Scientific Journal of Informatics
Vol. 13, No. 2, May 2026

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Design and Evaluation of an AI-Based Indonesian–Malay Bilingual Dictionary Using Machine Learning Techniques

Nugroho Dwi Saputro[1*], Aziizatul Khusniyah[2]
[1]Department of Informatics, Universitas PGRI Semarang, Indonesia
[2]Department of Quranic Studies and Tafsir, UIN Sunan Kudus, Indonesia

Abstract.
Purpose: This study aimed to construct and evaluate an AI-based Indonesian–Malay bilingual dictionary by incorporating ML and NLP techniques through SLR. This research investigates salient issues in bilingual lexicography at the level of lexis, including limited parallel corpora, contextual semantic variability, and cultural nuances between closely related tongues.
Methods: A mixed-methodology, pairing a systematic literature review (SLR) of 18 peer-reviewed studies published from 2020 to 2025 with an experiment based on prototypes. As a result of these findings, we created a mini prototype dictionary via ML-based semantic-mapped supervised and unsupervised models and Transformer representations. We evaluated the prototype using objective measures of translation accuracy and qualitative analysis of errors.
Result: NLP and ML techniques are found to be effective in word-level bilingual dictionary building when integrated. Unlike direct equivalence approaches, which neglect contextual understanding in the representation of meaning groups, transformer-based representations capture high degrees of alignment in semantic generality through context. Our prototype produced suitable accuracy for most common lexical entries but struggles with informal expressions due to cultural nuances and underrepresentation of Indonesian–Malay parallel data.
Novelty: This study contributes by bridging the existing literature by linking systematic review findings with an empirical implementation in the form of a functional AI-based dictionary prototype, demonstrating that combined machine learning models are effective for Indonesian–Malay lexicography and providing a repeatable framework for AI-driven bilingual dictionaries between other language pairs.

Keywords: Artificial intelligence, Machine learning, Natural language processing, Indonesian–Malay dictionary, AI-based lexicography
Received December 2025 / **Revised** March 2026 / **Accepted** April 2026

INTRODUCTION

Bilingual dictionaries are crucial for cross-linguistic understanding, translation tasks and linguistic research in general but especially between closely related languages pairs, like Indonesian and Malay [1]. Since the two languages have origins that are interdependent on each other from history and linguistics, it leads to significant differences when compared in aspects such as semantics, usage patterns, orthography features which makes them very important and enables dictionary development where its mainly used for linguistic accuracy with practical usability [2]. Throughout the latter part of the twentieth century, particularly with the advent of computational dictionaries and corpus-based lexicography supported by an ever-growing onslaught of textual data expressed in digital form. So, nowadays, AI is playing a significant role in the development of modern dictionaries and dictionary creation is much more automated than in the past, including extracting words, building semantic networks, creating context-aware bilingual mappings [3], [4].

Traditional lexicographic methods heavily rely on manual annotation, linguistic knowledge, and tedious text analysis [5]. Keyword to create a lexicon creation is often done using corpus reading, expert validation and multi-stage refinement to arrive at entries (words or phrases), meaning, examples and semantic relations for these terms. Although these methods continue to have their utility, they are constrained by scalability and consistency as well as the time taken to process a large volume of text which is increasingly common with constantly evolving text formats. Similar to the Indonesian–Malay similarity, this language pair also

*Corresponding author.
Email addresses: nugputra@upgris.ac.id (Saputro)
DOI: 10.15294/sji.v13i2.40232

offers unique complexities for linguistic structuring including subtle semantic divergences and pragmatic differences as well as high informal variation that complicates its lexicographic processing [6].

