COMSERVA:
Jurnal Penelitian dan Pengabdian Masyarakat
Volume 4 No. 1 Mei 2024 (221-230)

e-ISSN: 2798-5210
p-ISSN: 2798-5652

Enhancing Invoice Number Retrieval in Influencer Agency Transaction Records Using TF-IDF and SVC Model

Frisca Fitria
Universitas President University, Indonesia

*Email: frisca.fitria@student.president.ac.id
**Correspondence: Frisca Fitria*

DOI: 10.59141/comserva.v4i1.1339

ABSTRAK

This research explore the need for efficient invoice management in influencer marketing. The proposed solution using TF-IDF and SVC Model to effectively locate and handle crucial transaction information, including dates, brand names, invoice numbers, amounts, frequency and category. By using this approach to significantly improve the accuracy of identifying invoice numbers in influencer agency records. For the result this method are evaluated using precision accuracy and F1 Score metrics, to emphasize its effectiveness in current challenges of manual extraction. Overall, this paper offers practical solution to enhanced influencer agency transactions.

Kata kunci: TF-IDF, SVC Model, Invoice Number Retrieval, Influencer Agency, Transaction Records, Precision, Accuracy.

INTRODUCTION

Background

In modern advertising, influencer marketing has become essential, signalling a shift in consumer-brand dynamics (Yeo et al., 2023). The rising prevalence of influencer collaborations underscores the need for efficient transaction record management, vital for financial accountability and operational efficiency in influencer agencies. Given influencers' significant impact on consumer preferences and purchasing decisions, effective management of the endorsed products or brands is crucial. This research aims to enhance this process through innovative techniques, such as TF-IDF and SVC Model, addressing manual extraction challenges and offering practical solutions for streamlined influencer agency transactions.

Problem Statement

Manual extraction poses a hindrance to the seamless financial workflows in influencer marketing. To address this efficiency gap, this research advocates for the application of TF-IDF and SVC Model techniques. These advanced methods aim to enhance the accuracy of identifying invoice numbers within influencer agency records. This strategic approach aligns with the industry's escalating reliance on influencers for effective brand promotion, as highlighted by Yeo et al. in 2023. By leveraging innovative techniques, we aim to streamline and optimize the process of managing financial transactions within the dynamic landscape of influencer marketing.

Comparison with existing approach

(Manjari, Rousha, Sumanth, & Devi, 2020) revolutionize text summarization with TF-IDF and Selenium, contrasting with our proposed system's focus on accuracy in influencer agency transaction detail identification. While both utilize TF-IDF, our approach is tailored to address challenges specific to influencer agency transactions, optimizing accuracy in this domain.

(Yang & Long, 2023) enhance association rule mining with TF-IDF, whereas our system utilizes TF-IDF and SVC Model for transaction detail identification in influencer agency records. While both approaches utilize TF-IDF, our focus on accuracy within influencer agency transactions addresses different challenges and aims for specific optimizations.

(Pramono, Rohman, & Hindersah, 2013) modify TF-IDF for topic extraction, contrasting with our system's use of TF-IDF and SVC Model for transaction detail extraction in influencer agency records. While both utilize TF-IDF, our approach addresses challenges specific to influencer agency transactions, optimizing accuracy within this domain.

(Kadhim, 2019) compares TF-IDF and BM25 for feature extraction, whereas our system focuses on accuracy in influencer agency transaction detail extraction using TF-IDF and SVC Model. While both approaches utilize TF-IDF, our system addresses challenges specific to influencer agency transactions, optimizing accuracy in this domain.

(Ngan, Lee, & Khor, 2023) automate paper assignment using TF-IDF and classification, while our system enhances accuracy in influencer agency transaction detail extraction with TF-IDF and SVC Model. Though both employ TF-IDF, our approach addresses specific challenges in influencer agency transactions, optimizing accuracy for this domain.

