ISBN: 978-981-94-4198-3
2025 the 15th International Workshop on Computer Science and Engineering (WCSE 2025)
doi: 10.18178/wcse.2025.06.012

Prediction Questions for National Examination with Generalized Vector Space Model (GVSM) Algorithm

Nurma Ayu Wigati [1 +] and Ajeng Dwi Asti [2]

[1] Faculty of Mathematics and Natural Science, Universitas Indonesia, Depok, Indonesia
[2] Faculty of Information Technology, Universitas Andalas, Padang, Indonesia

Abstract. National Examination (UAN or UNAS or UN) is a government evaluation measurement tool used to determine the quality of education in Indonesia. Quality is evidenced when students can answer questions in the national examination based on materials aligned with the Graduate Competency Standards (SKL). Materials published in the past can also be republished for the next exam. This SKL is divided into several research topics that are relevant to current trends. This study aims to determine how effective the use of the Generalized Vector Space Model (GVSM) algorithm is in predicting each question that will appear. The GVSM algorithm determines the similarity between words that appear in one document and words that appear in another. The GVSM algorithm is used to determine the similarity between words that appear in one document and words that appear in another document. The evaluation results showed an accuracy of 0.75, a precision of 0.7321, and a recall of 0.7017.

Keywords: national examination, GVSM, prediction question, similarity word, evaluation

1. Introduction

Examinations are a way to evaluate the learning process. In the field of national education, the government introduced exams as a tool to measure the level of quality achieved in understanding the chosen learning field between teachers as educators and students as learners [1]. The Indonesian National Exam (commonly abbreviated as UAN, UNAS, UN) is administered to students in their final year of primary school (SD), junior high school (SMP), and senior high school (SMA). The national exam uses adjusted and created questions based on the National Education Standards Agency (BSNP). The questions asked to students must be in accordance with the Graduate Proficiency Standards (SKL), which are used as assessment material to monitor how well Indonesian students have understood the content taught during their schooling [2].

Each defined SKL contains a specific topic. Each question has a topic that describes an indicator of the graduate-level competency standard (SKL). The question generation process needs to categorize or group the questions into appropriate subjects based on the SKL topics. This is still done manually and there can be errors in the process of grouping questions, such as teachers not paying attention while inserting specific topics into questions or the grouping process taking time, making the question creation process inefficient.

A step to consider while grouping into topics of related subject areas is to look for similarities between one piece of information and another. Various methods are used in information similarity searching and information retrieval, such as the Vector Space Model (VSM) and the Generalized Vector Space Model (GVSM). VSM may be a strategy that measures the closeness between a report and an inquiry address entered by a client as a test archive by taking the cosine of the point between the vector shaped by the archive and the vector of the catchphrase entered by the client. However, VSM has a weakness. It assumes that each term in a document is independent. This method does not recognize the semantic relationship with other terms [3].

For example, if a user searches for the keyword "programming", the search results will show all question documents that contain only the word "programming", but there are still many question documents that are semantically related to the word "programming", such as "PHP", "Java", etc. In this case, the recall of the search results will decrease. Therefore, a method that can develop VSM by adding sensory features (synonyms)

+ Corresponding author.
E-mail address: nurma.ayu@sci.ui.ac.id

to this model is needed, namely GVSM [4]. GVSM is widely used to measure the similarity of words and produces significant results. Still, not many studies have measured document sections of a sequence of sentences and are limited to only one language, English [5]. This work is an extension of previous work that used only one of two languages, Indonesian and English.

Based on the depicted conditions and issues the yield is anticipated to permit us to analyze how viably the Generalized Vector Space Model (GVSM) calculation is in anticipating each address that will emerge. To answer the research objectives, our research includes the following: Section 1 provides the background in which the authors take up the application of GVSM to predict the similarity of questions in national examination questions, Section 2 deals with the theoretical foundations of text preprocessing and the application of GVSM. Section 3 portrays the investigative strategy and how the investigation was carried out. Chapter 4 describes the results carried out and the analysis of how the technical research is completed and how the output results are analyzed. Chapter 5 describes the conclusions drawn from the conducted research and future work.

