APPLICATION OF NATURAL LANGUAGE PROCESSING IN DOCUMENT VETTING

Glenda Rosy Clements[1], Shilpa Parkhi[2]

[1] Symbiosis International (Deemed University), Symbiosis Institute of Operations Management, Nashik, India

ABSTRACT

Natural Language Processing(NLP) is a theory-motivated range of computational techniques for the automatic analysis and representation of human language. NLP enables computers to perform a wide range of natural language related tasks at all levels, ranging from parsing and part of speech tagging to machine translation. The study discusses the design and application of NLP in Document Vetting processes. Among the numerous applications of NLP, this paper discusses the application of NLP in the document vetting process. The document vetting process is seen in many fields such as legal, finance, construction and others. The document vetting in a Bank Guarantee process is taken into consideration. An in-detailed process mapping of the Bank Guarantee is discussed, where in Document Vetting takes place in the process. As time progresses, organizations contain loads of documents in their repository. An algorithm for the score calculation for a document has been developed, to retrieve the best matched document from the repository with that of the input document. The NLP module is applied in this algorithm. The NLP module designed in this study is a customizable module and hence the organizations need not procure licenses of NLP software products. We suggest the use of NLP is in this process reduces the workload of employees, thereby increasing productivity and reducing cost. This paper is useful in the areas of banking and contract management.

Keywords

Natural Language Processing, Document Vetting, Tokenization, Text Classification, Sentiment Analysis

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

1. Introduction

In recent times, NLP has made great strides in making a difference in the way various business operate. Due to the rising demand in human-to-machine communications and enhanced algorithms, NLP is rapidly increasing. The NLP market is segmented, based on application (Machine translation, Sentiment analysis, Automatic Summarization, text classification, information extraction and others) and type (Text, Speech, Image)., in which the text and speech segment analysis is expected to grow rapidly. The end users of NLP could be Banking & Financial Institutions, IT and Telecom, Media and Entertainment, Education and many others. The global NLP market is expected to grow USD 41 billion by 2025, exhibiting a CAGR of 32.4 percent. An increase in the smart devices is very much likely to boost the growth of NLP. In 2018, the industry vertical, Banking and financial institutions contributed to the maximum of the global NLP market share. Banks use a branch of AI called NLP, to automate certain document processing activities and to increase the customer satisfaction level. The major applications include intelligent document search, investment analysis, developing chatbots and document vetting for various bank applications like Bank Guarantees & Letter of Credit. There are many applications that use NLP, from Google Translate to Smart Speakers.

Vetting can be defined as the process of conducting a detailed background check or fact check or a critical examination of documents related to law, property, banks, contracts, agreements and legal notices. The vetting done for the above mentioned documents is known as document vetting. The document vetting process is different for various process. In today's world, there are a numerous tools and products offering online document vetting. In this work, we shall discuss in-detail the process regarding the vetting of bank guarantee documents and the application of NLP in the process, to help calculate the score for the best matched document among thousands of documents in the repository, improves productivity.

2. NEED of the study

NLP has tended to view the process of language analysis as being decomposable into a number of, stages, mirroring the theoretical linguistic distinctions drawn between syntax, semantics and pragmatic. The stages of analysis in processing natural language are: Text pre-processing,

Tokenization, Lexical Analysis, Syntactic Analysis, Semantic Analysis, Pragmatic Analysis, and thus gives out the speaker's intended meaning [3]. There are four standard NLP tasks: Part-of-Speech tagging-tags each word with a unique tag which implies its syntactic role, Chunking-labelling segments a sentence with syntactic constituents, Named Entity Recognition-labels atomic elements in a sentence into categories and Semantic Role Labelling-gives a semantic role to a syntactic constituent of a sentence [1]. In a document vetting tool, one document is compared across thousands of documents in a repository and a sentence match, paragraph match and a keyword match is given out, to find out the best document match to the input document among the various documents in the repository. If an overall total score of the matched clauses in the document can be displayed, it would increase the efficiency. Based on the highest score of clause match of a document, the best matched document in the repository is given out. This score can be determined using NLP, as during the clause and paragraph match, there can be a number of sentiment words which need to be considered. This study investigates the application of NLP in document vetting processes would be helpful for the Banks operations team to verify the vetted Bank Guarantee Documents, which can save a lot of cost and money, thereby increasing the productivity.