(Zheng Huang et al., 2019) automate scanned receipt processing, whereas our proposed system aims to enhance accuracy in identifying transaction details within influencer agency records using TF-IDF and SVC Model. While both approaches focus on document analysis, our system is tailored to address specific challenges in influencer agency transactions, optimizing accuracy within this domain.

(Qiang & Zhong-min, 2022) extract symptom information from medical cases using TF-IDF and Word2Vec, contrasting with our system's focus on accuracy in influencer agency transaction detail extraction with TF-IDF and SVC Model. While both utilize TF-IDF, our approach addresses challenges specific to influencer agency transactions, optimizing accuracy within this domain.

(Nugawela, Abeywardena, & Mahaadikara, 2022) extract tabular data from images algorithmically, while our proposed system enhances accuracy in influencer agency transaction detail extraction using TF-IDF and SVC Model. While both focus on information extraction, our approach is tailored to address challenges specific to influencer agency transactions, optimizing accuracy within this domain.

(Qisong et al., 2023) automate NOTAM information extraction using NLP, whereas our system focuses on accuracy in influencer agency transaction detail extraction using TF-IDF and SVC Model. While both aim for automation, our approach addresses challenges specific to influencer agency transactions, optimizing accuracy within this domain.

Alisa et al. (2021) extract immunosuppressive properties from biomedical texts, contrasting with our system's focus on accuracy in influencer agency transaction detail extraction using TF-IDF and SVC Model. While both tackle information extraction, our approach is tailored to address challenges specific to influencer agency transactions, optimizing accuracy within this domain.

(Melnyk, Huymajer, Huemer, & Galler, 2023) propose a digital documentation system for construction management, while our system focuses on accuracy in identifying transaction details within influencer agency records using TF-IDF and SVC Model. While both aim to enhance efficiency, our approach addresses specific challenges in influencer agency transactions, optimizing accuracy within this domain.

(TongNan Huang, 2023) introduces blockchain-based transaction database encryption, contrasting with our system's focus on accuracy in influencer agency transaction detail extraction using TF-IDF and SVC Model. While both aim to enhance security and reliability, our approach addresses challenges specific to influencer agency transactions, optimizing accuracy within this domain.

(Li et al., 2020) propose a quaternary market transaction model for the electric heating market, contrasting with our system's focus on accuracy in influencer agency transaction detail extraction using TF-IDF and SVC Model. While both propose innovative transaction models, our approach addresses specific challenges in influencer agency transactions, optimizing accuracy within this domain.

Ni et al. (2023) introduce FLUID for continuous transaction processing in blockchains, while our system focuses on accuracy in identifying transaction details within influencer agency records using TF-IDF and SVC Model. While both aim to enhance efficiency in transactions, our approach addresses challenges specific to influencer agency transactions, optimizing accuracy within this domain.

(Reddy & Kumar, 2022) compare algorithms for credit card fraud detection, contrasting with our system's focus on accuracy in influencer agency transaction detail extraction using TF-IDF and SVC Model. While both aim to improve transaction security, our approach addresses specific challenges in influencer agency transactions, optimizing accuracy within this domain.

Overall, while existing research proposes innovative approaches to various transaction-related tasks, this research specifically targets accuracy in identifying transaction details within influencer agency records, utilizing TF-IDF and SVC Model tailored to this domain's challenges.

Furthermore, this research use detailed evaluation metrics, including accuracy, precision, F1 Score metrics and processing speed, provide a comprehensive assessment of the proposed solution's effectiveness.

Research aim and objectives

This research is dedicated to the rapid identification of invoice numbers within influencer agency records. Through the application of TF-IDF and SVC Model instead of Cosine Similarity, this approach prioritizes accuracy and confronts challenges linked to manual extraction. The core objective revolves around improving transaction record management, specifically within the domain of influencer marketing, by presenting practical and effective solutions. The incorporation of advanced techniques is geared towards streamlining processes and offering an efficient means of retrieving invoice numbers within the dynamic landscape of influencer agency transactions.