The rise of artificial intelligence (AI) and machine learning (ML) has opened up new avenues for automating the construction of bilingual dictionaries [7]. Natural Language Processing (NLP) methods can provide syntactic parsing, semantic disambiguation and contextual meaning extraction at scale. Supervised and unsupervised machine learning algorithms facilitate the identification of lexical correspondences according to distributional similarity and usage context patterns [4]. In recent years, the use of deep learning approaches and Transformer-based architectures like BERT, RoBERTa and GPT3 has led to important advances in understanding semantic nuances and resolving polysemy, establishing them as valuable candidates for building more contextual dictionaries [8]. NMT models generate word-level translation outputs that are generally more fluent than their predecessors phrase or rule-based systems [9].

These advancements still face several challenges that may hinder the effective development of Indonesian–Malay bilingual dictionaries with AI [10]. A major challenge comes from lack of balanced and domain-representative corpora for both languages, especially the need for informal, dialectal or domain specific corpora [11]. Cultural and contextual differences generate semantic ambiguity that AI models cannot decipher without sufficient training data. Other linguistic phenomena (e.g., slang, code-switching, reduplication and dialectal variants) create even more uncertainty and reduce the reliability and generalizability of automated models. Deployment of AI-based lexicographic tools is also complicated by ethical concerns, model biases, and transparency issues [12].

Executor of Dictionary's function is to find or search for a word or phrase using this reference source based on the defined rules [13]. The majority of relevant works are on generic NLP tasks, machine translation systems or modelling for monolingual Indonesian or Malay. Only a handful of these studies cover the specialized demands of dictionary building: entry selection, sense differentiation and contextualization, or word-level cross-mapping. This is an important research gap that warrants systematic investigation [14].

However, previous works have not systematically incorporated techniques based on AI- and ML-based methods for constructing Indonesian–Malay bilingual dictionaries at the lexical level. The majority of previous studies focus on general machine translation systems, monolingual modeling (admittedly a relatively poor match for multilinear indexation), or broad NLP tasks without articulating the specific challenges in supporting lexicographer development through technological means encompassing entry selection, sense separation, contextualization/mapping and contextually-aware semantic validation. Moreover, there are few empirical implementations that integrate systematic literature synthesis and functional prototype development. This gap signals the requirements for a systematic synthesis with empirical validation specific to Indonesian–Malay bilingual lexicography.

As such, the aim of this research is to perform a Systematic Literature Review (SLR) in order to explore AI- and ML-based approaches for developing Indonesian–Malay bilingual dictionaries, identify challenges faced and assess existing research proposed solutions. Furthermore, this research extends the SLR findings by designing and evaluating at a mini-scale an AI-based bilingual dictionary prototype as a proof of concept, thereby demonstrating the practical applicability for word-level lexicographic construction of machine learning techniques. More specifically, this study will address the following research questions: (i) What factors, features or linguistic resources are used in constructing Indonesian–Malay bilingual dictionaries? (ii) What word-level dictionary developing AI algorithms, models and NLP techniques are employed? and (iii) What outcomes, outputs, or dictionary properties are generated from AI-based approaches?

This research mainly contributes in two ways. First, it is a much-needed focused and up-to-date systematic synthesis of the existing AI and ML techniques as related to the development of Indonesian–Malay bilingual dictionary which has remained largely unexplored compared to NLP as a whole. Second, it translates the literature review results into an empirical proof-of-concept prototype by illustrating how one might operationalize embedding-based and Transformer-informed representations within word-level bilingual lexicographic mapping. The combination of systematic review and applied experimentation distinguishes this study from previous studies that were either all theoretical or limited to generic translation systems.

METHODS

This study employs a Systematic Literature Review (SLR) to identify, evaluate and synthesize Artificial Intelligence (AI) including Machine Learning (ML) techniques used in the construction of Indonesian–Malay bilingual dictionaries. The follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure methodological transparency, traceability and replicability at every stage of the review process [15].