2. Related Works

Text preprocessing has three main phases: case folding and tokenization, filtering, and stemming. Case sensitivity means that all characters in the document are converted to lowercase, and only characters from 'a' to 'z' are accepted. Non-alphabetic characters are considered delimiters. The tokenization phase is the phase where the input string is shortened based on each word that composes it [6]. The filtering phase is the phase where important words are selected from the token results. A stoplist algorithm (removing less significant words) can be used. Stoplists or stopwords are non-descriptive words that can be removed from the bag-of-words approach [7]. The stemming process can be performed for different languages, including the stemming process for Indonesian and English texts. The stemming process for Indonesian text is different from the stemming process for English texts. For English texts, only the process of removing suffixes is performed.

The Generalized Vector Space Model (GVSM) expands the standard Vector Space Model (VSM) by including extra data to expand the terms related to the displayed inquiry archive. The included data is semantic data from a word reference. The incorporation of semantic information performed when modifying the standard VSM improves the performance of text retrieval. The model is a new calculation performed to determine the semantic relatedness between existing terms [8]. SR is the relevance of terms between documents, d_i and d_j are the weights of question documents and queries (question inputs entered as test question documents) calculated using the TF-IDF formula, and n is the number of question documents [9]. SR can be calculated in two ways: by calculating the maximum value of semantic relevance between senses (synonyms) and the maximum value between concepts (interrogatives). SR is the semantic relationship that connects two senses (synonyms), s_1 and s_2. Using the maximum semantic relatedness between terms, if two words in a question do not contain the same word in the dictionary, $SR\ (T, S, O) = 1$. If the dictionary contains the same word and the other does not, then $SR\ (T, S, O) = 0$ [8].

3. Methodology

This study has 6 activity phases, including data collection and grouping, text preprocessing, GVSM modeling with SCM and SPE calculation, TF-IDF weighting, text similarity, and evaluation. The methodology of this study is shown in Figure 1.

3.1. Collecting Data

The data used in this study comes from a national exam collection by Ganesha Operation, one of the tutoring private institutions in the city of Solo, Indonesia, for senior secondary school students majoring in science and social science. The exam questions are categorized into specific topics based on indicators defined in SKL [1].

3.2. Text - Preprocessing

This process has three main phases: tokenization, filtering, and stemming. In this phase, we must perform a case-sensitive check to remove characters other than a-z. In the filtering phase, the Indonesian and English

literature is used as the stopword list [10]. In the stemming phase, Porter Stemming is used for English and Nazief & Adriani for Indonesian, using the PHP library Sastrawi. Table 1 is the example for stopword in Indonesia and table 2 is for English.

Table 1: The Example of Indonesian Stopword List

No	Indonesian Stopword
1	ada
2	adanya
3	adalah
4	adapun
5	agak

Table 2: The Example of English Stopword List

No	English Stopword
1	am
2	an
3	and
4	any
5	are

3.3. GVSM Modelling

To determine the degree of relatedness, we first compare related words in dictionaries to obtain the Semantic Relatedness (SR) value. The dictionaries we used are taken from [11] for Indonesian and [12] for English. In this step, we have to :

1. Calculate the SCM value to determine if there is a relationship between two words that have synonyms in the dictionary. Use the formula in Equation (1) for the calculation.

$$SCM\ (S,O) = \prod_{i=1}^{l} e_i \tag{1}$$

2. Calculate the SPE value to determine the synonym relationship between two same words. Use the formula in Equation (2) for the calculation.

$$SPE\ (S,O) = \prod_{i=1}^{l} \frac{2\ d_i\ d_{i+1}}{d_i + d_{i+1}} . \frac{1}{d_{max}} \tag{2}$$

3. The SR value is calculated using the formula in Equation (3) for the relationship between terms and the formula in Equation (4) for the relationship between meanings (synonyms).

$$SR\ (T,S,O) = \max\{SCM(S,O).SPE(S,O)\} \tag{3}$$

$$SR = tf - idf\ (t,d).\frac{2\ .d_i.d_j}{d_{max}\ (d_i + d_j)} \tag{4}$$

If the SR (Semantic Relevance) value is not 0, then GVSM modeling is performed. On the other hand, if the SR value is 0, then it is concluded that there is no meaningful relationship between the query (question input entered as the test question document) and the words in the document. Hence, modeling is performed using the VSM model. Modeling is calculated by directly providing the similarity weights between the words in the query (question input entered as the test question document) and the document without multiplying them by the SR value. The problem document is a collection of sentences from the question. On the other hand, the problem terms are a set of words that form connections in each question sentence.