Limitations

This research utilizes TF-IDF and the SVC Model to improve invoice number retrieval, acknowledging certain limitations. The study predominantly evaluates the accuracy and precision of the algorithms with a dataset of 20 samples, raising concerns about the generalizability of findings to real-world scenarios. Variations in actual influencer marketing data may impact outcomes. The evaluation emphasizes efficiency through mathematical metrics, potentially overlooking practical challenges inherent in diverse influencer marketing situations. These limitations underscore the importance of a comprehensive analysis, recognizing the need for broader considerations beyond the specific context of this research.

Hypothesis

Leveraging TF-IDF and the SVC Model for invoice number retrieval is anticipated to significantly enhance accuracy and precision in contrast to manual extraction. The proposed framework aims to achieve heightened efficiency in managing influencer agency transaction records.

Outline of the Research

This study introduces a practical solution aimed at enhancing the identification of invoice numbers within influencer agency transaction records. As the demand for influencer marketing continues to surge, the need for an efficient invoice management process becomes imperative. By leveraging TF-IDF and the SVC Model, this research addresses the crucial task of accurately locating and extracting invoice numbers. Additionally, it confronts challenges associated with manual extraction, evaluating the proposed approach using precision and accuracy metrics. Through a comparative analysis with existing methods in automated information extraction and invoice management, this paper positions itself as a faster and more practical solution for identifying invoice numbers in influencer agency records. The overarching goal is to improve the efficiency and effectiveness of invoice management processes.

METHOD

The methodology employed in this research is meticulously designed to bolster the retrieval of invoice numbers in influencer agency transaction records through the synergistic utilization of TF-IDF and Support Vector Classification (SVC) Model techniques. The methodology has been refined based on comprehensive insights from the research details provided above. The outlined steps encompass data collection, pre-processing, TF-IDF calculation, SVC Model implementation, and a thorough evaluation using diverse metrics.

1) Data Collection

The initial step involves gathering a diverse dataset of influencer agency transaction records from the internal database named "Cyclone Management." The dataset encompasses essential details such as including dates, brand names, invoice numbers, amounts, frequency and category.

2) Pre-processing of Data

The collected transaction records undergo a meticulous pre-processing phase. This involves cleaning and standardizing the records to ensure consistency. The pertinent information, including dates, brand names, invoice numbers, amounts, frequency and category, is extracted from the dataset.

3) TF-IDF Calculation

a. Term Frequency

The TF for each term in the documents stored in the "Cyclone Management" database is computed using the formula:

$$TF(t) = \frac{\text{Number of times term } t \text{ appears in document}}{\text{Total number of terms in document}}$$

b. Inverse Document Frequency (IDF)

The IDF for each term across all documents in the "Cyclone Management" database is calculated using the formula.

- $IDF(t) = \log_e \left(\frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}} \right)$

c. TF-IDF Score

The TF-IDF scores for each term in the sample documents are determined using the formula:

$$TF - IDF(t, d) = TF(t, d) \times IDF(t)$$

4) SVC Model Implementation

The Support Vector Classification (SVC) Model is initialized and optimized using grid search and cross-validation to achieve the best-performing model.

a. Linear Kernel

The linear kernel is defined by the dot product of the input samples.

$$K(x_i, x_j) = x_i \cdot x_j$$

b. Radial Basis Function (RBF) Kernel

The RBF kernel, also known as the Gaussian kernel, measures the similarity between two samples based on the Euclidean distance.

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

c. Decision Function Values

$$\sum_{i=1}^{n_{\text{SV}}} (\alpha_i \cdot y_i \cdot K(x, x_i)) + b$$

In the specific context, the decision function values are calculated for the test set (Xtest) using the trained SVC model (best_svc). These values serve as a basis for predicting the class labels of the test samples. The grid search utilized during the model training optimizes hyperparameters, such as C for regularization, ensuring the SVC model effectively learns and generalizes from the training data to make accurate predictions on new, unseen samples.