3. OBJECTIVES OF THE STUDY

Two objectives were identified to guide the study:
To perform process mapping for the document vetting process for BG applications
To propose and develop an algorithm for the calculation of clause match percentage in BG Documents, using NLP.

4. METHODOLOGY

The first step in the study was to perform a detailed process mapping of the document vetting process for BG applications. A detailed description of the document vetting process of BG applications in a bank are given in the following section.

4.1. Process mapping for BG application

The entire process of issuing a BG has been mapped. The three personas which we will be considering here are the Beneficiary, Applicant and Bank. The Bank has two teams involved, the Branch and the Operations Team. We also have the document vetting tool and the Database. As the entire process mapping is large, it will be divided into three parts.

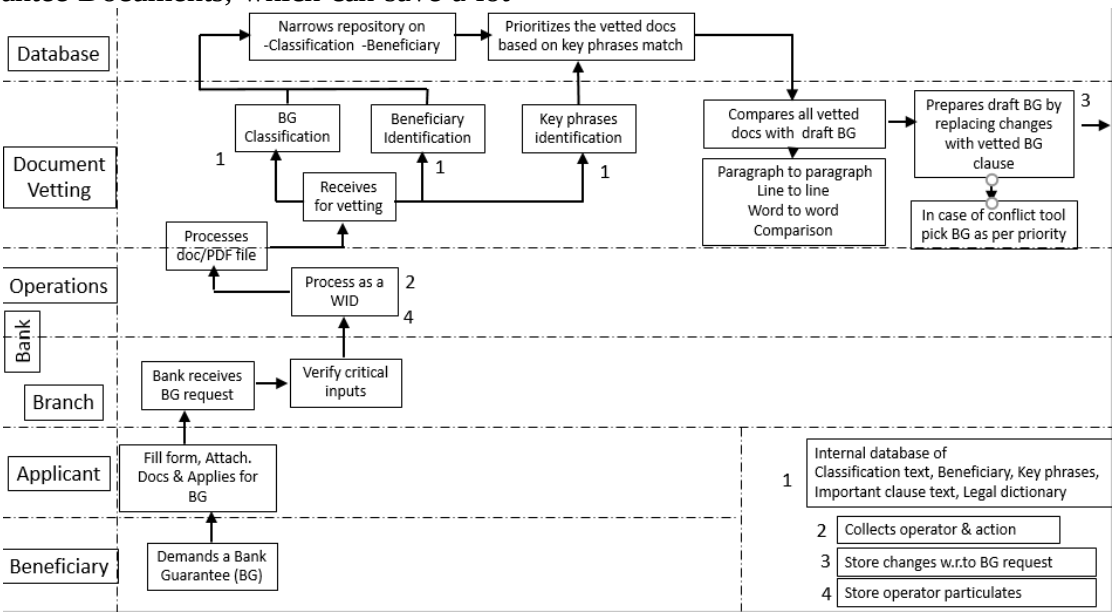

Figure 1 Part-I: Process Mapping

As seen in Figure 1, the (3) represents the database that stores changes whenever a change is made in the BG. The (2) and (4) represent the databases that store the operator's information and actions. The (1) represents the internal database of classification text, beneficiary, key phrases, important clause text, Legal Dictionary. The beneficiary demands for a BG to the applicant. The applicant fills the necessary forms, attaches the required document and applies for a BG to the

bank. The branch receives the BG request, verifies critical inputs and forwards the same to the bank Operations Team. The operation teams process it as a ticket or WID. The (2) and (4) represent the databases that store the operator's information and actions. The operations team processes the document and sends it to the Document Vetting tool for further process. The tool receives the document for vetting and with the help of the internal database, the tool classifies the BG and identifies the beneficiary, so that that the documents in the repository are narrowed down. Also, based on the key phrase identification, the Database prioritizes the vetted documents and sends it to the tool. The tool compares all the vetted documents with the input BG and prepares a draft BG by replacing the changes with the vetted BG Clause.

