AN AUTOMATED AND EASY ANALYSING SYSTEM FOR TEXT DATA IN TEXTUAL MINING

Uday Kumar N[1], Ramakrishna S[2]
[1]PG Student, Department of Computer Science, Sri Venkateshwara University Tirupati
[2]Professor, Department of Computer Science, Sri Venkateshwara University Tirupati

Abstract
Electronic literary reports are among the most well-known showing content open through e-learning stages. Educators or students with various dimensions of information can get to the stage and feature segments of literary substance which are esteemed as especially important. The featured reports can be imparted to the learning network in help of oral exercises or individual learning. Notwithstanding, features are frequently deficient or unsatisfactory for students with various dimensions of learning. This paper tends to the issue of anticipating new features of halfway featured electronic learning reports. With the objective of advancing showing content with extra highlights, content order methods are abused to consequently examine parts of records enhanced with manual features made by clients with various dimensions of information and to create specially appointed expectation models. At that point, the produced models are connected to the staying substance to propose features. To improve the nature of the learning background, students may investigate features produced by models custom-made to various dimensions of information. We tried the expectation framework on genuine and benchmark reports featured by space specialists and we looked at the execution of different classifiers in producing features. The accomplished outcomes exhibited the high precision of the forecasts and the appropriateness of the proposed way to deal with genuine instructing reports.

I. INTRODUCTION

E-Learning stages are unpredictable frameworks gone for productively supporting learning exercises with the assistance of electronic gadgets (for example workstations, tablets, cell phones). Contrasted with conventional ways to deal with learning, they disentangle the connection among educators and students [1], in light of the fact that they permit (I) imparting electronic instructing materials to numerous clients, (ii) get to video addresses and other showing content through electronic gadgets (PCs, PCs, tablets, cell phones), and (iii) trading inputs on practices, works out, or hypothetical exercises through devoted correspondence channels. The most ordinarily shared electronic instructing materials are literary reports [2]. They envelop address notes, digital books, logical articles, or specialized reports. Be that as it may, because of the consistently expanding measure of electronic records retrievable from heterogeneous sources, the manual assessment of these training materials may turn out to be for all intents and purposes unfeasible. Thus, there is a requirement for robotized examination answers for break down electronic showing content and to naturally construe possibly helpful data. In this paper we address the issue of consequently producing report features. Features are graphical signs that are normally misused to check some portion of the printed substance. For instance, the most noteworthy pieces of the content can be underlined, shaded, or circumnavigated. The significance of content features in learning exercises has been affirmed by past examinations

on instructive brain research (for example [3]) and visual archive examination (for example [4]). The featured archives can be effectively shared among educators and students through e-learning stages [2]. Be that as it may, the manual age of content features is tedious, i.e., it can't be connected to expansive record accumulations without a noteworthy human exertion, and inclined to mistakes for students who have restricted learning on the archive subject. Robotizing the procedure of content featuring requires creating progressed scientific models ready to (I) catch the hidden relationships between's printed substance and (ii) scale towards vast report accumulations. The commitment of this paper is twofold: (1) It proposes to utilize content characterization strategies to computerize the way toward featuring learning records. (2) It considers the capability dimension of the featuring clients to drive the age of new features.

Target 1 - Highlight age dependent on arrangement strategies. Given a lot of in part featured learning records we go for naturally creating new features by applying grouping procedures. Classifiers are built up information mining calculations which have discovered application in different application spaces. Their appropriateness to printed information is built up [5]. Beginning from a lot of physically featured sentences, we manufacture a theoretical model, called classifier, which joins all the notable data expected to naturally foresee whether a sentence ought to be featured or not. Our methodology is information driven and (nearly) language-autonomous, i.e., it doesn't depend on cutting edge language handling strategies. In particular, we examine the substance of recently featured reports going over a similar theme to think about the connections between's the event of terms (or arrangements of terms) in sentences and the nearness/nonattendance of features. Such relationships will be abused to anticipate new features.

