International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 11, November- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

AUTOMATED DOCUMENT POCESSING USING MACHINE LEARNING TECHNIQUES IN BUSINESS SECTOR

Sudip Kumar[1], Dr. Pushpendra Sarao[2]
Department of Computer Science
[1,2]Shri Venkateshwara University, Gajraula (Uttar Pradesh)

ABSTRACT

The creation of an automatic document processing system is a problem from an organizational standpoint. The processing of information involves the participation of employees at every stage. Because of the ongoing evolution of companies' activities, constant updating is required. A method of processing text that is sustained through automation requires upkeep. The primary purpose of this investigation is to carry out research on the application of machine learning techniques in the business sector, specifically with the intention of automating document processing. A machine learning classification model that makes use of SVM and Word2vec has been designed for the aim of categorizing business documents. This model was developed for the goal of classifying business documents. During the course of the trial, the proposed model was able to get a score of 0.872 on the Marco F1 scale.

Keywords: *Automated Document Processing; Machine Learning; SVM; Word2vec.*

INTRODUCTION

The purpose of software documents is to provide data pertaining to the software structure which it refers. This purpose can be characterized as "software documentation." The information included inside it is frequently vital to the efficient and successful use, maintenance, and development of a system. manuals are created at every stage of production, including design (requirements), development, and even after the product is complete in the form of user manuals. Documentation is written in a variety of formats (Aghajani et al., 2019; Kuziemski and Misuraca, 2020). The creation and distribution of documentation are essential components of any software system as well as software engineering. The reality of it and how thoroughly it covers the topic both contribute to its quality (Syed et al., 2020). In actual practice, however, there is not enough time or resources available to write the relevant documentation. This is due of the dynamic nature of software projects and the varied requirements placed on their documentation, which together create an environment that is particularly difficult to work in (Wamba-Taguimdje et al., 2020; Davis et al., 2019).

Figure 1.1: Documentation in Business[1]

[1]https://blog.prototypr.io/software-documentation-types-and-best-practices-1726ca595c7f?gi=140ec6f45611

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 11, November- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

The previous literatures that are relevant to this topic are discussed in further detail in the next section.

LITERATURE REVIEW

AUTHORS AND YEAR	METHODOLOGY	FINDINGS
Rohaime et al., (2021)	In this study, it was recommended that partially structured information be extracted from images of bills using OCR (optical character recognition) and computer science (AI). The automated robotics (RPA) bot technology was created as robotics to help with the data entry process in the company's system.	The results of the system indicated that the process can be finished in fewer than thirty seconds while maintaining an accuracy rate of one hundred percent.
Burger et al. (2023); Brown et al., (2020)	This article presented a realistic case study of systematic literature reviews, often known as SLRs, in order to serve as a guide for the application of AI during the process.	Both the benefits and drawbacks of artificial intelligence (AI) in its current state are discussed by the writers, both of which should be considered in any AI-related research.
Garrel & Jahn (2023)	The approach that has been suggested improves the performance of cyber security systems by bolstering their ability to ward off intrusions.	The CS-FSM improves data privacy by 18.3%, scalability by 17.2%, risk reduction by 13.2%, data protection by 16.2%, and attack avoidance by 11.2%.

Optical character recognition (OCR) and machine learning are two technologies that allow businesses the capacity to automate a range of the operations that they now utilize, according to previous research. The firms in question might benefit from this. However, it is all too easy to fail when attempting to put artificial intelligence into a commercial context due to the complex nature of the organizational and technological components. Therefore, the primary goal of this project is to do research on machine learning techniques for automated document processing.

METHODOLOGY

The Engineering System - Multidomain Decision Matrix (ESMDM) is used to analyse the most crucial aspects of the system, and the Object Process Methodology (OPM) is used to characterize the system's components. Support vector machines, or SVMs, are used in this thesis's machine learning model as the classifier. Word2vec offers the function, which approximates the document embedding's. The organization that serves as the source of the data utilized in testing and training is known by the acronym RVL-CDIP, which stands for Ryerson Vision Lab Complex Document Information Processing. RVL-CDIP is a collection of digitized images of business documents from the 1990s cigarette industry (Devlin et al., 2018). This thesis takes use of the Tesseract OCR technology to extract text from the digitized documents

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 11, November- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

included in the collection. To assess which rule-based classifier performs better, the F1-scores of the proposed model and two alternative classifiers are contrasted.