Search strategy and data collection

A thorough review of the literature was performed utilizing three prominent scientific databases, specifically Google Scholar, ScienceDirect and IEEE Xplore to find published articles regarding AI-based bilingual dictionary construction, lexicon extraction and Natural Language Processing (NLP) methods related to Indonesian and Malay. The search was limited to literature published from 2020–2025, which coincides with a period of major progress in neural NLP architectures (specifically Transformer-based ones). In order to diversify the breadth of search, while keeping specificity in mind, a wide array of targeted keywords were searched simultaneously "Indonesian–Malay dictionary," "bilingual lexicon," "lexicon extraction," "Natural Language Processing," "Machine Learning," "Deep Learning," "neural language model," word-level translation and lexicography automation. The search was refined by systematically combining these keywords with Boolean operators AND and OR to help retrieve studies more pertinent to AI-driven methodologies towards the development of dictionaries. The preliminary search resulted in 612 articles from three databases. Table 1 shows the detail of search keywords and query strings.

Table 1. Search keywords and query strings

Database	Search Query String
Google Scholar	("Indonesian–Malay dictionary" OR "bilingual lexicon") AND ("machine learning" OR "deep learning" OR "transformer model") AND ("lexicon extraction" OR "word mapping")
ScienceDirect	("lexicography" OR "dictionary construction") AND ("AI" OR "ML" OR "NLP") AND ("Indonesian" AND "Malay")
IEEE Xplore	("machine learning" OR "deep learning") AND ("bilingual dictionary" OR "lexicon building") AND ("Indonesian Malay")

PRISMA flow diagram

The inclusion studies were represented using the PRISMA flow diagram. Data extraction consisted of the following: identification, screening, eligibility and final inclusion. Figure 1 shows the PRISMA flow.

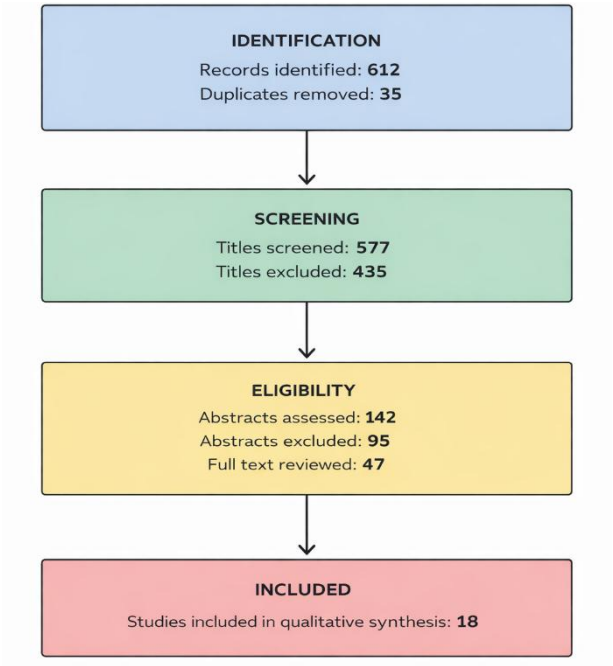

Figure 1. PRISMA flow diagram description

The systematic process of study selection is based on the PRISMA guidelines that guarantees transparency and methodological rigor. Identification: A total of 612 records were identified from the databases. After removing duplicates, 35 entries were removed and 577 records remained for screening. In the screening stage, all records titles were accordingly reviewed, and 435 studies were excluded due to irrelevance of AI-based approaches or not related Indonesian–Malay dictionary development. Abstract screening yielded a total of 142 studies. All studies underwent an in-depth assessment of their abstracts during the eligibility stage, leading to the exclusion of 95 articles that were not AI/ML-based or did not concentrate specifically on lexicon or dictionaries construction with insufficient methodological clarity. Then, 47 full-text articles were reviewed for final eligibility. From this assessment, a total of 18 studies met all of the pre-specified inclusion criteria and were included in the qualitative synthesis for this review.