Our methodology is relevant to homogenous records (i.e., archives going over a similar subject), since it depends on recurrence-based content investigations. For effortlessness, from this point forward we will accept that a sentence is featured if no less than a bit of its printed substance is featured. The augmentation of the proposed way to deal with reports featured at various granularity levels (for example at the dimensions of single words or of passages) is direct and its outcomes are talked.

To fabricate the classifier we tried numerous procedures, among which Bayesian classifiers [6], choice trees [7], Support Vector Machines [8], rule-based [9], Neural Networks [9], and acquainted classifiers [10]. To describe the sentences of the learning archives, the classifier thinks about the accompanying highlights: (I) the events of single terms (unigrams), (ii) the event of arrangements of terms (ngrams), and (iii) the dimension of information of the client who featured the sentence (if accessible). We tried our methodology on benchmark reports featured by area specialists, i.e., the Document Understanding Conference 2005 SCU-stamped records [11]. In particular, we looked at the execution of different classifiers in producing highlights. The classifiers accomplished great exactness esteems in anticipating features.

Target 2 - Highlight age driven by the learning dimension of the featuring clients. The dependability and ease of use of content features firmly rely upon the dimension of skill of the featuring clients [12]. For instance, because of their capability on the secured subject, master clients can deliver more dependable features than fledglings.

Be that as it may, at times, the features made by clients with lower dimensions of information can be valuable for supporting learning exercises also. For instance, they may cover foundation learning regularly dismissed by cutting edge per users.

Learning stages regularly enable clients to determine their present information level on explicit subjects. At times, this data isn't expressly accessible, however it tends to be either deduced from the client job (for example scholastic educator, understudy of a B.Sc. University-level course) or surveyed utilizing specially appointed assessment procedures (for example [13]).

Our point is to abuse the data about the dimension of information of the featuring clients amid feature age and investigation. Since clients with a similar information level are destined to feature similar pieces of the content [12], we learn one grouping model for each dimension. Each model catches the basic relationships covered up in the content featured by clients with a similar dimension. Consequently, per-level models produce features custom-made to various dimensions of information. To improve the nature of the learning background, students may play out a for each dimension investigation of the recently produced features by adjusting the dimension of investigation to their necessities.

II RELATED WORK

Some efforts to automatically generate highlights of generic documents have already been made. For example, in [14]–[16] information highlighting facilities have been proposed to assist users in evaluating relevance of accessed documents. The accessed documents are identified by a search engine in response to a user query. The parts of the text that are deemed as worth highlighting are identified by matching salient keywords in contextual vocabularies. In [17] the authors addressed the complementary issue of automatically recording the marks applied to paper documents on their electronic originals. In this paper, highlight generation is data-driven and not driven by user-generated queries. This approach, unlike keyword driven ones, does not require any a priori knowledge on the learner's interests and is applicable to a broader set of users. The main contribution of this work is in the area of **learning analytics**, which entails the measurement, collection,

analysis, and reporting of data about learners and their contexts [18]. It combines different disciplines such as computer science, statistics, psychology, and pedagogy. A prominent branch of research, called educational data mining, concerns the application of data mining techniques to data generated from educational settings (e.g. universities) [19]. Learning analytics tools have different goals, among which (a) the

analysis and prediction of students' performance (e.g. [20], [21]), (b) the improvement of the quality of the learning experience by offering personalized and/or subject-wise services (e.g. [22], [23]), and (c) the extraction of salient content from large teaching data and its exploitation through online or mobile platforms (e.g. [24], [25]). The system proposed in this paper falls into category (c). The tool proposed in [25] focuses on automatically answering to learners' questions by applying text summarization techniques, while in [24] summaries of textual documents are generated to improve the accessibility of the learning materials through mobile devices. Unlike [24], [25], the approach proposed in this paper is not query-driven and relies on text classification techniques rather than on summarization algorithms.