5) Predictions and Ranking

a. Prediction Function

The prediction function applies the trained SVM model (best_svc) to the TF-IDF vectors of the testing data (Xtest). The SVM model utilizes its decision function, which was learned during training, to assign a predicted label to each document in the testing set. The predicted labels are stored in the vector y.

$$\hat{y} = \text{predict\_labels}(X_{\text{test}}, \text{best\_svc}) = \text{best\_svc.predict}(X_{\text{test}})$$

b. Ranking Function

Ranked_indices=

$$\text{rank\_indices}(X_{\text{test}}, \text{best\_svc}) = \text{argsort}(\text{best\_svc.decision\_function}(X_{\text{test}}))$$

The ranking function operates on the decision function values produced by the trained SVM model (best_svc) for the TF-IDF vectors of the testing data (Xtest). These decision function values represent the model's confidence or certainty about the classification of each document.

The function uses the argsort operation to sort the indices of the testing samples in ascending order based on their decision function values. The result is the vector ranked_indices where each element corresponds to the position of a document in the sorted order. The lower the index, the higher the decision function value, indicating a higher confidence in the predicted category.

6) Evaluation Metrics

a. Accuracy:

Accuracy measures the proportion of correctly predicted instances among the total instances. In the code, it is calculated by dividing the number of correct predictions (accuracy_score) by the total number of predictions.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

b. Precision:

Precision evaluates the accuracy of positive predictions. It is calculated by dividing the number of true pos.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

c. F1 Score:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1 Score is the harmonic mean of precision and recall. It considers both false positives and false negatives. The code calculates it using the precision and recall scores. The formula emphasizes the balance between precision and recall.

7) System Implementation

a. Invoice Number Retrieval

The identified documents in the "Cyclone Management" database are scanned, and the invoice numbers are extracted as potential matches for each set of sample data.

b. Integration into Transaction Record Management System:

The TF-IDF and SVM retrieval system is seamlessly integrated into the existing transaction record management system. This integration ensures compatibility through the design of standardized data formats and interfaces, facilitating smooth interaction with other modules of the system.

8) Processing Speed Calculation

The time taken for the entire process, encompassing both manual extraction and the proposed system, is recorded. The speedup percentage, a measure of efficiency improvement, is calculated using the following formula:

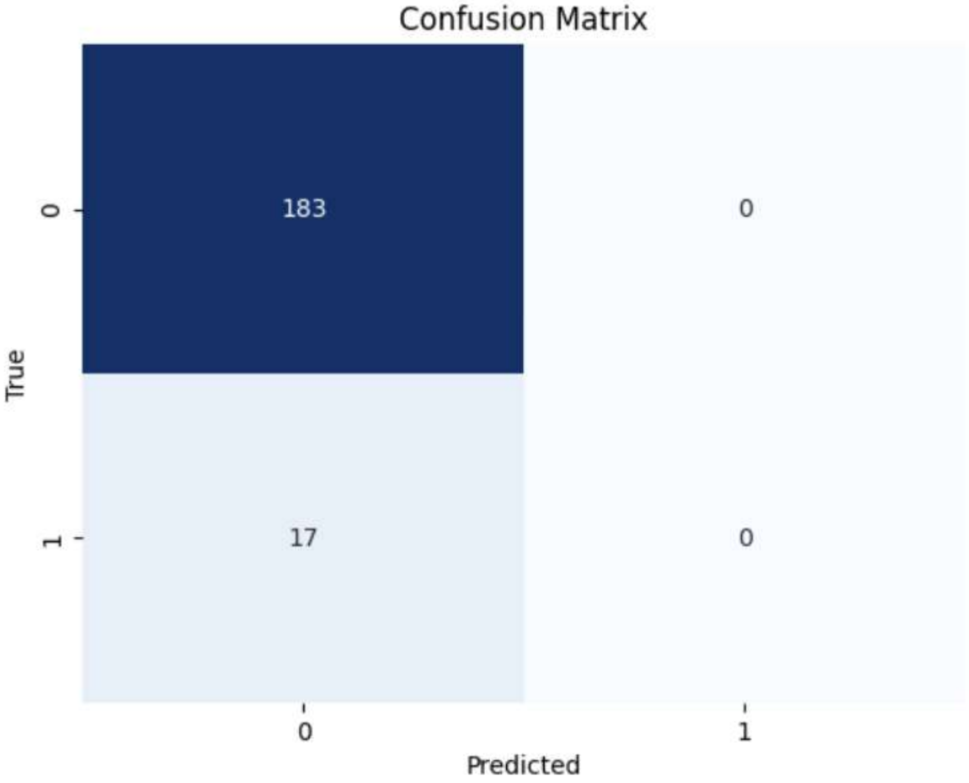
$$\text{Speedup Percentage} = \left(\frac{\text{Time taken for Manual Extraction - Time taken for Proposed System}}{\text{Time taken for Manual Extraction}}\right) \times 100$$

