EMITTER International Journal of Engineering Technology
Vol. 10, No. 1, June 2022, pp. 47~62
DOI : 10.24003/emitter.v10i1.679

Text Mining for Employee Candidates Automatic Profiling Based on Application Documents

Adhi Dharma Wibawa[1,2], Arni Muarifah Amri[2], Arbintoro Mas[2], Syahrul Iman[2]

[1]Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[2]Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
Correspondence Author : adhiosa@te.its.ac.id

Received February 2, 2022; Revised March 5, 2022; Accepted April 6, 2022

Abstract

Opening job vacancies using the Internet will receive many applications quickly. Manually filtering resumes takes a lot of time and incurs huge costs. In addition, this manual screening process tends to be inaccurate due to fatigue conditions and fails in obtaining the right candidate for the job. This paper proposed a solution to automatically generate the most suitable candidate from the application document. In this study, 126 application documents from a private company were used for the experiment. The documents consist of 41 documents for Human Resource and Development (HRD) staff, 42 documents for IT (Data Developer), and 43 documents for the Marketing position. Text Processing is implemented to extract relevant information such as skills, education, experiences from the unstructured resumes and summarize each application. A specific dictionary for each vacancy is generated based on terms used in each profession. Two methods are implemented and compared to match and score the application document, namely Document Vector and N-gram analysis. The highest the score obtained by one document, the highest the possibility of application to be accepted. The two methods' results are then validated by the real selection process by the company. The highest accuracy was achieved by the N-Gram method in IT vacancy with 87,5%, while the Document Vector showed 75% accuracy. For Marketing staff vacancy, both methods achieved the same accuracy as 78%. In HRD staff vacancy, the N-Gram method showed 68%, while Document Vector showed 74%. In conclusion, overall the N-gram method showed slightly better accuracy compared to the Document Vector method.

Keywords: Document Vector, N-Gram, Euclidean distance, Text Mining, Specific Dictionary, Candidate Automatic Profiling

1. INTRODUCTION

In the era of VUCA (Volatility, Uncertainty, Complexity, and Ambiguity) and digitalization, organizations are required to respond more quickly, collaboratively, flexibly, and adaptively to the changes in the environment [1]. This era describes a business situation that moves toward uncertainty and quickly changes to cause anxiety. Organizations that are hierarchical and bureaucratic might lose business momentum because of the delay in responding and defeating innovation. Transformation of adaptive organizations needs an agile team, and people with the best qualifications, experiences, and skills will be required. These requirements even relate to the recruitment of the company's employee candidates. HR recruitment is expected to adapt in this VUCA era to get the best team with a quick and precise selection, including online recruitment, so that information might spread widely and provide opportunities to job seekers to compete after opening a job vacancy. Through online recruitment, a candidate can also easily send a letter of application to be received directly by the company.

In this recruitment process, the essential things performed by HRD would be profiling employees before passing through the stages of the next test. The profiling of employee candidates is knowing and matching candidate profiles process with the existing requirements. One of the methods employed in this profiling is profile matching. In general, [2] the profile matching process compares individual competencies with competencies that must be possessed by prospective employees based on the requirements of each division [3]. The greater the resulting weight, the greater the chance is for such a candidate to be accepted. Profile matching is also a decision-making mechanism assuming that 'the illustrated ideal predictor variable should be below the minimal level' must be met or passed [4].

Currently, profiling in several companies is still performed by manually checking the documents one by one and then matching them with the existing requirements. However, this method seems complicated, and it takes much time. Moreover, with the massive number of cover letter documents, the checking accuracy might be decreased due to staff fatigue. Companies will find it difficult to get candidates quickly and precisely with such manual profiling. The real impact will be failing to get the exact and proper candidate for each vacancy.

This research aimed to introduce automatic profiling to get the best candidates using the text processing method concerning the constraint mentioned. In this article, two methods of text processing are applied to find the best match between the application documents with the specific criteria determined by the HRD of the company. A specific dictionary is generated as the basis for scoring the result of those two techniques, namely Document vector, and N-Gram analysis.