3.4. TF – IDF Weighting

This process is performed to obtain a value for each successfully extracted term. To perform the TF-IDF weighting method, following these steps that we do:

1. First, for the occurrence of a word in each document, calculate the TF and DF using the formulas in (5) and (6).

$$tf\ (t,d) = f(t,d)\ (5);\ df\ (t) = f\left(d_{ij}\right) \tag{6}$$

2. Calculate the IDF, which is the logarithm of the proportion of address records containing the significant term to the entire number of address archives handled. Since the method includes different archives, we calculate the IDF according to equation (7).

$$idf\ (t) = \log \frac{N}{df\ (t)} \tag{7}$$

3. Calculate the weights between terms in the existing question documents using equation (8).

$$tf - idf\ (t,d) = tf\ (t,d)\ \times idf\ (t) \tag{8}$$

In this phase, a TF-IDF weighting system is created by prioritizing the term weights in the existing question documents and ignoring the term weights in the question documents in the query in question (question input is input as a test question document). This is because the difference in the resulting log ratio is not very large, 0.1. In this way, a higher cosine similarity value is achieved than if we consider the weights of the terms in the query in question (question input as test question document) and the weights of the terms in the existing question documents in question. A problem document is a collection of question sentences. A problem term is a set of words that are connected within each problem sentence. TF-IDF weights are modeled using GVSM and VSM modeling. If the SR value is equal to 0, GVSM modeling is used, and if the SR value is equal to 0, VSM modeling is used. Hence, the SR value is ignored for TF-IDF weights in the VSM model.

3.5. Text Similarity

This process is performed to calculate the similarity of existing question documents. In the process of similarity calculation, the cosine similarity model is applied. This can be seen from the formula (9) used when the SR value is 0, which means there is no relationship between the trained question document and the query (the question input entered as the test question document). On the other hand, the formula (10) shows that when the SR value is not 0, there is a relationship between the trained question document and the query (the question input entered as the test question document).

$$sim\ (d_i, d_j) = \frac{\sum_{i=1}^{n} d_i \sum_{j=1}^{n} d_j}{\sqrt{\sum_{i=1}^{n} (d_i)^2 \sum_{j=1}^{n} (d_j)^2}} \tag{9}$$

$$sim(d_i,d_j) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} d_i d_j SR}{\sqrt{\sum_{j=1}^{n} d^2{}_i \sum_{j=1}^{n} d^2{}_j}} \tag{10}$$

3.6. Evaluation

The evaluation used to test and analyze the study results was performed using k-fold cross-validation tests and accuracy calculations using a confusion matrix showing the accuracy, hit rate, and F-score of all tests performed [13]. To calculate accuracy, we use equation (11), precision (12), recall (13), and f-measure (14).

$$akurasi = \frac{TP\ (tema_1) + TP\ (tema_2) + \cdots + TP\ (tema_n)}{Total\ (tema_1) + Total\ (tema_2) + \cdots + Total\ (tema_n)} \times 100\% \tag{11}$$

$$p\ (tema_n) = \frac{TP\ (tema_n)}{Terprediksi\ (tema_n)} \tag{12}$$

$$r\ (tema_n) = \frac{TP\ (tema_n)}{Total\ (tema_n)} \tag{13}$$

$$F\ (tema_n) = \frac{2p\ (tema_n) * r\ (tema_n)}{p\ (tema_n) + r\ (tema_n)} \tag{14}$$

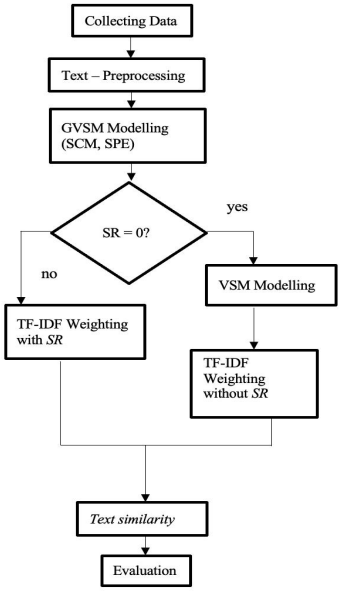

Fig. 1: Methodology Research

4. Results and Discussion

The system processes and uses the text document of questions to check each entered question and classify it as matching with SKL or not. The description of the system we created is shown in Figure 2.