A higher speedup percentage indicates a more efficient and faster system, showcasing the advantages of the proposed TF-IDF and SVC Model-based approach.

RESULT AND DISCUSSIONS

The result and discussion of this research are presented values for the retrieval of invoice numbers in influencer agency transaction records highlight several key areas including evaluation metrics, brand analysis, dashboard visualization and processing speed comparison.

1. Evaluation Metrics

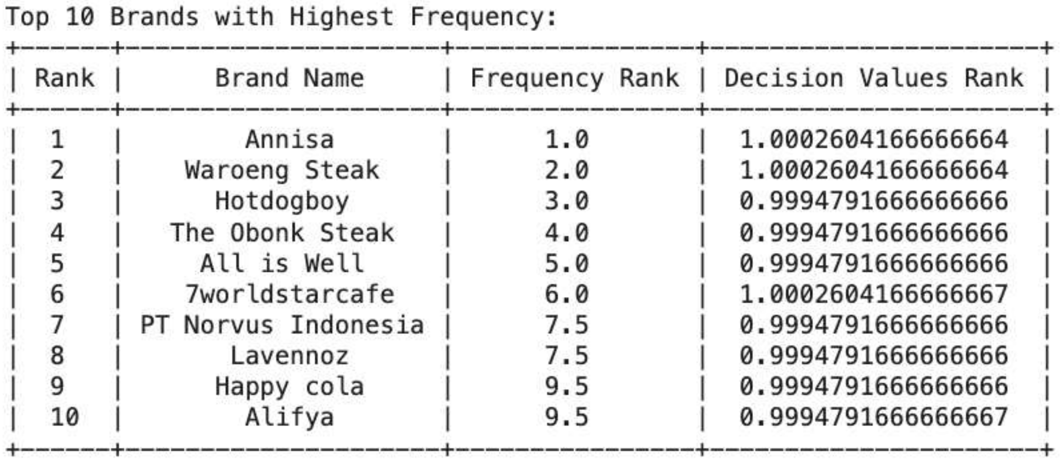


```
Accuracy: 0.915
Precision: 1.0
F1 Score: 0.0
```