2. RELATED WORKS

Text mining is becoming more popular recently due to the big data environment that is expanding into many fields. Since many companies and organizations realize the importance of data and data analytics, data regarding all the processes that happen inside the organization are now becoming a treasure [5, 6, 9, 23-27]. All organizational activities now have become a primary need to be saved for future analysis. Text mining has been used by many scientists to analyze and obtain new insights from the piles of text files. Some examples of the application of text mining in the big data environment are sentiment analysis and classification, article classification, spam, and fake news detection, argument extraction, exploring the social Issues, logs mining, and search personalization, including article summarization and automatic recommendation [5]. With the development of ICT infrastructure and the massive use of the internet and social media recently, many businesses and companies have looked at that as new opportunities not only to promote their products but also to make their jobs done effectively. For example, when a company is looking for new employee candidates, social media such as Instagram, Facebook, Linkedin, Whatsapp, etc have helped them efficiently. Once a new flyer has been released by one company, within seconds it will reach all parts of the world. By looking at this example, this study is exploring text mining techniques to make an automatic candidate profiling by using application documents. Some previous studies concerning similar works have been done by [6-7, 23-27]. Some studies use local language-based text mining [6], while others use English-based text mining. For example, Kino et al [23] explored the key factors from the candidate profile dataset regarding travel time, job location, job type, and candidate skill that affect the selection process. Another study was done by [24] also performed text mining but in different analyses by developing a web application so that all candidates can be matched at the first approach before continuing into the further application process. Another different study was also done by [7]. In this article [7], the author implemented text mining on curriculum vitae (CV) documents to extract the matching skill by using the Tf-IDF technique. Even though there have been numerous studies in similar concerns, however, the experiment design and goal are mostly different.

In this article, we conducted an experiment to automatically profile the candidates based on their application documents by using two techniques of text mining, namely Document Vector analysis and N-Gram analysis. A number of 126 application documents from 3 different job positions, namely IT, Human Resources Staff, and Marketing Staff were analyzed. Features were extracted by using these two techniques and at the final stage, scoring was done on those two techniques. The higher the score of a document, the higher the rank obtained for being accepted. The ranks from both methods are then being validated by the real data released by the HRD from the manual selection process.

3. ORIGINALITY

This study explores text mining techniques to score pdf documents from candidate applications by using two techniques, namely Document Vector analysis and N-Gram analysis. In the final stage, we will evaluate and compare the two techniques in scoring the documents to extract the exact profile that matches the best with the requirements determined by the company HRD division. Compared to some previous studies [5-7, 23-27], this study uses application documents based on Indonesian text. Specific dictionary regarding each job vacancy is defined by the HRD division. For example, from IT Vacancy, the specific competency and skills that are needed by the company are database programmer and database administrator. Some specific terms are set as a reference for obtaining the document scores. For Document Vector analysis, the Knime application is used to preprocess the application document until calculating the distance between the document and the terms set (specific dictionary). Euclidean distance is implemented to calculate the difference [8]. In the second method, we implemented the N-Gram technique. In the N-Gram technique [22], all processes are done by using python programming. In the preprocessing technique, we use sastrawi library as Indonesian stopword and stemming function. Since most of the skills needed are in English terms, after performing preprocessing stage (data cleansing), we extract the English terms by running the python function for English terms identification. N-Gram method is then implemented to obtain the filtered skills from the resulting English terms. Similar to the Document Vector method, the final score in the N-gram method is also done by matching the final extracted text with the specific dictionary regarding the vacancy skill sets defined by the HRD division. The more the terms are matched with the specific dictionary terms defined, the higher the score obtained by each document.

The novelty of this study is, we evaluate two text mining techniques in performing automatic profiling on application documents that are based on the Indonesian language. A Specific dictionary that consists of specific terms in representing the skills needed for each job vacancy is used to match the final score resulting from the two text mining techniques. In addition, Document Vector and N-Gram analysis are two text mining techniques that need low computation time. When the result is promising, these techniques are applicable to be used in real automatic document profiling when a company received thousands of pdf application documents.

4. SYSTEM DESIGN

4.1 Research Method

This article consists of two methods of text mining, wherein each method consists of 6 stages, starting from data collection to validation, as shown in the following figure (figure 1 and figure 2), and the details of these stages will be explained in the following sub-chapters.