Text classification aims at defining an abstract model of a set of classes, called classifier, which is built from a set of labeled textual data, i.e., the training set. The classifier is then used to

appropriately classify new textual data for which the class label is unknown. In our context, the training
set consists of a set of document sentences manually labeled as highlighted or non-highlighted by teachers or learners with
different levels of knowledge. The prediction task focuses on deciding whether a sentence belonging to a non-highlighted (portion of) document is worth being highlighted or not.

Many text classifiers have been proposed in literature. Amongst others, Support Vector Machines (SVMs) (e.g. [8]) and Neural Networks (NNs) (e.g. [26]) are commonly the mostly used classification models, because they are able to perform fairly accurate predictions. Alternative solutions
include Bayesian algorithms (e.g. [27] and Decision trees (e.g. [28]). A survey of text classification techniques is given in [5]. Some attempts to use existing classification algorithms in learning analytics have already been performed.

Text summarization entails generating a concise summary of a collection of textual documents. Sentence-based summarizers are automated tools that generate a summary consisting of a selection of the most significant document sentences in the collection. Many summarization approaches have been proposed in literature. Depending on the strategy used to perform sentence selection, they can be classified as (i) Clustering-based approaches (e.g., [30]), if they exploit clustering algorithms to group similar sentences and then pick the most significant sentences within each group. (ii) Graph-based approaches (e.g., [31]), if they
rely on graph indexing algorithms. (iii) Optimization-based strategies, if they exploit Singular Value Decomposition [32] or Integer Linear Programming [33], or similar strategies to select salient document sentences. (iv) Itemset-based approaches (e.g., [34]), if they exploit frequent item sets, which represent sets of document terms of arbitrary length, to capture the underlying correlations among multiple terms. While the classification problem addressed by this paper is a prediction task based on past humanly generated predictions (i.e., the set of previously highlighted sentences), in the summarization problem the goal is to describe most salient document features without any a priori information. An experimental comparison between text summarizers and classification techniques in the context

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.

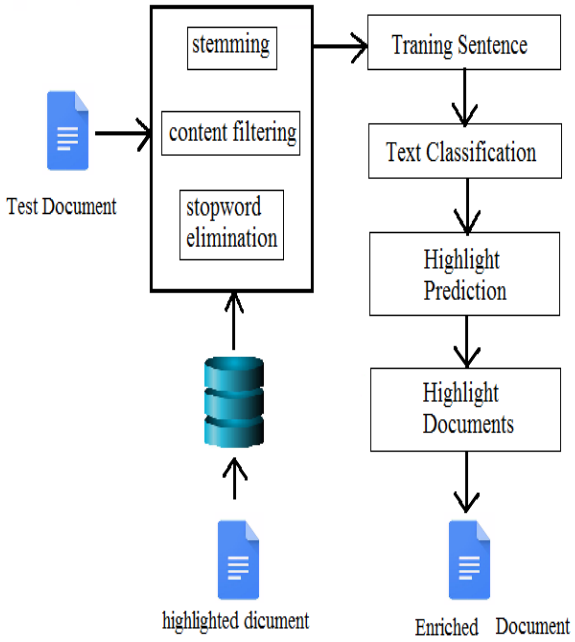
III PROPOSED SYSTEM

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.

IV METHODOLOGY

Architecture:



1. 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.

2. Text Preparation

To predict highlights from learning documents, the HIGHLIGHTER system considers the following features: (i) the occurrences 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 the sentence (if available). To properly handle textual features during sentence classification, few basic preparation steps are applied. First, non-textual content occurring in the text is automatically filtered out before running the learning process. Then, two established text processing steps are applied: (i) stemming and (ii) stop word elimination.

3. 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.

4. 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.

5. 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.

Algorithm

1. Wordnet Stemming and Stop words Algorithm: for English-written documents. To cope with documents written in different languages, different stemming and stop word elimination algorithms can be straightforwardly integrated as well. To analyze the occurrence of single terms in the sentence text, after stemming and stop word elimination the sentence text is transformed into a term frequency-inverse document frequency.
2. 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
3. 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

V 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.