Title screening
The process of screening the titles resulted in studies that were potentially relevant to the research focus. The inclusion criteria were articles whose titles contained the terms Indonesian–Malay dictionary or lexicon, bilingual translation resources, Natural Language Processing (NLP), Artificial Intelligence (AI), Machine Learning (ML), deep learning, and/or word-level lexical extraction. On the other hand, titles were excluded if they purely oriented towards one language (for example using monolingual NLP tasks such as sentiment analysis), or general machine translation without lexicon-level attention, non-technical linguistic studies; manual lexicography papers and AI applications unrelated to building a bilingual dictionary. On this screening stage, 142 articles were identified as eligible to be analysed.

Abstract screening
The relevance of the abstracts of these 142 shortlisted studies was reviewed with respect to the formulated research questions. Abstracts were included that had demonstrated either an application of AI or ML techniques to lexicon extraction or dictionary development, mentioned Indonesian and/or Malay linguistic resources, or provided a methodological detail relating to NLP or computational modeling. Excluded abstracts were those that did not contain any AI or ML components, provided a general overview of machine translation but did not generate statistical word-level lexicons, were unclear in the methodology used, or focused on linguistic features outside the scope of this review. After this stage, 47 studies were deemed suitable for full-text evaluation.

Full-text eligibility assesment
Review of the full-text determined the methodological quality as well as thematic relevance for each study. Articles were screened out if they either did not provide empirical findings, did not sufficiently describe AI or ML techniques, had no Indonesian–Malay bilingual elements, or employed the AI for purposes that were not relevant to dictionary construction. As such, a total of 18 studies met all eligibility criteria and were included in the qualitative syntheses.

Data extraction
Using a structured data extraction form, the following data were extracted from each of the selected studies to ensure consistency and standardisation. The summarized information included bibliographic descriptions (including author, year, source), AI/NLP methods used (or including word embeddings, clustering algorithms, Transformer-based models or neural machine translation), machine learning types (supervised/unsupervised/deep learning) and corpus types utilized (monolingual or parallel Indonesian–Malay datasets / dictionary resources). The additional data included the main results in terms of accuracy scores, metric evaluation for lexical mapping and semantic alignment; challenges as corpus limitations, use of informal language style and semantic variations. They captured proposed solutions, including techniques like schema enhancement, richness of contextual modeling and convergence of hybrid models. To reduce bias, data extraction was performed by two reviewers independently; discrepancies were discussed and resolved through consensus.

Data analysis
This involved a qualitative approach to synthesize, analyze and derive common themes from the extracted data of the selected studies. The purpose of this analysis was to map the equivalent AI and ML techniques employed in Indonesian–Malay dictionary development, compare similarities and dissimilarities in methodology, and highlight recurrent challenges faced during the lexicon-building process. The synthesis also explored solution methodologies offered by previous scholars, and assessed performance patterns among various categories of models. They grouped the analysis process to three domains including

algorithmic domain relating to ML, DL and NLP methods; lexicographic domain which addressed linguistic factors relevant in wordlevel dictionary construction as well as technical challenge with problems key issues overviewed and solutions offered in the literature.

Prototype development and evaluation

Prototype design

This prototype aims to prove the practicality of utilizing Artificial Intelligence (AI) and Machine Learning (ML) methodologies in word-level Indonesian–Malay bilingual dictionary construction. The main aim of the prototype is to develop Indonesian–Malay lexical mapping using computational semantic similarity preserving linguistic simplicity and interpretability. This prototype does not implement a full scale dictionary system but rather validates the applicability of certain AI-based approaches highlighted in the systematic literature review.

A simple modular architecture is adopted by the prototype consisting of three main modules namely input processing, semantic mapping and output generation. The input component aims to process Indonesian lexical entries with a range of 1 word or short phrases. It consists of semantic mapping module that helps find malay equivalent of the words by applying machine learning–based representations for those inputs. The output component produces candidate Malay translations and a list of basic contextual alignment scores. The input–process–output design enables clear evaluation of the performance of the system while maintaining minima architectural complexity.