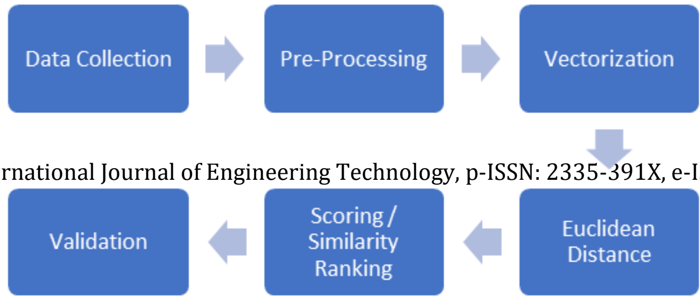


Figure 1. Document Vector Method

For the N-gram method, there are several steps, including the following:

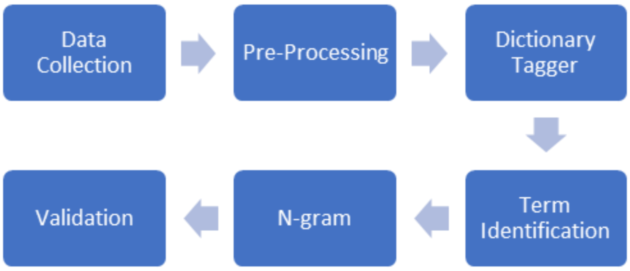


Figure 2. N-Gram Method

4.1.1 Data Collection

This research aims to compare the accuracy resulting from the two methods, therefor both methods use the same data. This research collected secondary data from a private company. There were 128 application documents, consisting of 43 documents for HRD staff, 40 documents for IT (Data Developer), and 45 documents for the Marketing position. The stages to be carried out were as follow:

- Interviewing with the HRD team, especially for generating the specific dictionary for each job vacancy.
- Collecting data from HRD regarding the three vacancies, with details as follow:

Table 1. Data Job Vacancy

Position	Submitted Resume
IT (Data Developer)	42
HRD Admin	41
Marketing	43

a. Data Developer

In this era of big data, being exposed to the most recent technology is essential. Google and Facebook are examples of companies that have applied such big data for a long time, and of course, they have been dealing with petabytes of data [5]. Human resources with analytical skills and creativity are required in applying Big Data technology. The abilities or skills required include determining a new method that can be performed to collect, interpret and analyze data, as well as computer programming skills and business skills or the understanding toward business goals [9]. Business intelligence specialists, data developers, data science, and data analyst are the skills that are looked at by the company.

For this reason, the company states such necessities into the candidate's requirements. Here is a dictionary regarding the requirements needed:

Table 2. Requirement IT Position

Specific Dictionary Determined for IT Requirement	
Hadoop, Olap, Tableau, PowerBI	MongoDB
AWS, Azure	Oracle
Data Synchronization	RDBMS
Informatics	PostgreSQL
Springboot	Redis
Phyton	ERD
Apache, Nifi	Kafka, Spark, Flink

b. Marketing

The attempt to get attention and retain customers is one of the essential things to do in this ever-increasing competitive world of business [10]. Therefore, marketing is needed to help smoothen the marketing in this company.

Marketing is an effort to introduce products to customers. Actions to be carried out include several activities ranging from promotion, distribution, sales to the products development strategy. The marketing group is responsible for customer loyalty, negotiating contracts, and implementing sales. Marketing is responsible for the first step to building customers [11]. Marketing plays a significant role in the development of a business. It can be said that there is no successful business without an adequate effort toward marketing. The benefits of having marketing include:

- Increased sales
- Having a good relationship with consumers
- As a means of doing branding
- Developed products

In this research, the company needed to get a candidate in marketing who was also proficient in digital marketing or had ever had experiences regarding such a position. Accordingly, this company provided criteria that match current needs. The following is a dictionary of requirements that exist in marketing in this study:

Table 3. Requirement Marketing

Specific Dictionary Determined for Marketing Vacancy	
Sales, Marketing, Promotion	SEO, Google AdSense
Adobe photoshop	Corel draw
Innovative, Creative	Loyalty
Target, Discipline	Analytical
Digital marketing	Teamwork, Organizational
Ms. Office, PowerPoint	Content Writer
Negotiation	Presentation, Relation
Thorough	Neat

c. HRD Staff of Administration

Human Resources Development is a series of organizational activities carried out at a specific time and designed to produce behavioral change. It has been concluded that HRD focuses on improving the performance of employees in the company [12]. The tasks and activities of HRD are immensely complex and usually equipped with a complete team ranging from managers to administrators. In addition, administrators also give help with new employee recruitment. Social and communication skills are highly needed, besides, the ability to operate a computer. Specifically, when the HRD staff must work with the Human Resource Information System (HRIS). With an appropriate Human Resource Information System, Human Resource staff enables far more strategic functions in the organization [13]. So, one of the requirements for HRD administration would be to adapt quickly to technology, especially technologies in Human Resource Information systems. The followings are the requirements for HRD in this study:

Table 4. Requirement HRD

Specific Dictionary Determined for HRD Staff Vacancy	
HRIS (Human Resource Information System	Psychology, Administration, Law
Insurance	Payroll, Attendance
Neat, Thorough, Discipline	Ms office, Ms excel
Pivot, Vlookup	Recruitment, Training
Communication	Filing, Document Archive
E-mail	Evaluation, Performance

4.1.2 Pre-Processing (Document Vector Method)

On this method, Knime open software is used for document preprocessing and vectorization. Based on figure 1, after data collection, the next process is pre-processing which includes document cleaning and filtering [14]. This phase is the most critical and complex process that leads to the representation of each document by a selected set of index terms [15]. The pre-processing technique implemented in this method are:

- Preparing all the application documents and specific dictionaries in one set of documents to be processed in the data preparation stage.
- Punctuation erasure: it removes punctuation marks or symbols in the dataset. Symbols such as ".", "," will be erased in this stage.
- Filtering (Number filter and N-Char filter) Filters all terms contained in the input documents that consist of digits, including number, decimal separators "," or "." and possible leading "+" or "-". There is also an option to filter all terms that contain at least one digit. This phase will filter all terms that will not be used in the analysis step.
- Case Folding: it changes all the letters in the text into small or capital ones (lowercase) [16]. In this research, all documents will be converted to lowercase.
- Stopword removal: to eliminate words that are considered not to give any effect, usually replacing liaison. In this study, the stopwords list used is the Indonesian stoplist [17].
- BOW (Bag of Words): is a model representing an object as a bag (multiset) of the word. At this stage, each sentence in the Application Document will be turned into separate words [18].

4.1.3 Document Vectorization

Document vectorization is a process of transforming text data into a numerical data representation. In Knime, this process can be done by applying a node called document vector. This node creates a binary or numerical vector representation for each term or document, based on the filtered bag of word input data table. Vectorization enables the machines to understand the textual contents by converting them into meaningful numerical representations [19]. In this stage, all the application documents are converted into vectors. The dimension of the vectors for one document will be a 1xN vector, with N being the number of distinct terms in the BoW (Bag of Words). In Knime, the Node for the document vector is divided into 3 parts, the input part (input port), the setting, and the output part.

1. Input Port, the input port of document vector Node is the bag of words of the documents. The form of the input port is a table containing MxN matrix data (M is the number of rows, whereas N is the number of columns). The 1st row of M is representing the specific dictionary terms, the 2nd row of M is representing the first application document, followed by the 3rd application document, the 4th application document, etc, consecutively until all the application documents are processed. For example, if the number of application documents is 42, then the MxN matrix will be 42xN, with N being the number of terms extracted as the BoW of each document.

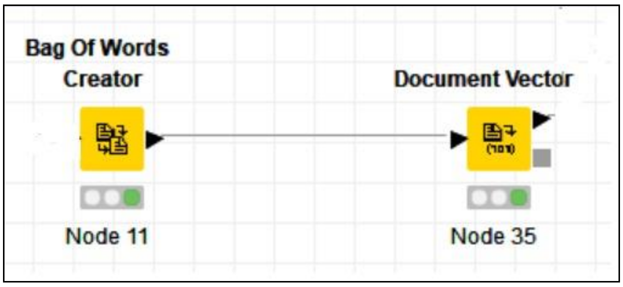
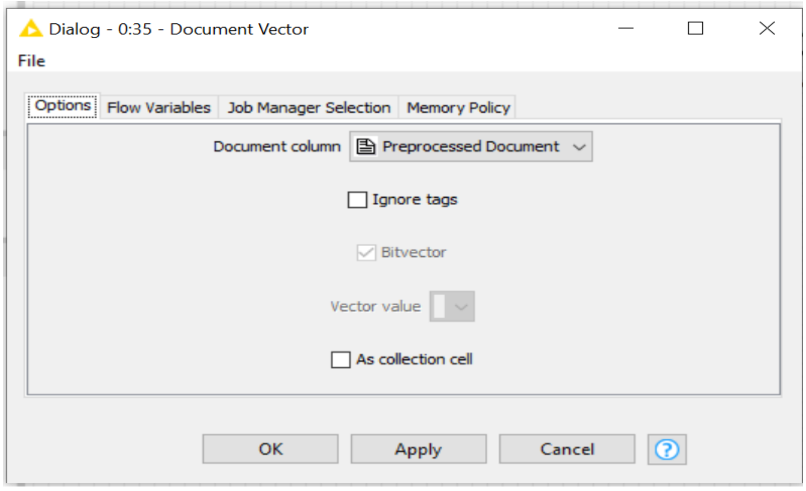