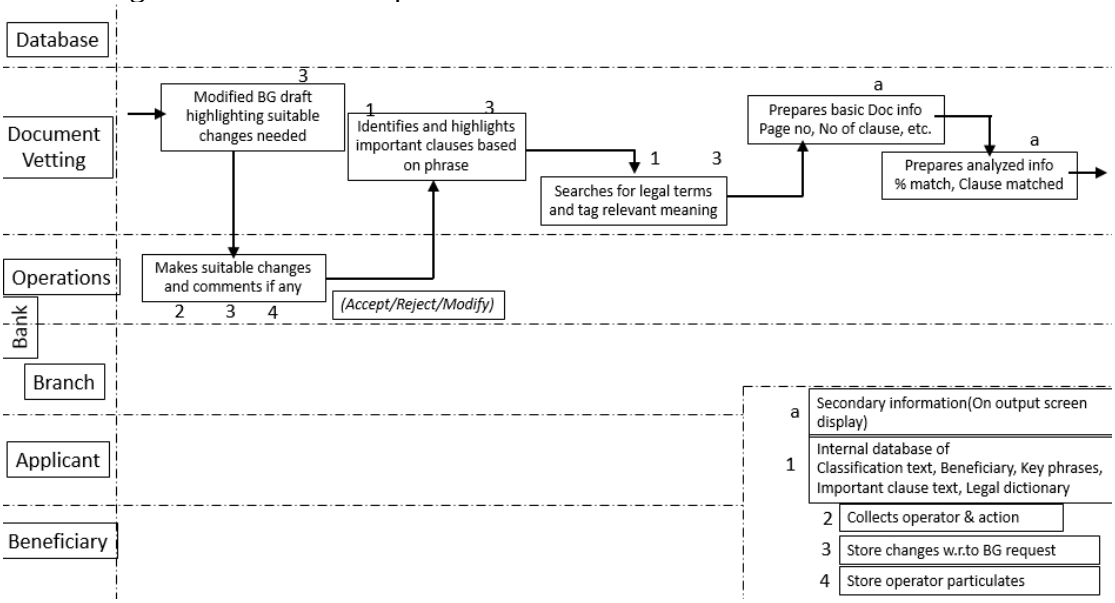

Figure 2 Part-II: Process Mapping

The figure 2 displays the second part. The modified BG Draft is generated highlighting the suitable changes which are needed. The operation team receives the same from the tool and either accepts or rejects or modifies the changes and adds his comments and processes it back to the tool. The tool identifies and highlights important clauses such as 'Auto Renewal' and then searches for legal terms from the Legal Dictionary and tags the relevant meaning. The (a) is the secondary information which would be displayed on the output screen. The tool prepares the basic doc info, no of clauses and % match of clauses and displays the same on the screen, for the operations team to view the results.