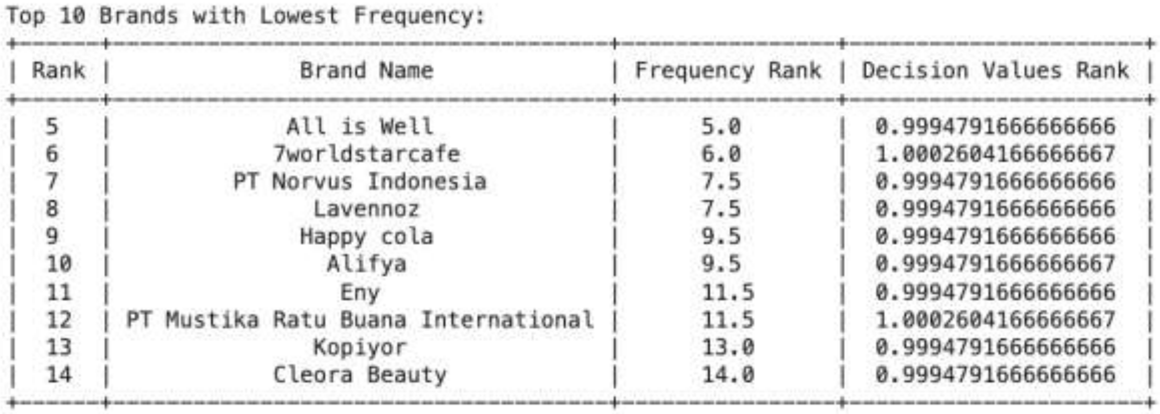
Accuracy of TF IDF on proposed methods is 91,15%, it is relatively provide high correct in predicting influencer agency transaction. Precision stands at 100% it is accurate and reduce the chances of false positive. F1 Score 0.0 the model is struggles to achieve a balance between precision and recall.

2. Brand Analysis

It shows result after implement TF-IDF and SVC Model. The data shows including rank, brand name, frequency rank and decision values rank. Consist of 2 table including top 10 brands with highest frequency and top 10 brands with 10 lowest frequency.



```
Top 10 Brands with Highest Frequency:
+------+--------------------+---------------+---------------------+
| Rank |     Brand Name     | Frequency Rank | Decision Values Rank |
+------+--------------------+---------------+---------------------+
|  1   |       Annisa       |      1.0      |  1.0002604166666664  |
|  2   |   Waroeng Steak    |      2.0      |  1.0002604166666664  |
|  3   |      Hotdogboy     |      3.0      |  0.9994791666666666  |
|  4   |   The Obonk Steak  |      4.0      |  0.9994791666666666  |
|  5   |     All is Well    |      5.0      |  0.9994791666666666  |
|  6   |    7worldstarcafe  |      6.0      |  1.0002604166666667  |
|  7   | PT Norvus Indonesia |     7.5      |  0.9994791666666666  |
|  8   |      Lavennoz      |      7.5      |  0.9994791666666666  |
|  9   |     Happy cola     |      9.5      |  0.9994791666666666  |
|  10  |       Alifya       |      9.5      |  0.9994791666666667  |
+------+--------------------+---------------+---------------------+
```

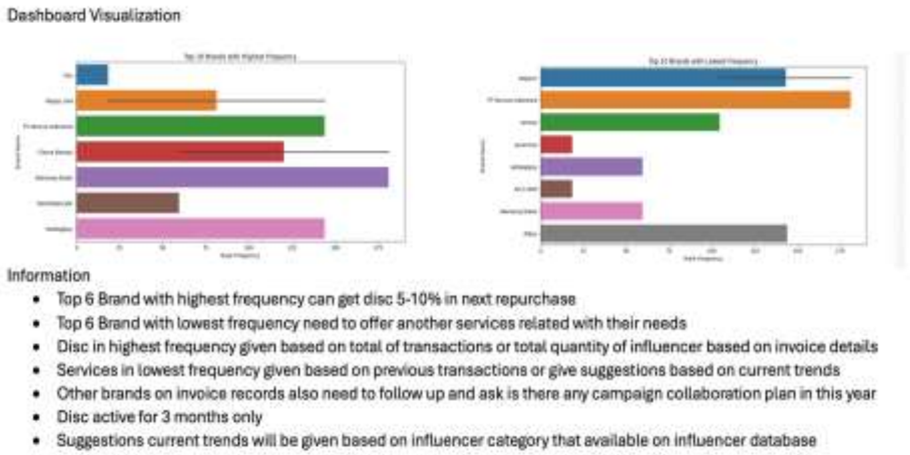


```
Top 10 Brands with Lowest Frequency:
+------+----------------------------------+----------------+----------------------+
| Rank |            Brand Name            | Frequency Rank | Decision Values Rank |
+------+----------------------------------+----------------+----------------------+
|  5   |           All is Well            |      5.0       |  0.9994791666666666  |
|  6   |          7worldstarcafe          |      6.0       |  1.0002604166666667  |
|  7   |        PT Norvus Indonesia       |      7.5       |  0.9994791666666666  |
|  8   |             Lavennoz             |      7.5       |  0.9994791666666666  |
|  9   |            Happy cola            |      9.5       |  0.9994791666666666  |
|  10  |              Alifya              |      9.5       |  0.9994791666666667  |
|  11  |               Eny                |      11.5      |  0.9994791666666666  |
|  12  | PT Mustika Ratu Buana International |   11.5      |  1.0002604166666667  |
|  13  |             Kopiyor              |      13.0      |  0.9994791666666666  |
|  14  |           Cleora Beauty          |      14.0      |  0.9994791666666666  |
+------+----------------------------------+----------------+----------------------+
```