Figure 3. Document Vector Node in Knime

The document vector Node is configured as follows:



Document column: Preprocessed Document

□ Ignore tags

☑ Bitvector

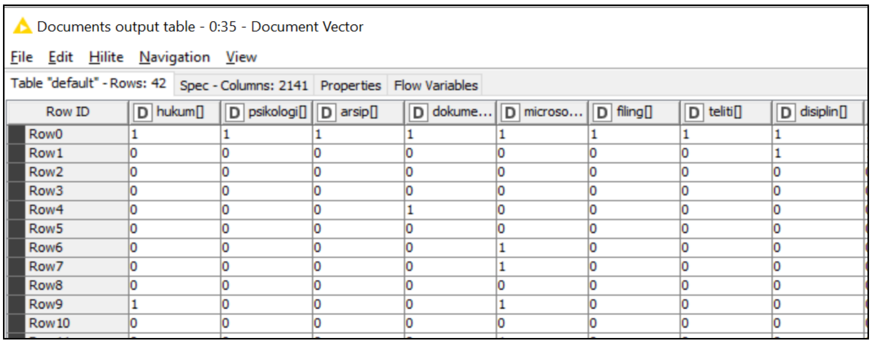
Vector value

□ As collection cell

Figure 4. Document Vector Setting

Based on figure 4, the Bitvector setting is checked and vector value will be created indicating whether certain terms is contained in a document.

2. Output Port, the output port of document vector Node will be a table that contains the relationship between each application document with the specific dictionary (defined terms based on each vacancy). Value “1” indicates that a specific term in the dictionary is matched with the terms in the application document, vice versa for value “0” (see figure 5). Figure 5 below shows a 10 sample output table in document vector output:



Documents output table - 0:35 - Document Vector

Row ID	D hukum▯	D psikologi▯	D arsip▯	D dokume...	D microso...	D filing▯	D teliti▯	D disiplin▯
Row0	1	1	1	1	1	1	1	1
Row1	0	0	0	0	0	0	0	1
Row2	0	0	0	0	0	0	0	0
Row3	0	0	0	0	0	0	0	0
Row4	0	0	0	1	0	0	0	0
Row5	0	0	0	0	0	0	0	0
Row6	0	0	0	0	1	0	0	0
Row7	0	0	0	0	1	0	0	0
Row8	0	0	0	0	0	0	0	0
Row9	1	0	0	0	1	0	0	0
Row10	0	0	0	0	0	0	0	0

Figure 5. Document Vector output sample

4.1.4 Euclidean Distance

Euclidean distance is a distance calculation of two points in the euclidean space. Euclidean distance is most often used to compare the profile of across variables [20]. The distance between vectors X and Y is defined as follows:

$$\sqrt{(x^2 - x^{1)2} + (y^2 - y^{1)2}} \qquad (1)$$

Where X = (X_1, X_2, ..., X_j) is variable in document than
y= y_1, y_2,..., $y_{j)}$ is variable in the central point

In other words, Euclidean Distance is the square root of the sum of squared differences between corresponding elements of the two vectors. In this article, Euclidean distance calculates the score between the document vector and specific dictionary in each vacancy.

4.1.5 Scoring based on Similarity Ranking

At this stage, ranking is carried out based on the similarity results between the document vector and the specific dictionary generated on each job vacancy. The number of documents being ranked was based on the number of applications selected by the HRD using manual screening (see table 4).