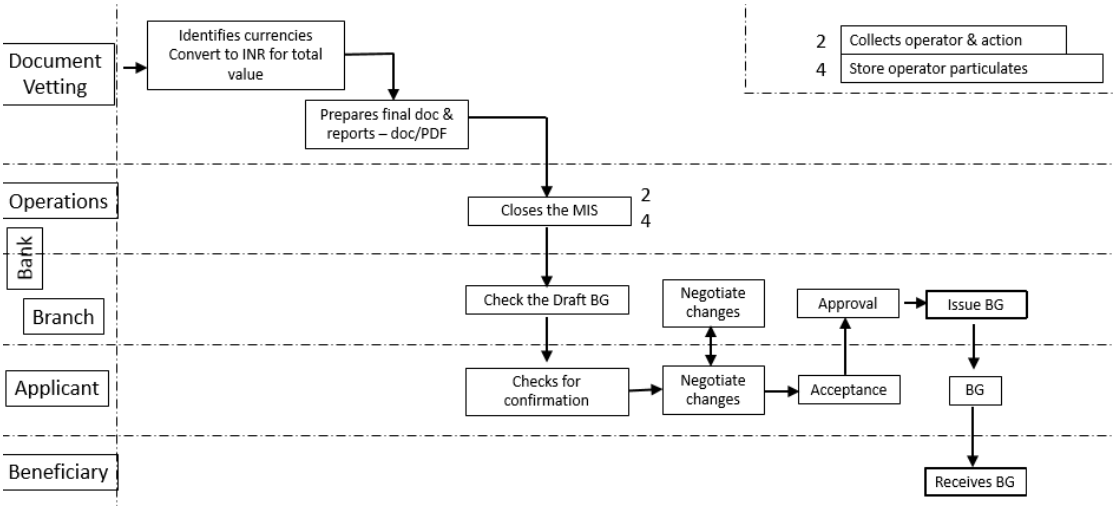

Figure 3 Part-III: Process Mapping

Figure 3 displays the third part, the tool identifies the different currencies in the BG and converts it to INR to obtain the total value of the BG. The document vetting tool prepares the final document and sends it to the orations tea. The operations

team closes the ticket and sends the vetted BG to the branch. The branch hands over the BG to the Applicant for confirmation. After negotiations between the applicant and the branch, the applicant accepts the BG and the Bank approves and issues the BG to the applicant, which is finally handed over to the beneficiary.

4.2. Score Calculation for Clause Match

As we have seen, the in detail process of issuing BG and how the process of document vetting is utilized in this. By the application of NLP in document vetting, it can be able to come up with a clause % match of the documents in the repository. By choosing he best clause % matched document among thousands of documents in the repository saves the time of the operations team, which in turn increases productivity.