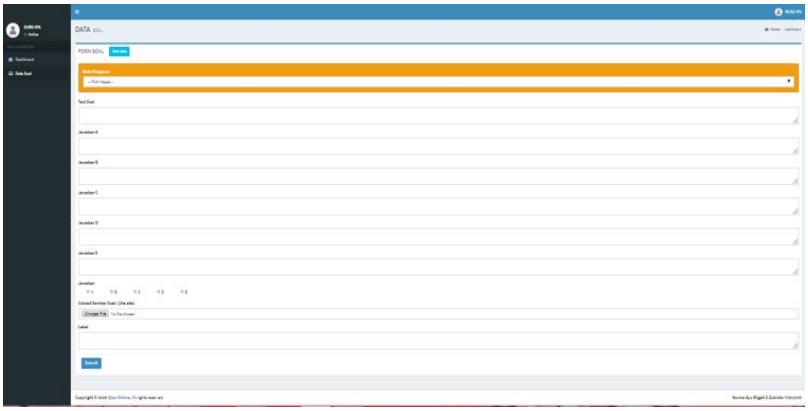

Fig. 2: Prediction System

The questions in the SKL national examination are divided into different topics. Not all topics are used in the study. The existing list of topics contains 124, but the topics used in this study are reduced to 122 as they are aligned with the indicators established in SKL. The topics not used are Listening Comprehension I and II of the English subjects. For example, we take one question from the Indonesian language question database, "Ratusan kompor berjajar pada areal sepanjang 1.000 meter, di depan sembilan warung di ujung Jalan Dewi Sartika, Cawang, Jakarta Timur. Warung, atau lebih tepat disebut bengkel, itu hanya berukuran 2x3 meter persegi Pada ruangan yang sederhana itu tampak penuh kompor bermacam-macam ukuran, tapi warnanya sama: biru. Di situ juga ada peralatan untuk memperbaiki kompor yang rusak. Isi paragraf tersebut adalah ..." After we finish process case folding and tokenization, filtering, and stemming, otherwise the result this first step is "ratus-kompor-jajar-areal-meter-sembilan-warung-ujung-jalan-dewi-sartika-cawang-jakarta-timur-warung-tepat-bengkel-ukur-x-meter-persegi-ruang-sederhana-tampak-penuh-kompor-macam-macam-ukuran-warna-biru-situ-alat-baik-kompor-rusak-isi-paragraf". In this sentence we find these word "macam" has synonim with "warna". We look for the relation synonim for each words and the result of SCM and SPE for modelling GVSM is 0.8708.

$$SR = tf - idf_{(warna,macam)} \cdot \frac{2.d_{(warna)}.d_{(macam)}}{d_{(max)}\,(d_{(warna)} + d_{(macam)})} = (1{,}814164326).\frac{2*3*2}{5\,(3+2)} = = 0{,}8708$$

The TF - IDF weighting for every words related with the synonyms can be seen in table 3.

Table 3: Result of TF – IDF Query

Bobot *TF – IDF* (w_q)	*Query*
1,8142	ratus
0	kompor
3,0183	areal
2,7173	meter
3,8782	sembilan
…	…
1,1262	baik
1,8142	rusak
0,8880	isi
0,5326	paragraf

We use GVSM for modeling, so the formula is Equation (10). The GVSM model calculates similarity by multiplying the weight of each term in the corresponding document by its Semantic Relevance (SR) value. Then, we divide it by the square root of each associated question document, calculate the square, multiply it by the query (the question input entered as the test question document), and calculate the square. For many Indonesian technical documents, the sum is 1044. In the example of calculating the text similarity of multiple documents, the result is:

$$sim(d_2,d_j) = \frac{\sum_{i=1}^{1044} d_2 \sum_{j=1}^{1044} d_j \, SR}{\sqrt{\sum_{j=1}^{1044} {d^2}_2 \sum_{j=1}^{1044} {d^2}_j}} = \frac{0{,}7885 * 0{,}8708}{\sqrt{204{,}7297^2 * \ 151{,}5523^2}} = \frac{0{,}7885 * 0{,}8708}{14{,}3084 * \ 12{,}3105} = \ 0{,}0039$$