Table 5. The number of accepted applications by the HRD through the manual selection process

Vacancy	Number of Document applications	The selected document by HRD (manual screening)
Data Developer	42	16
HRD Staff	41	19
Marketing Staff	43	27

4.2 Pre-processing for N-Gram Method

The N-Gram method is shown in figure 2 above. The preprocessing stage is almost the same as in the Document Vector method. However, there are several additional steps in the preprocessing, namely as follows:

- Sentences tokenize is the stage of enumerating a document into a per sentence with a sentence marker (delimiter) being a dot (.). It makes the system can identify English terms in each sentence. We look for English terms is because all of the competencies in the application documents are presented in English terms, such as Data Synchronization, Database Management, Computer Network, Etc.
- Word Tokenize is to break the words in a document into a single word so that the English words in the document can be tagged using the English dictionary. This step is the follow-up of the previous step (Sentence Tokenize). For this method, python programming is used.
- Dictionary Tagger, dictionary tagger is tagging English words/phrases in a document with an English dictionary to identify English terms. Meanwhile, words/phrases other than English will be replaced with the number 0 (zero) used as a separator between one term and another term.
- Term Identification, when all the English phrases have been identified and separated by using "code 0", term identification is the process of getting the English terms correctly from the phrase, because sometimes the English phrase consists of 3 or 4 words, for example: "network maintenance in government office". In this example, we need to get only the term: network maintenance. We performed this step by using verb phrase chunking.

4.3 N-Gram Method

The dictionary tagging process aims to identify terms based on predefined keywords. Dictionary taggers are also known as part of speech (POS), Part-of-speech tags words in a sentence with a predetermined set of tags [21]. Each marked word will represent each dictionary entity to be identified. The method used in marking the words in the document is the N-gram method. Since the specific dictionary for each vacancy consists of one to three words (N=1 - 3), it takes the value of the N-Gram profile, namely unigram, bigram, and trigram. The N-Gram method works as follows: when the words in the specific dictionary consist of one word, the N-Gram profile value used by the system is unigram (N=1). For two words, it will use bigram (N=2), and when the words in the specific dictionary consist of three words, then the profile value used is a trigram (N=3). In principle, we apply all N values (from 1 to 3) to obtain the terms from the application document to be compared with those from the specific dictionary. The more the same terms found from the application document, the higher score was given.

On the validation stage, all the application documents were then ranked from the highest score to the lowest and cut with the same number as the number of applications accepted by the HRD (see table 5). The validation is done by matching all the selected documents through the N-Gram method with the documents accepted by the HRD using a manual screening process. To calculate the accuracy, we use the equation below:

$$\text{Accuracy} = \frac{\textit{The number of matched documents}}{\textit{The number of documents accepted by the HRD using manual process}}$$

5. EXPERIMENT AND ANALYSIS

The result of our experiment is shown in Table 6. Column iii is the number of documents selected by the Document Vector method, while column iv is by the N-Gram method. Actually, both methods selected the same number of documents as the number of documents accepted by the HRD (see column v below), however, only the matched names are used for the accuracy calculation. The unmatched names were removed from the analysis. For example, when we compare the number of selected documents by the N-Gram method in Data Developer vacancy (see column iv), we found only 14 documents. This is because from the 16 documents selected and ranked by the N-Gram method, there were two documents that were not accepted by the HRD using a manual selection process. In this case, we then calculate there were only 14 documents that matched with the selected documents by the HRD. Using the same way, we then calculated the accuracy from the Document Vector method in all vacancies and presented it in Table 7. There are several aspects that we can discuss regarding this result. Overall, we find that the accuracy from both methods was not as high as we expected. We find that the specific dictionaries defined by the HRD for selecting the proper candidate take an important role in this process because this dictionary is the ground truth for comparing the textual content of each application document. Some drawbacks such as different terms that contain the same meaning could be a challenge in text processing. For example, the term *database programmer* and *database administrator* could have the same meaning, but in the matching process those two terms could end up on a different side. In addition, when we look at the *insurance* term in the specific dictionary and when this term consists of only one word, this term will be detected by the N-Gram algorithm when using Unigram. However, the standalone term like the *insurance* in a document could mean many things. The representation of term *insurance* could be: *I have experience working in an insurance company*, or *I have experience as an administrative staff in an insurance company.* These two text expression example shows the potential of error when applying unigram in the N-Gram algorithm. However, in this experiment, all the specific dictionary terms are given by the company and we could not interfere.