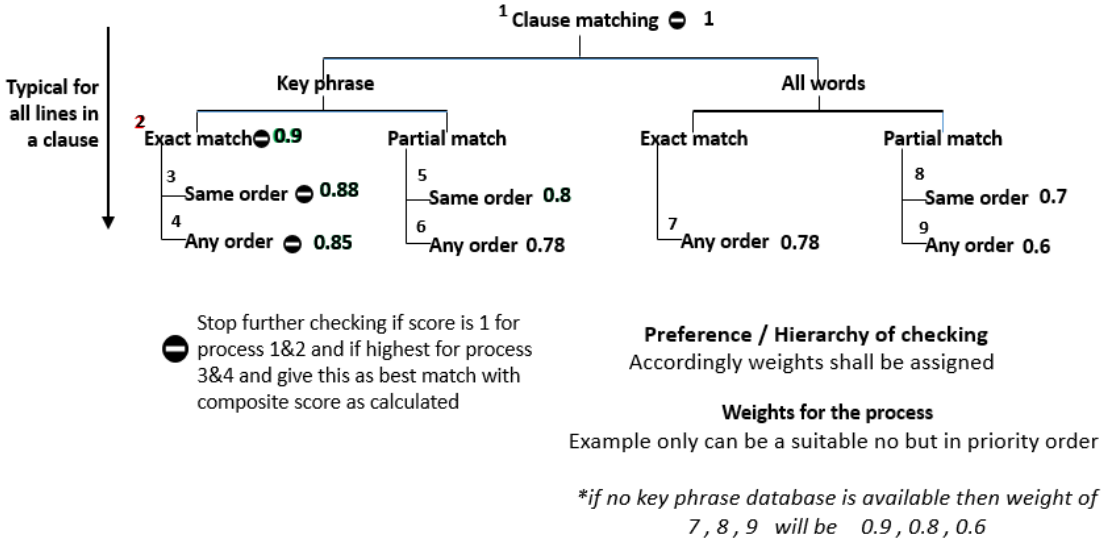

Figure 4 Hierarchy Order for Clause Match

As seen in Figure 4, for checking the clause match, the process is divided into two categories. Key Phrase Matching and All Words Matching. The numbers 1-9 represent the hierarchy the tool should be checking the clauses. The weights for each process (0.6-1) are based on the hierarchy order, these weights are assumed and can be given as per user. The black icons represent the tool to stop further process if the score calculated in process 1 & 2 which are Clause Matching and Key Phrase Exact match process is 1, as this would give the best match document. Also for Exact Match-same order and Exact Match-any order, if the score is highest the tool has to stop the further process. These 9 processes are explained in detail.

Clause Matching: A process where the entire clause in the input document is matched to the entire clause in the document present in the repository. Hence score is 1.

Key Phrase Matching:

Exact Match: A process where the particular key phrase in the input document say, “Bank Guarantee equal in value to all payments” is matched with the key words without break in order exactly in the document present in the repository. Hence score is 1.

Exact Match-Same order: A process where the particular key phrase in the input document say, “Bank Guarantee equal in value to all payments” is matched with the key words with a break, meaning there are new words between these keywords, but are in order exactly in the document present in the repository. We use the below formulas to calculate the score in this process.

Relative Number = (No. of in-between words) * No. of breaks

Score = 1/ Relative Number

When the highest score is obtained for a document among multiple documents in the repository, the tool exits the process.

Exact Match- Any order: A process where the particular key phrase in the input document say, “Bank Guarantee equal in value to all payments” is matched with the key words with a break, meaning there are new words between these keywords, but are in any order in the document

present in the repository. We use the below formulas to calculate the score in this process.
Relative Number = (No. of in-between words) * No. of breaks
Score = 1/ Relative Number
When the highest score is obtained for a document among multiple documents in the repository, the tool exits the process.
Partial Match – Same Order: A process where the particular key phrase in the input document say, "Bank Guarantee equal in value to all payments" is matched with only few key words with a break, meaning there are new words between these keywords, but are in order exactly in the document present in the repository. We use the below formulas to calculate the score in this process.
Relative number = (No of sets –Set no) *No of words in set
Where Set no = Total no. of keywords in input document- no. of missing keywords in the repository documents.
Score = Set relative no / ((n-1) * (n-2))
Where n= total no. of keywords in input document
Though the highest score is obtained in this process, the tool doesn't exit the process, but continues to the next process, as there are chances to find a better score meaning a better matched clause.
Partial Match – Any Order: A process where the particular key phrase in the input document say, "Bank Guarantee equal in value to all payments" is matched with only few key words with a break, meaning there are new words between these keywords, but are not in order in the document present in the repository. We use the below formulas to calculate the score in this process.
Relative number = (No of sets –Set no) *No of words in set
Where Set no = Total no. of keywords in input document- no. of missing keywords in the repository documents.
Score = Set relative no / ((n-1) * (n-2))
Where n= total no. of keywords in input document
Though the highest score is obtained in this process, the tool doesn't exit the process, but continues to the next process, as there are chances to find a better score meaning a better matched clause.
After this process, a similar process is followed for the remaining three process, but instead of considering only keyword, all words in the clause are considered, which are:
All words- Exact Match-Any Order
All words- Partial Match- Same Order
All words- Partial Match- Any Order
After calculating the scores in each of the process, the composite score content is calculated as follows:

$$\text{Composite score / line content} = \frac{Wt1 * Score1 + Wt2 * Score2 + Wt3 * Score3 + \dots + Wtn * Score}{Wt1 + Wt2 + Wt3 + \dots + Wtn}$$

Where Wt1, Wt2…are the assumed weights of each of the processes.