This operation is performed when the calculation of similarity values by cosine similarity between the query (question input entered as the test question document) and each term in the question document is completed. The ranking is obtained from the previously performed cosine similarity calculation. The ranking is arranged from largest to smallest cosine similarity value. There are a total of 1044 documents related to the Indonesian language. After calculating the cosine similarity value of each document, the ranking results in the 500th document being the first document with the largest cosine similarity value (0.0910) and the last document sequence with the smallest cosine similarity value is the 1041st document, which has a cosine similarity value of 0. The ranking results of the cosine similarity values for the topic "Indonesian language" are shown in Table 4.

Table 4: Results Cosine Similarity

Cosine Similarity	Documents
0,0910	D500
0,0880	D801
0,0709	D39
....	....
0	D1041

The confusion matrix displays the accuracy, precision, recall, and F-measure for all the test questions administered.

$$akurasi = \frac{5010}{6680} = 0{,}75;\ p\ (tema_1) = \frac{32}{80} = 0{,}4000;\ p\ (tema_2) = \frac{33}{83} = 0{,}3976$$

$$r\ (tema_1) = \frac{32}{63} = 0{,}5079;\ r\ (tema_2) = \frac{33}{64} = 0{,}5156$$

$$F\ (tema_1) = \ \frac{2 * \ 0{,}4000 * 0{,}5079}{0{,}4000 + 0{,}5079} = 0{,}4475\ ;\ F\ (tema_2) = \ \frac{2 * \ 0{,}3976 * 0{,}5156}{0{,}3976 + 0{,}5156} = 0{,}4490$$

The average precision value across subjects was 0.7321, the recall was 0.7017, and the F-measure across subjects was 0.7178. The Vector Space Model (VSM) is a type of algebraic modeling that is often used to describe text or documents in vector form. This method is often used to determine the similarity value between the training question document and the query (question input as the test question document). The similarity

score of a string or word is higher the more frequently that string occurs in the document. VSM is widely used in information filtering, indexing, and relevance assessment. The calculation of the VSM algorithm is very strict because it only considers the frequency of word occurrence. It also needs to consider the semantic relationship between words (distant, equal, close). To overcome these weaknesses, the Generalized Vector Space Model (GVSM) algorithm was developed, which uses a similarity value in meaning in the process of calculating the semantic correlation (SC) between several synonyms. A word is entered only if it is also related to the query (the question input is entered as the test question document). The more synonyms of high-frequency words related to the query (the question input entered as the test question document) there are in a sentence, the higher the SC value of that sentence. The similarity of text in one document to text in another, called text similarity, can be measured using a similarity or distance function. The similarity and distance functions that can be used include Dice, Jaccard, Euclidean distance, Pearson correlation, and cosine similarity. The cosine similarity function is the best distance function computation for grouping purposes. In VSM, vectors between terms are treated as orthogonal pairs. However, this assumption is highly unrealistic since terms in a language usually have some degree of relevance to each other.

In this study, errors still occurred in predicting exam questions due to the uneven distribution of question data and the phenomenon of model overfitting. Therefore, to reduce bias and improve prediction accuracy, data normalization should be applied during the weighting process. Overfitting is a problem that occurs when a model learns too much from training data that contains noise and irrelevant variation, resulting in a model that places too much emphasis on specific patterns in the training data and cannot generalize to new data. In this context, overfitting can cause a model to be effective only for questions like those in the training data and less effective for new, different questions. Data normalization mitigates this problem by ensuring that the training data is proportionally distributed and more representative, helping to reduce the tendency of the model to memorize certain irrelevant patterns.

Data normalization also plays an important role in improving processing efficiency in terms of speed and accuracy. Normalized data with different scales and value ranges can be treated equally, allowing the model to process information more comprehensively and systematically. This significantly reduces processing time while reducing the chance of errors that may affect the result.