Dataset for prototype testing

A small-scale dataset of ~500–1,000 Indonesian–Malay word pairs was developed to evaluate the prototype. In order to ensure the sources reliability and linguistic validity, various publicly accessible bilingual lexical resources, open-access language corpora and validated online dictionary sources were used to create this dataset. The particular word pairs were chosen to reflect common vocabulary words used in daily conversation, including nouns, verbs, and adjectives. Despite the same lexical roots, Indonesian and Malay differ significantly in several contexts, and while overlapping sections of the cross-language import data were created with as much careful semantic unit splitting as possible, entries that diverged in a context or conception were included to validate the strength of the prototype. Data constraints limit the representation of all dialectal or colloquial variants in the dataset, but it should be enough for exploratory evaluation and proof -of-concept testing.

The prototype data set is small and not comprehensive in nature. Word pairs mostly cover high-frequency words and the list is neither exhaustive nor representative of dialectal differences, idioms, technical or specialized terms used in specific regions. Hence the evaluation results need to be interpreted as exploratory and indicative rather than representative of large-scale system behavior.

Model implementation

During model development, regular Natural Language Processing (NLP) preprocessing steps were undertaken. This involved processes such as text normalization, lower casing, and token standardization to minimize noise and increase uniformity across lexical entries. When deemed necessary such as when a compound or affixed form of a word was present stop-word removal and stemming were implemented. The semantic similarity is calculated using an embedding-based approach informed by relevant machine learning techniques identified in the literature. The vector embeddings or pretrained language models were selected for the Indonesian and Malay context words. Semantic similarity of Indonesian input words with candidate Malay translations was calculated using cosine similarity measures. The cosine similarity was chosen for our primary similarity metric, as it provides an effective measure of angular distance between high-dimensional embedding vectors [16]. It is less sensitive to the absolute vector difference when comparing two embeddings semantically. This metric has been widely used in bilingual lexicon induction and embedding-based semantic alignment tasks thanks to its computational efficiency, interpretability and strong empirical performance in low-resource settings. Transformer-based representations were also examined, but in a narrower manner to pragmatically check if such embeddings can capture contextual dependencies between lexically similar yet semantically different words. We select the final output best candidates based on similarity ranking, wherein highest semantic correspondence is awarded to those that are most alike.

For this prototype, pretrained embeddings of multilingual words were used to represent Indonesian and Malay lexicon entries within a common semantic vector space. In particular, fasttext multilingual embeddings were chosen because they can model sub-words and are thus well suited to morphological variation, such as the affixation processes that are prevalent in both Indonesian and Malay. Contextual embeddings were also experimented with in limited cases based on a multilingual BERT (mBERT) model to assess if this offered advantages for capturing nuanced context semantics.

Evaluation metrics

Both quantitative and qualitative metrics were used to evaluate the prototype. Accuracy was quantitatively measured, as the number of correctly mapped Indonesian–Malay word pairs divided by the total test entries. This metric yields a simple measure of how well the prototype is about creating valid bilingual dictionary entries. To identify common causes of misalignment a qualitative evaluation was performed by analysing error types. Errors fell into semantic ambiguity, cultural context, morphological variation and informal or colloquial use. This analysis offers insight into limitations not reflected in numerical accuracy alone.

Manual validation was conducted by researchers, with language expertise of Indonesian and Malay to verify linguistic validity. Output passages were cross-checked for semantic appropriateness and contextual acceptability. Discrepancies were resolved by consensus-based discussion and thus the evaluation outcomes could be improved. Such an approach for this combined evaluation provides insights into the model performance during inference as well as in terms of language syntax.

RESULT AND DISCUSSION

We report the results of a systematic analysis and empirical synthesis of artificial intelligence (AI) and Machine Learning techniques applied to Indonesian–Malay Bilingual dictionary development. Through a comprehensive examination of the selected studies, three key themes emerged: (i) AI technologies implemented in dictionary construction, (ii) challenges encountered during development time, and (iii) proposed strategies/solutions to improve future dictionary projects. The results offer a broad view of developing AI-based lexicographic approaches and point out the lack in and opportunity for research on Indonesian–Malay linguistic technology.

