highlighting learning documents. (2) It considers the proficiency level of the highlighting users to drive the generation of new highlights.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

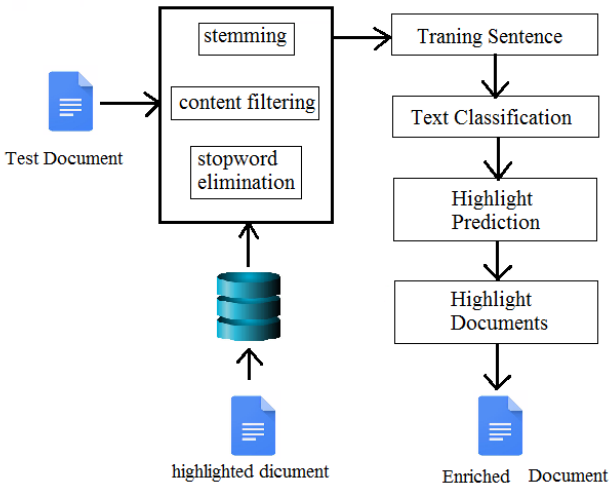
## 2. EXISTING SYSTEM

We address the issue of automatically generating document highlights. Highlights are graphical signs that are usually exploited to mark part of the textual content. For example, the most significant parts of the text can be underlined, colored, or circled. The importance of text highlights in learning activities has been confirmed by previous studies on educational psychology and visual document analysis. The highlighted documents can be easily shared between teachers and learners through e-learning platforms. However, the manual generation of text highlights is time-consuming, i.e., it cannot be applied to very large document collections without a significant human effort and prone to errors for learners who have limited knowledge on the document subject. Automating the process of text highlighting requires generating advanced analytical models able to (i) capture the underlying correlations between textual contents and (ii) scale towards large document collections.

## 3. PROPOSED SYSTEM

The manually highlighted documents are first collected into a training dataset. Some established text processing steps are then applied to prepare the raw data to the next classification process. Classification entails learning a model from the subset of document sentences that have been manually highlighted by human experts. The model is exploited to analyze new sentences of the collection and decide whether they are worth being highlighted or not based on their content and, possibly, based on the level of knowledge of the highlighting user. Finally, learners are provided with highlights corresponding to different levels of knowledge.

## 4. ARCHITECTURE




## 5. IMPLEMENTATION

### Data Representation

For each sentence of the training and test document collections we consider the following attributes: (i) the textual content, (ii) the presence of highlights, and (iii) the level of knowledge of the user who highlighted the sentence (if any). The training data consists of a set of records.

### Text Preparation

To predict highlights from learning documents, the HIGHLIGHTERsystem considers the following features: (i) theoccurrences of single terms (unigrams) in the sentence text,(ii) the occurrence of sequences of terms (n-grams), and (iii)the level of knowledge of the user who highlighted thesentence (if available). To properly handle textual featuresduring sentence classification, few basic preparation stepsare applied. First, non-textual content occurring in the textis automatically filtered out before running the

learningprocess. Then, two established text processing steps areapplied: (i) stemming and (ii) stopword elimination.

**Feature Selection**

To predict the class value of the test records, features in the training dataset may have different importance. Some of them are strongly correlated with the class and, thus, their presence is crucial to perform accurate predictions. Others are uncorrelated with the class. Hence, their presence could be harmful, in terms of both accuracy and efficiency of the classification process.

**Text Classification**

Classification is a two-step process which entails: (i) Learning a model from the training dataset, called classifier, which considers the most significant correlations between the class and the other data features, and (ii) assigning a class value to each record in the test dataset, based on the previously generated model. To investigate the use of text classification algorithms in highlight prediction, we learn multiple benchmark classifiers relying on different techniques.

**Per-Level Document Highlighting**

If in the training dataset there is no information about the level of knowledge of the users, one single classification model is generated and used to predict new highlights. Otherwise, the knowledge level of the highlighting users is considered because it is deemed as relevant to perform accurate highlight predictions.

## 6. ALGORITHM

**Wordnet Stemming AndStopwords Algorithm**: for English-written documents. To cope with documents written in different languages, different stemming and stopword elimination algorithms can be straightforwardly integrated as well. To analyze the occurrence of single terms in the sentence text, after stemming and stopword elimination the sentence text is transformed into a term frequency-inverse document frequency.

**Data Mining Algorithm:**Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database

**Clustering Algorithm:**Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields

## 7. CONCLUSION

This paper proposes highlighter, a new approach to automatically generating highlights of learning documents. It generates classification models tailored to different levels of knowledge from a set of highlighted documents to predict new highlights, which are provided to learners to improve the quality of their learning experience. A performance comparison between various classifiers on benchmark data and an analysis of the usability of the proposed approach on real document collections have been performed. In the current version of the system, highlights are not personalized. Specifically, the same highlights are deemed as appropriate for all the users having the same level of knowledge.

## 8. FUTURE WORK

We aim at tailoring the automatically generated highlights to specific users. Therefore, we would like to generate not only unified and per-level models, but also user-centric models. Furthermore, we currently ignore the presence of textual annotations, which could enrich the document content with additional notes or rephrases. We plan to analyze such automatically generated content to gain insights into the level of knowledge of learners.

## REFERENCES

[1] J. L. Moore, C. Dickson-Deane, and K. Galyen, "E-learning, online learning, and distance learning environments: Are they the same?" The Internet and Higher Education, vol. 14, no. 2, pp. 129 – 135, 2011.

[2] F. Gr¨unewald and C. Meinel, "Implementation and evaluation of digital e-lecture annotation in learning groups to foster active learning," TLT, vol. 8, no. 3, pp. 286–298, 2015.

[3] S. Elliott, Educational Psychology: Effective Teaching, Effective Learning. McGraw-Hill, 2000.

[4] A. B. Alencar, M. C. F. de Oliveira, and F. V. Paulovich, "Seeing beyond reading: A survey on visual text analytics," Wiley Int. Rev. Data Min. and Knowl. Disc., vol. 2, no. 6, pp. 476–492, Nov. 2012.

[5] C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in Mining Text Data, C. C. Aggarwal and C. Zhai, Eds. Springer US, 2012, pp. 163–222.

[6] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in ECML, 1998, pp. 4–15.

[7] J. R. Quinlan, "Induction of decision trees," Mach. Learn., vol. 1, no. 1, pp. 81–106, Mar. 1986.

[8] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology, vol. 2, pp. 27:1–27:27, 2011.

[9] W. W. Cohen, "Fast effective rule induction," in In Proceedings of the Twelfth International Conference on Machine Learning. Morgan Kaufmann, 1995, pp. 115–123.

[10] E. Baralis, S. Chiusano, and P. Garza, "A lazy approach to associative classification," IEEE TKDE, vol. 20, no. 2, pp. 156–171, 2008.

[11] Document Understanding Conference, "HTL/NAACL workshop on text summarization," 2004.