5. Conclusion

This consideration demonstrates that the Generalized Vector Space Model (GVSM) calculation is viable and has the potential to anticipate exam questions that will show up in future exams. However, the results obtained are still not fully optimal. This study goes through six main stages: data collection, text preprocessing, GVSM modeling to calculate the similarity coefficient matrix (SCM), and semantic proximity evaluation (SPE) to measure the similarity between words in related sentences, inverse word frequency document frequency (TF-IDF) weighting, calculation of text similarity, and evaluation of the results. Based on the evaluation performed, the algorithm produces an accuracy value of 0.75, precision of 0.7321, recall of 0.7017, and F-measure of 0.7178. These values indicate that the algorithm is effective, but there is still room for improvement. Factors that lead to suboptimal results include failure to apply word or sentence normalization during the weighting process. In addition, it is shown that an imbalance in data distribution also affects the model's performance and should be improved in future research. Future research is expected to include the process of automatic recognition of exam questions using multimodal data such as images and audio, allowing for automation of question data or direct import into the system. Furthermore, integrating other classification algorithms, such as Naive Bayes, Decision Tree, or ID3, is proposed to improve accuracy and obtain more optimal results.

6. References

[1] Kementerian Pendidikan dan Kebudayaan Republik Indonesia, “SKL UN 2013,” Badan Standar Nasional Pendidikan. Accessed: Dec. 25, 2024. [Online]. Available: http://bsnp-indonesia.org/2012/11/20/kis-kisi-un-tahun-pelajaran-20122013/

[2] D. Srinursih, “Ujian Nasional Menjadi Timbangan Kemampuan Siswa,” Kementerian Pendidikan dan Kebudayaan. Accessed: Dec. 12, 2024. [Online]. Available: https://www.kemdikbud.go.id/main/blog/2016/05/ujian-nasional-menjadi-timbangan-kemampuan-siswa
[3] J. Waitelonis, C. Exeler, and H. Sack, “Linked Data Enabled Generalized Vector Space Model To Improve Document Retrieval,” in *Proc. of 3rd Int. Workshop on NLP&DBpedia 2015*, Oct. 2015.
[4] Wong, W. Ziarko, and P. C. N. Wong, “Generalized Vector Space Model In Information Retrieval,” in *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR ’85*, 1985, pp. 18–25.
[5] X. Wang, Q. Chen, and Y. Yang, “Word vector embedding and self-supplementing network for Generalized Few-shot Semantic Segmentation,” *Neurocomputing*, vol. 613, p. 128737, Jan. 2025, doi: 10.1016/J.NEUCOM.2024.128737.
[6] M. Moqbel and A. Jain, “Mining the truth: A text mining approach to understanding perceived deceptive counterfeits and online ratings,” *Journal of Retailing and Consumer Services*, vol. 84, p. 104149, 2025, doi: 10.1016/J.JRETCONSER.2024.104149.
[7] R. Shao, P. Lin, and Z. Xu, “Integrated natural language processing method for text mining and visualization of underground engineering text reports,” *Autom Constr*, vol. 166, p. 105636, Oct. 2024, doi: 10.1016/J.AUTCON.2024.105636.
[8] G. Tsatsaronis, I. Varlamis, and M. Vazirgiannis, “Text Relatedness Based on a Word Thesaurus,” *Journal of Artificial Intelligence Research*, vol. 37, pp. 1–39, 2010.
[9] C. I. Nakpih, “A modified Vector Space Model for semantic information retrieval,” *Natural Language Processing Journal*, vol. 8, p. 100081, Sep. 2024, doi: 10.1016/J.NLP.2024.100081.
[10] Damian Doyle, “Stopword Lists.” Accessed: Nov. 02, 2024. [Online]. Available: https://www.ranks.nl/stopwords/
[11] D. Sugono and Pusat Redaksi Tim Bahasa, “Tesaurus Bahasa Indonesia Pusat Bahasa,” Sugiyono, Y. Maryani, Dra. M. T. Qodratillah, A. Budiwiyanto, D. Puspita, D. Amalia, and T. Santoso, Eds., Pusat Bahasa Departemen Pendidikan Nasional, 2008.
[12] Oxford Press University, *The Oxford Thesaurus An A-Z Dictionary of Synonyms*. Oxford Press University, 1997.
[13] D. M. Powers, “Evaluation : From Precision, Recall, and F-Factor to ROC, Informedness, Markedness & Correlation,” *Journal of Machine Learning Technologies*, vol. 2, pp. 37–63, 2011.