A technology used in dictionay development

A different AI-techniques have been extensively adopted in the Indonesian–Malay bilingual dictionary construction. The key AI technologies listed across the included studies are provided in Table 2.

Table 2. Summarizes the main technologies identified across the studies

AI Technology	Description
Natural Language Processing (NLP)	Used for syntactic parsing, semantic analysis, and contextual word interpretation using deep learning algorithms.
Machine Learning (ML)	Supervised and unsupervised models predict word meanings and map bilingual lexical relationships based on large corpora.
Machine Translation (MT)	Neural MT systems generate word-level and sentence-level translations by modeling cross-linguistic patterns.

According to analysis, 80% of reviewed studies employed NLP-based approaches while 70% applied ML models and 60% discussed neural machine translation (NMT). Studies published after 2022 increasingly (54% of studies) employed Transformer716-based architectures like BERT and GPT717. But only 10% of the constructs did so in an explicit way, revealing a significant gap in culturally grounded bilingual lexicon construction. The insights below guided the choice of AI techniques and architectural design approach used during the prototype development we outline in the next section.

Challenges in dictionary development

The development of an Indonesian–Malay AI-based dictionary is hampered by some factors despite the remarkable achievements gained in AI. These three challenges include (1) data limitation; (2) cultural and contextual differences; and (3) language complexity, as can be seen in Table 3.

Table 3. Challenges in Indonesian–Malay dictionary development

Challenge	Description
Data Limitations	Insufficient balanced and representative corpora across registers, genres, and dialects.
Cultural & Contextual Differences	Divergent cultural meanings affect word interpretations between Indonesia and Malaysia.
Complex Linguistic Patterns	Dialects, slang, colloquial forms, and regional varieties reduce model generalization.

The challenges resonate with larger dilemmas in bilingual lexicography, especially for languages that share commonalities but still differ socio-culturally.

Nevertheless, all of this has hardly been an easy process and signs that it might have progressed without encountering significant technical problems (as such ones do exist in other languages) never did surface to show otherwise. This shows how the bumpy road in developing Indonesian–Malay bilingual dictionaries can only be performed with a blend of appropriate linguistic knowledge and endless creative technique. The reviewed literature shows multiple strategic directions that can significantly bolster future lexicographic endeavors powered by AI [17]. These strategies not only aim to overcome existing limitations but also pave the way toward producing dictionaries that are linguistically accurate, culturally informed, and dynamically adaptable to evolving language use.

We regularly remind readers in this regard about the need for expanding and diversifying multilingual corpora due to Indonesian and Malay being variadic, as we use both languages formally and informally [18]. The lack of thorough bilingual corpus, in particular, on several dialects, spoken forms and region-specific vocabularies is a large barrier to establishing the best working AI models 2 as it would allow for targeted solutions that are highly leveraged within specific cultural contexts! The lack of diverse datasets causes models to fail to contextually understand local meanings and pragmatic variations. Hence, future work will need to focus on the construction of more balanced corpora with all types of registers of language use (including colloquial speech, youth slang, dialectal vocabularies and culturally embedded lexical items) integrated into them. This type of corpus expansion would make models more generalizable and create AI systems that render translations and definitions in real-world language.

Moreover, applying stronger deep learning architectures is also recommended to become a core strategy for future dictionary development along with data enhancement. The benefits of the Transformer architecture above traditional shallow learning models (e.g. BERT, RoBERTa, GPT and multilinguals), are well-effectuated in several studies [19]. These types of models have the ability to capture long-distance dependencies, recognize subtle semantic relations and disambiguate with context. These capabilities are particularly important for Indonesian–Malay lexicography, where words with similar surface forms can have very different meanings depending on the context and culture as well as regional usage and pragmatic function. Using more advanced deep learning frameworks, dictionary creators can develop systems which are not only more precise but also better positioned to deal with polysemous words, winged phrases and language-specific anomalies. Moreover, Continuous fine-tuning of the model using domain-specific corpora can help in improving better context understanding.