When we compare with the Document Vector method, the way Document Vector method in calculating the distance between the extracted text from the application document and the specific dictionary terms is by using Euclidean Distance. In the extracted text, each word will be scored by value 1 when the word is matched with the words in the specific dictionary terms. So in the end, each application document will create a set of values that consists of 1 and 0. The distance is then being calculated from that set score values. The perfect value will be obtained when the score is equal to the number of words in the specific dictionary. We notice also some drawbacks in this algorithm, for example, when two application documents show the same score in the Euclidean Distance calculation, it does not mean that the terms listed are also the same. This means that some candidate competencies could be different even if they have the same score in the rank. Furthermore, some value "1s" could be spread out along the specific dictionary terms or could also be in one side of areas on the specific dictionary terms. These two conditions certainly tell us the different candidate competencies.

Table 6. The Result of document selection from two methods

Vacancy	Number of Document applications	Selected Document by Document Vector	Selected Document by N-Gram	The selected document by HRD (manually screening)
i	ii	iii	iv	v
Data Developer	42	12	14	16
HRD Staff	41	14	13	19
Marketing Staff	43	21	21	27

Those discussed aspects are the main cause of the resulted accuracy we obtained during this experiment. Several findings for future improvement need to be addressed such as the terms listed in the specific dictionary should be evaluated in a more accurate and precise way. Some terms, whether in one, two, or three words, that have the same meaning with different words expression need to be considered and included in the specific dictionary, in both methods.

Table 7. The Accuracy comparison from the two methods

Vacancy	The accuracy of the Document Vector method	The accuracy of the N-Gram method
Data Developer	75%	87,5%
HRD Staff	74%	68%
Marketing Staff	78%	78%

6. CONCLUSION

This paper explores the potential of text mining on document selection for automatic candidate profiling when a company receives many applications that need to be processed efficiently. Since manual document screening needs a lot of effort, time, and human resources, text mining for selecting the best candidate can be a good option for the company's efficiency and effectiveness. N-Gram and Document Vector methods could be used to extract the terms from the application document to profile the candidate for the further selection process. Considering this experiment result we found that both methods, the Document Vector, and the N-Gram method showed a high potential to be implemented in the real automatic screening process for obtaining the best candidate. N-Gram method overall shows a slightly better result compared to the Document Vector result. However, some optimization needs to be done, because the average accuracy from both methods has not yet reached 100%. Overall, this result gives us a better understanding of how to implement text mining to automatically profile the candidate based on their document applications more precisely. The specific dictionary terms regarding the candidate competencies, the complete terms that have different expressions but with the same meaning, and a more precise calculation technique regarding the matching process between the extracted terms and the specific dictionary are aspects that need to be optimized to get better accuracy.

Acknowledgments

There is no conflict of interest in writing and publishing this article. This study is funded by Kementerian Pendidikan dan Kebudayaan Nasional Indonesia (Kemendikbud-Ristekbrin).

REFERENCES

[1] P. Hendrarso, **Meningkatkan Kualitas Sumber Daya Manusia di Perguruan Tinggi menuju Era VUCA : Studi Fenomenologi Pada Perguruan Tinggi Swasta**, Prosiding Seminar Stiami, vol. 7, no. 2. 2020.

[2] S. R. Astari, "**Penerapan Profile Matching Untuk Seleksi Asisten Laboratorium**," Telematika, vol. 16, no. 1, p. 1, 2019, doi: 10.31315/telematika.v16i1.2987.

[3] J. Kuswanto, "**Penerimaan Karyawan Baru Menggunakan Metode Profile Matching**," J. Ilm. Sist. Informasi, Teknol. Inf. dan Sist. Komput., vol. 15, no. 2, pp. 85–97, 2020.

[4] E. Sutinah, "**Sistem Pendukung Keputusan Menggunakan Metode Profile Matching dalam Pemilihan Salesman Terbaik**," Informatics Educ. Prof., vol. 2, no. 1, p. 234409, 2017.

[5] Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). **Text mining in big data analytics.** Big Data and Cognitive Computing, 4(1), 1–34. https://doi.org/10.3390/bdcc4010001

[6] Wosiak, A. (2021). **Automated extraction of information from Polish resume documents in the IT recruitment process**. Procedia Computer Science, 192, 2432–2439. https://doi.org/10.1016/j.procs.2021.09.012

[7] Alanoca, H. A., Vidal, A. A. R. de C., & Saire, J. E. C. (2020). **Curriculum Vitae Recommendation Based on Text Mining**. http://arxiv.org/abs/2007.11053

[8] A. Aditya, B. N. Sari, and T.N Padilah, "**Perbandingan pengukuran jarak Euclidean dan Gower pada klaster k-medoids**," Jurnal Teknologi dan Sistem Komputer, vol. 9, no. 1, pp. 1-7, 2021.