Each table includes of relevant information in providing insight to follow up different brands within the transaction records.

3. Dashboard Visualization

This section show visualization such as top 10 brands based on frequency and information to enhance order frequency.



Dashboard Visualization

Information

- Top 6 Brand with highest frequency can get disc 5-10% in next repurchase
- Top 6 Brand with lowest frequency need to offer another services related with their needs
- Disc in highest frequency given based on total of transactions or total quantity of influencer based on invoice details
- Services in lowest frequency given based on previous transactions or give suggestions based on current trends
- Other brands on invoice records also need to follow up and ask is there any campaign collaboration plan in this year
- Disc active for 3 months only
- Suggestions current trends will be given based on influencer category that available on influencer database

Information are included discount for brand based on their frequency of transactions, discount given has term and conditions such as low and highest get differ discounts only active for 3 months, suggestions given based on conditions the available data on influencer database.

This dashboard has aims as data decision making for follow up brand based on transaction records with more targeted.

4. Processing Speed Comparison

Invoice number still saved manually it needs to compare it one another it spend 20 minutes, this proposed system only takes 0,44 minutes ~ 1 minutes.

$$\text{Speedup Percentage} = \left(\frac{\text{Time taken for Manual Extraction - Time taken for Proposed System}}{\text{Time taken for Manual Extraction}}\right) \times 100$$

Speedup Percentage: [(Time taken for Manual Extraction - Time taken for Proposed System) / Time taken for Manual Extraction] * 100

`Speedup Percentage = [(1200 − 44) / 1200] * 100 ≈ 97,81%`

Proposed methods are faster rather than manual extraction it is improve 97,81% compare with manual extraction.

CONCLUSION

The application of TF-IDF and SVC Model for invoice number retrieval in influencer agency transaction records has significant advancements in accuracy and efficiency. With 91.15% accuracy and a precision score of 100%, it has good performance in accurately identifying invoice numbers. While F1 Score has challenge in achieving a balance between precision and recall. This research provide dashboard for facilitate data visualization as solution to improves transaction record management within the influencer marketing domain especially in follow up brand for next collaboration based their data on invoice record. This integration to enhances brand analysis and order frequency insights. This solution only took 0,44 minutes, it already improved in processing speed which before has 20 minutes it decrease time taken with 97.81% compared with manual extraction. Future research may be maximize the model to achieve balance between precision and recall to enhancing the overall performance. Also exploring integration of Natural Language Processing (NLP) can be areas for further improvements. Overall, this research presents a practical and effective solution to enhance invoice management processes in influencer agency transactions, contributing to rapidly and growing of financial workflows in realm of influencer marketing.