Precision, remembering, and F-measure are used for evaluations of achievement. The proposed approach is evaluated against two widely used rule-based machine learning algorithms, RIPPER and PART, using the RVL-CDIP collection. The dataset's advertising, scholarly report, scholarly the publication's release the norm, news article, budget, payment, survey responses and job formats were taken out and used for the assessment. The percentage of correctly positive outcomes to all correctly and incorrectly successful outcomes is known as the "precision" ratio. The organization's official representation is as the following:

Precision = $TP/ (TP + FP)$ ---------- (1)

$wh\ ereTPisTruePositiveandFPisFalsePositive.$

To calculate recall, divide the number of true positives by the total number of true positives and false negatives.

Recall = $TP/ (TP + FP)$ ----------(2)

$wh\ ereTPisTurePositiveandFNisFalseNegative$

The F1-score is calculated by multiplying the precision and memory scores by two, then dividing that total by the precision and recall scores.

$F_1 = 2\ (TP) / (FP + FN + 2(TP))$ ---------- (3)

$wh\ ereTPisTurePositive, FPisFalsePositiveandFNisFalseNegative$

The A macro score for F1 is the measure used to evaluate the overall effectiveness of the predictive model because the one in question is a multi-class coder. The free total of all the categories is used to get the Marco F1 value. Due to the information set is well-balanced and all of the categories contribute an equal amount to the assessment, the Macro F1-score was chosen to represent the results of this study.

RESULTS AND DISCUSSIONS

In this study, a comparison is made between the suggested model, which makes use of Support vector algorithms (SVM) are used as the classification system, record bedding is used to indicate the highlight. IPPER and PART are among the two governed by rules predictors that are widely used. The proposed model's F1 score was over all these groups is displayed in figure 2. The model's performance is at its lowest in the category of news articles, which gets a score of 0.84. There is a possibility that the explanation is due to the fact that the news story makes extensive use of text but does not contain concentrated document embedding's.

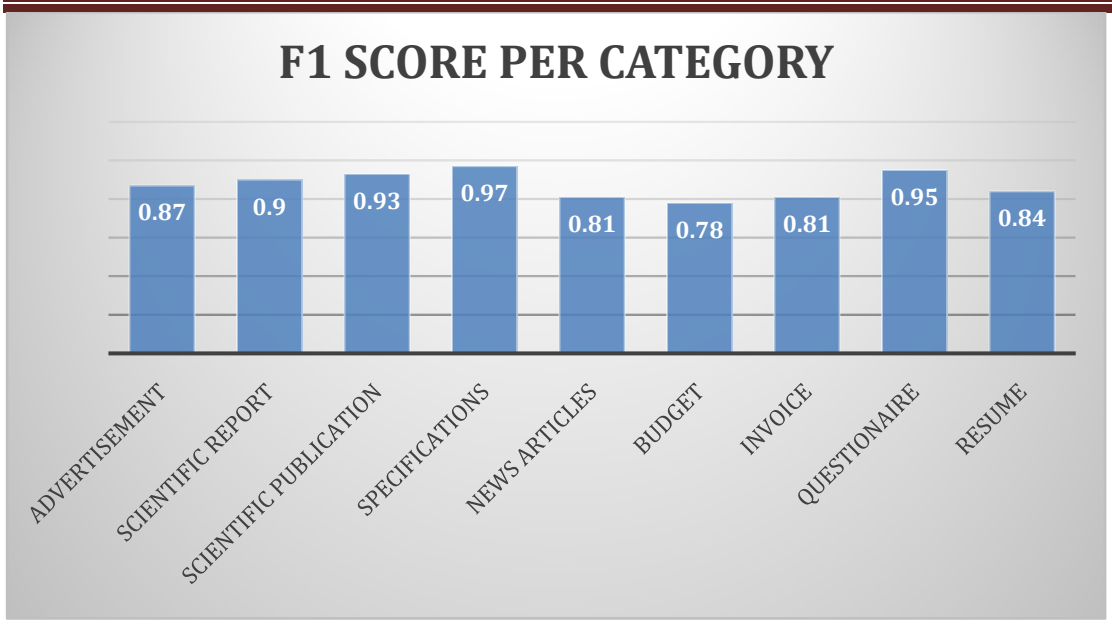

Figure 2: F1-score per Document Category for the Proposed Model.

The following graphic displays the F1-score comparing the one suggested and both of the two rules-driven frameworks. Classifiers using rules perform better than the proposed model in the resume classification category. The category of resumes has a number of different words, including "universe," which is the stem result of "university," and "biograph," which is the stem result of "biography." One of the possible reasons for this is that both words are stem results. These are two words that are frequently utilized in the drafting of resumes. The terms "universe" and "biograph" can be found as criterion in the majority of the rules.