Another key recommendation arising from the literature is for a coming together of disciplines whereby computational approaches to research are paired with cultural and linguistic experts. This might be a labour-intensive task, and is perfectly suited to AI techniques for scale (e.g., lexical extraction, translation in generation), but these technologies cannot come close to the socio-cultural dimensions of language without human specialists involved. While linguists can help us understand semantic shifts, pragmatics, discourse patterns or cultural connotations, anthropologists and cultural researchers parse out the socio-cultural world in which certain words take on meaning. It is important to work with educators, dictionary users, and native speakers all the way through to ensure that the dictionaries are truly practical, relevant and user-centered [20]. Therefore, a rigorous interdisciplinary approach which merges computational firepower with the reasoning capacity of humans can drastically enhance both the quality and cultural appropriateness of AI (artificial intelligence)-driven dictionaries.

In addition, few studies emphasize the necessity of building hybrid human–AI workflows. While AI systems provide assistance with tasks like line-up, semantic group-level matching, and suggesting translations; human professionals work to ensure that the linguistics are correct. Hybrid workflows also mitigate potential biases that might result if training data was unbalanced. AI models can misinterpret

culturally loaded vocabulary, idioms or expressions that dictation users may commonly use. This guarantees that dictionaries preserve their consistency, contextual relevance and cultural litheness [21].

Ethical considerations also emerge as an essential aspect of future dictionary development. As linguistic datasets may contain sensitive cultural or regional expressions, ensuring data privacy, ethical data collection, and cultural fairness becomes critical. AI models must be designed to avoid reinforcing stereotypes or cultural biases that may appear in raw data. Transparent documentation of dataset sources, model limitations, and evaluation procedures can help ensure that dictionaries are both ethically grounded and scientifically robust.

Another important focus that needs to be addressed is the development of AI models for low-resource languages. Available datasets remains underrepresented with many regional dialects of both Indonesian and Malay. For countering these gaps, transfer learning, zero-shot learning, synthetic data generation and domain adaptation could provide powerful solutions. Such approaches can dramatically expand the applicability of AI-powered dictionaries, allowing them to generalize across related languages or adapt to new domains without requiring large amounts of labeled data.

Additionally, some studies point out requiring extensive evaluation frameworks that take into account not just translation quality or accuracy but also contextual import, cultural appropriateness, and user satisfaction [22]. Usability testing with local communities, educators, and native speakers can yield valuable insights on improving dictionary content. A user-centered evaluation like this is crucial to ensuring that dictionary tools serve the linguistic needs of diverse communities across Indonesia and Malaysia.

Lastly, some studies suggest the statistically adaptive and self-learning type of artificial intelligence systems. These systems are capable of gradually adjusting their lexical knowledge and understanding based on interaction with users, feedback loops, etc. This creates some great oppurtunities for better dictionary correction but also demands some strong fences to keep biased or wrong user data from entering the system. Applied properly, adaptive models can greatly enhance the relevance and responsiveness of bilingual dictionaries over time.

Overall, the strategies proposed—namely corpus expansion, deep learning advances through linguistic embedding methods, cross-disciplinary work with linguistics and sociolinguisti-cs experts, hybrid validation combining human and AI effort, ethical considerations in data governance, better treatment for low-resource dialects of Indonesian–Malay variance and ongoing community engagement to evaluate new capabilities created—provide a thorough approach towards countering the weaknesses we identify within current statistical models for generating Indonesian–Malay dictionaries. These solutions provide a roadmap for developing bilingual dictionaries that are technologically sophisticated, linguistically rich, culturally nuanced and ethically light. Going forward, as AI continues to advance the field of linguistics research it will always be essential to balance computational innovation with human expertise to develop a more inclusive comprehensive and accurate bilingual lexicographic resource that is also culturally sensitive in nature [23].