5. RESULTS AND DISCUSSION

Natural Language Processing Module

After an in detailed research regarding NLP, a NLP module that consists of total five modules, has been developed, which shall be discussed. Modules 1-4 will be used for Key Phrase Matching and modules 1-5 will be used for all words matching. Formulas to calculate the final scores to arrive at the best matched document has been developed.

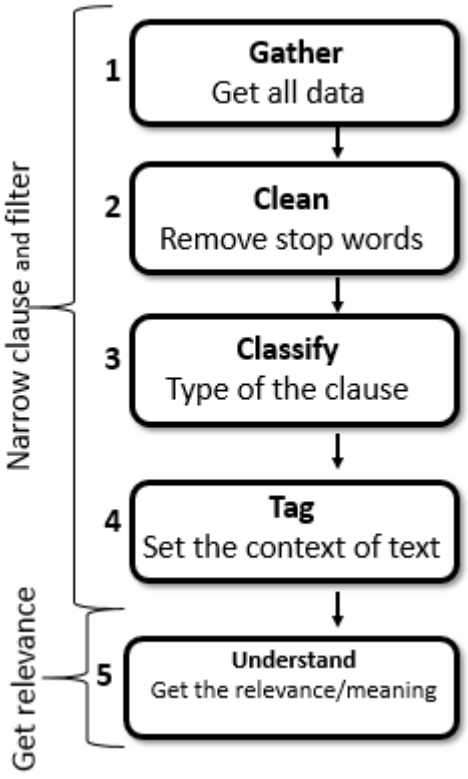

Figure 5 Modules in NLP

Figure 5 displays the five modules in NLP. As, the figure shows the modules 1-4 are used to narrow and filter the documents and the fifth module here is to get relevance or understand the meaning of

the clause and finally how the score of the clause is calculated. We shall look into each of these modules in detail with an example to understand better.

Module 1: Gather- In this Module, the system gathers all the data and tokenizing of the text is done. Tokenizing of data meaning, each word is separated as tokens. Let us consider an example.

Before Tokenizing: The Contractor is required to put up a Bank Guarantee equal in Value to all payments made before Systems Acceptance

After Tokenizing: The | Contractor | is | required | to | put | up | a| Bank| Guarantee |equal| in | Value | to | all | payments | made | before | Systems | Acceptance

Module 2: Clean- In this module, the system removes the no meaning words such as is, to, as etc.,

There is a separate database which is used here called the Stop Words Database which contains the stop words. Consider the previous example.

Before Cleaning: The | Contractor | is | required | to | put | up | a| Bank| Guarantee |equal| in| Value | to | all | payments | made | before | Systems | Acceptance

After Cleaning: The | Contractor | is | required | to | put | up | a | Bank| Guarantee |equal| in| Value | to | all | payments | made | before | Systems | Acceptance

Module 3: Classify- In the module, the system classifies the clause type based on the internal databases. This would have a separate database known as the Classification Database which would contain words like Fees, Responsible, guarantee etc., This would also have a synonyms database to identify similar meanings of the word like payments- funds, Agreement- Acceptance etc. Let us consider an example here.

Before Classification: Contractor | required | put | up| Bank| Guarantee | equal | Value | payments

After Classification: Contractor | required | put | up| Bank| Guarantee |equal| Value | payments | made | before | Systems | Acceptance

Here the clause is classified as "Guarantee" clause.

Module 4: Tag- In this module, the system sets the context of the text in the clause. Based on the semantic words, the sentence is tagged. A database known as Tag Database is used here where under tags such Stakeholders- The words bank, applicant, customer is used, Scope- limited to, liability, mistake. In this module also, the synonyms database is used. This can be explained with the help of an example.

Before Tagging: Contractor | required | put | up | Bank| Guarantee |equal | Value | payments | made | before | Systems | Acceptance

After Tagging: Contractor | required | put | up| Bank| Guarantee |equal| Value | payments | made | before | Systems | Acceptance

Here the words Contractor and Bank are tagged under Stakeholders and the words Value payments, systems, acceptance are tagged under Scope.