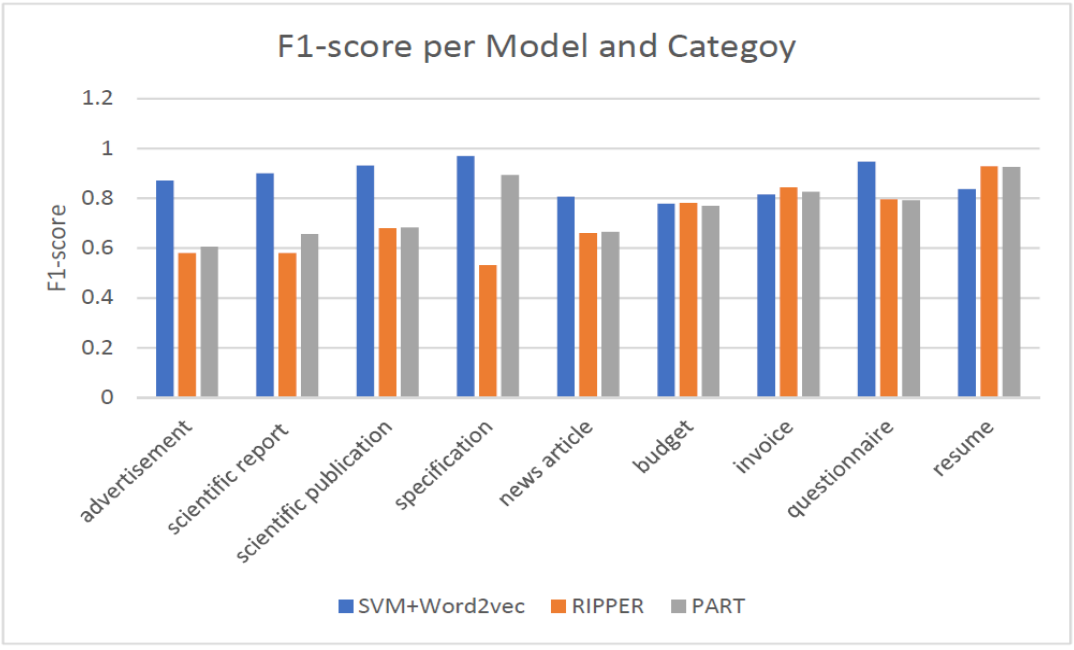

Figure 3: "F1-score Comparison between SVM, RIPPER and PART"

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 11, November- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

CONCLUSION

A machine learning classification model that makes use of SVM and Word2vec has been designed for the aim of categorizing business documents. This model was developed for the goal of classifying business documents. During the course of the trial, the proposed model was able to get a score of 0.872 on the Marco F1 scale. The other significant finding was that rule-based classifiers can also give good performance, provided that the document category in question has its own set of distinct keywords. This was an important finding because it was previously unknown that rule-based classifiers could.

REFERENCES

1. Aghajani, E. *et al.* (2019) 'Software documentation issues unveiled', *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)* [Preprint]. doi:10.1109/icse.2019.00122.
2. Wamba-Taguimdje, S.-L. *et al.* (2020) 'Influence of Artificial Intelligence (AI) on firm performance: The Business Value of AI-based transformation projects', *Business Process Management Journal*, 26(7), pp. 1893–1924. doi:10.1108/bpmj-10-2019-0411.
3. Kuziemski, M. and Misuraca, G. (2020) 'AI governance in the Public Sector: Three tales from the frontiers of Automated Decision-making in Democratic settings', *Telecommunications Policy*, 44(6), p. 101976. doi:10.1016/j.telpol.2020.101976.
4. Syed, R. *et al.* (2020) 'Robotic Process Automation: Contemporary themes and challenges', *Computers in Industry*, 115, p. 103162. doi:10.1016/j.compind.2019.103162.
5. Rohaime, N. A.et al. (2022, November). Integrated Invoicing Solution: A Robotic Process Automation with AI and OCR Approach. In *2022 IEEE 20th Student Conference on Research and Development (SCOReD)* (pp. 30-33). IEEE.
6. Burger, B.et al. (2023). On the use of AI-based tools like ChatGPT to support management research. *European Journal of Innovation Management*, *26*(7), 233-241.
7. Garrel, J. V., & Jahn, C. (2023). Design framework for the implementation of AI-based (service) business models for small and medium-sized manufacturing enterprises. *Journal of the knowledge economy*.
8. Brown, T. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.
9. Davis, B.et al. (2019, September). Deep visual template-free form parsing. In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 134-141). IEEE.
10. Devlin, J.et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*