BIBLIOGRAPHY

Huang, TongNan. (2023). Transaction Database Encryption Technology based on Blockchain Technology. *2023 8th International Conference on Information Systems Engineering (ICISE)*, 342–345. IEEE. doi: 10.1109/ICISE60366.2023.00078

Huang, Zheng, Chen, Kai, He, Jianhua, Bai, Xiang, Karatzas, Dimosthenis, Lu, Shijian, & Jawahar, C. V. (2019). Icdar2019 competition on scanned receipt ocr and information extraction. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1516–1520. IEEE. doi: 10.1109/ICDAR.2019.00244

Kadhim, Ammar Ismael. (2019). Term weighting for feature extraction on Twitter: A comparison between BM25 and TF-IDF. *2019 International Conference on Advanced Science and Engineering (ICOASE)*, 124–128. IEEE. doi: 10.1109/ICOASE.2019.8723825.

Li, Shupeng, Liu, Yingchao, Ren, Shuai, Huo, Xianxu, Yu, Jiancheng, Li, Qing, & Zhou, Ying. (2020). New Assumption of Regenerative Electric Heating Market Model from the Perspective of Transaction Cost. *2020 Asia Energy and Electrical Engineering Symposium (AEEES)*, 375–379. IEEE. doi: 10.1109/AEEES48850.2020.9121468.

Manjari, K. Usha, Rousha, Syed, Sumanth, Dasi, & Devi, J. Sirisha. (2020). Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm. *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, 648–652. IEEE. doi: 10.1109/ICOEI48184.2020.9142938

Melnyk, Oleksandr, Huymajer, Marco, Huemer, Christian, & Galler, Robert. (2023). Digitalization in the Construction Industry: The Case of Documentation and Invoicing in Tunneling. *2023 IEEE 25th Conference on Business Informatics (CBI)*, 1–10. IEEE. doi: 10.1109/CBI58679.2023.10187588

Ngan, Seon Choon Han, Lee, Ming Jie, & Khor, Kok Chin. (2023). Automating Conference Paper Assignment Using Classification Algorithms Incorporated with TF-IDF Vectorisation. *2023 IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, 1–6. IEEE. doi: 10.1109/ISIEA58478.2023.10212219

Nugawela, Methmi, Abeywardena, Kavinga Yapa, & Mahaadikara, Hansika. (2022). Algorithmically Navigating Complex Tabular Structures in Images for Information Extraction. *2022 3rd International Informatics and Software Engineering Conference (IISEC)*, 1–6. IEEE. doi: 10.1109/IISEC56263.2022.9998220

Pramono, Luthfan Hadi, Rohman, Arief Syaichu, & Hindersah, Dan Hilwadi. (2013). Modified weighting method in TF* IDF algorithm for extracting user topic based on email and social media in Integrated Digital Assistant. *2013 Joint International Conference on Rural Information & Communication Technology and Electric-Vehicle Technology (RICT & ICeV-T)*, 1–6. IEEE. doi: 10.1109/rICT-ICeVT.2013.6741547

Qiang, Cheng, & Zhong-min, Du. (2022). A NLP Application of Automated Symptom Information Extraction from TCM Medical Cases. *2022 IEEE 2nd International Conference on Computer Systems (ICCS)*, 35–39. IEEE. doi:

10.1109/ICCS56273.2022.9988199.

Qisong, Hu, Zixuan, Wu, Wen, Yin, Minrui, Chen, Guoqiang, Chen, & Yang, Yang. (2023). Research on NOTAM Information Extraction of Civil Aviation with NLP. *2023 IEEE 5th International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, 520–523. IEEE. doi: 10.1109/ICCASIT58768.2023.10351768

Reddy, P. Raghavendra, & Kumar, A. Sivanesh. (2022). Credit Card Fraudulent Transactions Prediction Using Novel Sequential Transactions by Comparing Light Gradient Booster Algorithm Over Isolation Forest Algorithm. *2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, *2*, 563–567. IEEE. doi: 10.1109/ICIPTM54933.2022.9754211.

Yang, En, & Long, Zhaohua. (2023). Research on the Weighting Method Based on Tf-IDF and Apriori Algorithm. *2023 IEEE 6th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, 1003–1005. IEEE. doi: 10.1109/ICISCAE59047.2023.10393523.