Module 5: Relevance – This module is the crux of the entire process. Based on the semantic words, the system gets the sentiment score using DB. This module us used to get the relevance or meaning of the clause. There is a sentiment database which contains sentiment words such as Before, after, more than equal to, less than and many other such words. These words have been segregated and given a score ranging in the scale from -1 to +1. Here a few words have been considered for example.

Before vs After – hard negative,

More than equal vs More than – soft positive

May or may not vs May not – probable negative

The score is allotted as follows:

- Hard negative -1
- Soft negative -0.5
- Probable negative -0.25
- Hard positive +1
- Soft positive +0.5
- Probable positive +0.25

The calculation of sentiment score of the clause is done in three steps:

1. Using the synonyms database and the words identified in Modules 3 & 4, the important words are tagged in the input document and the repository document
2. Using the sentiment database, the sentiment words are tagged in the input document and the repository document.
3. The sentiment words are attached to the closest important words identified in both the input document and the document in repository. Based on the shortest distance between the important word and Sentiment words the important words are related to sentiment words starting with the last important word.

Now the sentiment score is calculated, for each tag identified, the following formula is used:

Modulus (Sentiment input document – Sentiment repository)

Sentiment input document- Depending on the sentiment word whether hard negative or soft negative or probable negative or hard positive or soft positive or probable positive, the score is given to the sentiment word in the input document.

Sentiment repository- Depending on the sentiment word whether hard negative or soft negative or probable negative or hard positive or soft positive or probable positive, the score is given to the sentiment word in the repository document.

The summation of all the sentiment word scores is the total sentiment score of the clause. After calculating the sentiment scores, the composite score/line relevance is calculated by using the following formula:

$$\text{Composite score / line relevance} = \frac{\left\{ W_t * \frac{\text{No of matched tag}}{\text{Total tags}} + W_c * \frac{\text{No of matched Classif}}{\text{Total Classif}} \right\} / W_s * \sum \left| \text{(Sentiment Base} - \text{Sentiment repository)} \right|}{W_t + W_c + W_s}$$

Wt- Weight of tag

Wc- Weight of classification

Ws- Weight of Sentiment

These weights can be assumed as per the user.

Once the Composite score content and composite score relevance is calculated for each document in the repository, the priority Index is Calculated, to give out the best match of documents based on the priority achieved.

Priority index = Composite score content * Composite score relevance

It can be understood from the modules of NLP which were designed above to obtain the overall document percentage match in the document vetting process, organizations can save cost as there is no need to procure NLP licensed packages. Taken together, these findings illustrate that the application of NLP in the document vetting process can be beneficial to various other sector including legal and finance.

6. CONCLUSION

Based on the outcome of the study, the application of NLP in document vetting is recommended for use in various other sectors like legal and contract management also. Moreover, the modules and formulas developed for the calculation of the overall document percentage match in the present study highlight the increase in productivity on work performance which saves time. The detailed process mapping of the BG application and the role of document vetting in the process has made it easier to identify the application of NLP in the process. However, this finding is limited only to the document vetting process. Further research is needed to replicate the above results in various other processes.

REFERENCES

[1] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of machine learning research.

[2] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations

[3] Nitin Indhurkhya, Fred J. Damerau (2010). Handbook of Natural Language Processing. Second Edition.

[4] Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), 544-551.

[5] Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3), 55-75.

[6] Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., & Zettlemoyer, L. (2018). Allennlp: A deep semantic natural language processing platform. arXiv preprint arXiv:1803.07640.

[7] Goldberg, Y. (2016). A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research, 57, 345-420.

[8] Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266.

[9] Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational linguistics, 21(4), 543-565.

[10] Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd international conference on Knowledge capture (pp. 70-77).

[11] Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. In COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics.

[12] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (pp. 55-60).

[13] Chowdhary, K. R. (2020). Natural language processing. In Fundamentals of Artificial Intelligence (pp. 603-649). Springer, New Delhi.