[9] A. Ali, J. Qadir, R. ur Rasool, A. Sathiaseelan, A. Zwitter, and J. Crowcroft, "**Big data for development: applications and techniques**," Big Data Anal., vol. 1, no. 1, 2016.

[10] D. Rapitasari, "**Digital marketing Berbasis Aplikasi Sebagai Strategi Meningkatkan Kepuasaan Pelanggan**," J. Cakrawala, vol. 10, no. 2, pp. 107–112, 2016.

[11] Kotler, P., Rackham, N., & Krishnaswamy, S. (2006). **Ending the War Between Sales and Marketing.** www.hbrreprints.org

[12] Kasmawati, "**Pengembangan Sumber Daya Manusia Dalam Organisasi Pendidikan Islam**," J. UIN Alaudin, vol. VIII, no. 2, pp. 392–402, 2019.

[13] I. A. Zarqan, "**Human Resource Development in the Era of Technology**; Technology's Implementation for Innovative Human Resource Development," J. Manaj. Teor. dan Terap. | J. Theory Appl. Manag., vol. 10, no. 3, p. 217, 2017.

[14] M. Habibi, "**Implementation of Cosine Similarity in an automatic classifier for comments**," JISKA (Jurnal Inform. Sunan Kalijaga), vol. 3, no. 2, p. 110, 2019.

[15] D. Soyusiawaty and Y. Zakaria, "**Book data content similarity detector with cosine similarity** (case study on digilib.uad.ac.id)," Proceeding 2018 12th Int. Conf. Telecommun. Syst. Serv. Appl. TSSA 2018, 2018.

[16] R. Saptono, H. Prasetyo, and A. Irawan, "**Combination of cosine similarity method and conditional probability for plagiarism detection in the thesis documents vector space model**" J. Telecommun. Electron. Comput. Eng., vol. 10, no. 2–4, pp. 139–143, 2018.

[17] A. W. Pradana and M. Hayaty, "**The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts**," Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control, vol. 4, no. 3, pp. 375–380, 2019.

[18] S. Sohangir and D. Wang, "**Improved sqrt-cosine similarity**

measurement," J. Big Data, vol. 4, no. 1, 2017.

[19] A. K. Singh and M. Shashi, "**Vectorization of text documents for identifying unifiable news articles**," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 7, pp. 305–310, 2019.

[20] Singh Lehal M, Kumar, A, Goyal, V, "**Comparative Analysis of Similarity Measures for Extraction of Parallel Data**", International Journal of Control and Automation, Vol. 12, No. 6, pp. 408-417, 2019.

[21] A. Koochari, A. A. Gharahbagh, and V. Hajihashemi, "**A Persian part of speech tagging system using the long short-term memory neural network**," 6th Iran. Conf. Signal Process. Intell. Syst. ICSPIS 2020, 2020, doi: 10.1109/ICSPIS51611.2020.9349556.

[22] Wikipedia/wiki/n-gram

[23] Kinoa, Y., Kurokia, H., Machidab, T., Furuyab, N., Takanob, K., "**Text Analysis for Job Matching Quality Improvement**," Int'l Conf. on Knowledge Based and Intelligent Information and Engineering Systems, 2017.

[24] Almada, R. V., Elias, O. M., G ´omez, C. E., Mendoza, M. D., L ´opez, S. G., **Natural Language Processing and Text Mining to Identify Knowledge Profiles for Software Engineering Positions**, 5th 81st Int'l Conf. in Software Engineering Research and Innovation (CONISOFT), 2017.

[25] S A Md Nasir, W F Wan Yaacob, and W A H Wan Aziz. **Analysing Online Vacancy and Skills Demand using Text Mining.,** Journal of Physics: Conference Series., 1496 (2020), IOP Publishing, doi:10.1088/1742-6596/1496/1/012011

[26] Debortoli S, Müller O and vom Brocke J., (2014). **Comparing business intelligence and big data skills: a text mining study using job advertisements.** Business & Information Systems Engineering 6(5)

[27] Karakatsanis I, AlKhader W, MacCrory F, Alibasic A, Omar M A, Aung Z and Woon W L. (2017)., **Data mining approach to monitoring the requirements of the job market: A case study.** Information Systems Vol 65 p1-6.