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.

VI 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.

[12] Y. H. Lee, G. D. Chen, L. Y. Li, Nurkhamid, C. Y. Fan, and K. H. Chiang, "The effect of utilizing the learning skill of highlighting and constructing a map in a networked hyperlink condition on learning performance," in 2012 IEEE 12th International Conference on Advanced Learning Technologies, July 2012, pp. 546–548.

[13] D. Suthers and K. Verbert, "Learning analytics as a" middle space"," in Proceedings of the Third International Conference on Learning Analytics and Knowledge, ser. LAK '13. New York, NY, USA: ACM, 2013, pp. 1–4.

[14] N. Milic-Frayling and R. Sommerer, "Facility for highlighting documents accessed through search or browsing," Feb. 9 2010, uS Patent 7,660,813.

[15] A. Patel and D. desJardins, "Highlighting occurrences of terms in documents or search results," Dec. 14 2010, uS Patent 7,853,586.

[16] B. Cragun and P. Day, "Apparatus and method for automatically highlighting text in an electronic document," Mar. 20 2007, uS Patent 7,194,693.

[17] J. J. Hull and D.-S. Lee, "Simultaneous highlighting of paper and electronic documents," in Pattern Recognition, 2000. Proceedings. 15th International Conference on, vol. 4, 2000, pp. 401–404 vol.4. [18] R. Ferguson, "Learning analytics: Drivers, developments and challenges," Int. J. Technol. Enhanc. Learn., vol. 4, no. 5/6, pp. 304–317, Jan. 2012.

[19] G. Siemens, "Learning analytics: Envisioning a research discipline and a domain of practice," in Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge, ser. LAK '12. New York, NY, USA: ACM, 2012, pp. 4–8.

[20] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," Procedia Computer Science, vol. 72, pp. 414 – 422, 2015.

[21] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek, "Improving retention: Predicting at-risk students by analysing clicking behavior in a virtual learning environment," in Proceedings of the Third International Conference on Learning Analytics and Knowledge, ser. LAK '13. New York, NY, USA: ACM, 2013, pp. 145–149.

[22] X. Zhou, B. Wu, and Q. Jin, "Open learning platform based on personal and social analytics for individualized learning support,"

in 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing, Aug 2015, pp. 1741–1745.

[23] A. Ezen-Can, K. E. Boyer, S. Kellogg, and S. Booth, "Unsupervised modeling for understanding mooc discussion forums: A learning analytics approach," in Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, ser. LAK '15. New York, NY, USA: ACM, 2015, pp. 146–150.

[24] G. Yang, D. Wen, Kinshuk, N.-S. Chen, and E. Sutinen, "Personalized text content summarizer for mobile learning: An automatic text summarization system with relevance-based language model," in Technology for Education (T4E), 2012 IEEE Fourth International Conference on, July 2012, pp. 90–97.

[25] S. Saraswathi, M. Hemamalini, S. Janani, and V. Priyadharshini, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2011, vol. 193.

[26] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85 – 117, 2015.

[27] P. Frasconi, G. Soda, and A. Vullo, "Text categorization for multipage documents: A hybrid naive bayes hmm approach," in Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, ser. JCDL '01. New York, NY, USA: ACM, 2001, pp. 11–20.

[28] C. Apt´e, F. Damerau, and S. M. Weiss, "Automated learning of decision rules for text categorization," ACM Trans. Inf. Syst., vol. 12, no. 3, pp. 233–251, Jul. 1994.

UDAY KUMAR he is a master of Computer Science (M.Sc) pursuing in Sri Venkateswara University, Tirupati, A.P. He received Degree of Bachelor of Science in 2017 from Rayalaseema University, Kurnool. His research interests are Cloud Computing, and Machine Learning.