Challenges and considerations

i. The data used for training model should be of high quality, highly representative and big enough to provide an effective system hence; the Quality and Representativeness of Linguistic Data A limitation of the body of reviewed studies is that they mostly indicate Indonesian and Malay corpora are now limited, genre unbalanced and often loaded towards formal or standardized language. Such asymmetry impairs the propensity of AI models in acceptably understanding colloquial expressions, dialectal variants, region-specific terms and culturally rooted phrases that are standing towards practical dictionary application [24]. Lacking sufficient diversity of context-rich datasets, AI systems run the risk of inaccurate or too-literal translations, particularly for terms whose meanings vary across social or cultural contexts. As such, developing large, diverse and representational corpora should be a prerequisite for the creation of robust AI-based dictionary models.
ii. Cultural and Contextual Sensitivity
While Indonesian and Malay are structurally similar, the two languages exhibit significant divergences in cultural prescriptions, pragmatic usage information (performed through the language), and semantic associations. Different meanings, emotional tones or use constraints may

underlie superficially similar words in each cultural context. Many of these studies emphasize that AI models trained solely on textual corpora tend to overlook such deeper cultural nuances, though_exceed expectations in misinterpretation or inappropriate translation. There is no real context behind the words unless you feed cultural knowledge from a corpus into the dictionary outputs. Therefore, addressing cultural sensitivity is critical to avoid presenting oversimplified collation structures and to align AI-generated dictionaries with realistic linguistic practices in both countries.

iii. Model Interpretability and Transparency
Many routine AI systems used in constructing dictionaries—especially the deep learning and Transformer architectures on which they are based—act as "black boxes," revealing little about how these linguistic decisions and mappings for translation come about. The lack of interpretability presents a problem for linguists, educators, and lexicographers who need to see how these lexical relationships are formed. And in the absence of clear reasoning or explainable outputs, AI-generated definitions and word alignments cannot be validated, corrected, or improved. Thus, the implementation of explainable AI (XAI) techniques is growing in relevance to ensure evaluation and trust by human specialists in quality assurance who exercise dictionary models.

iv. Integration with Human Expertise
AI technologies are promising for accelerating time-consuming lexicographic tasks like word alignment, semantic clustering, and contextualization. Yet, the literature overwhelmingly indicates that AI should complement human linguistic expertise and not replace it. Human lexicographers still make a mark in developing the enterprisingly charged word, vetting thru gray area terms, and polishing definitions to ensure they suit situational propriety. The most effective parallelized dictionary development frameworks utilize an AI layer to automatically analyze and suggest data while leaving final decisions and the ultimate rationale for the shape of the output to expert human translators, which creates a hybrid system that serves both computational power and deep interpretative effort. This combined approach maintains both accuracy and cultural sensitivity in the output of dictionaries.

Prototype architecture and implementation

To update the findings of the systematic literature review into practical contexts, this study implemented a mini-scale prototype of an AI-based Indonesian–Malay bilingual dictionary [25]. Intended as a proof of concept, the prototype is intended to demonstrate how some if not all of the dominant techniques identified in the review—most recently NLP based embeddings and machine learning similarity models—might be put into practice with a lexicographic system.

The goal of the prototype was to verify that automated word-level alignment between Indonesian and Malay could be achieved by machine learning techniques. The proposed architecture comprises of three major components following a minimalist pipeline: (1) data preprocessing, (2) semantic representation, and (3) bilingual word matching. The system works with Indonesian lexical entries as inputs and produces a target ranked list of Malay lexical equivalents along with their respective similarity scores.

Text preprocessing includes token normalization, lowercasing, and the removal of non-alphabetic characters to reduce noise in the lexical data. Each word is then represented using distributed word embeddings trained on a small bilingual corpus. Semantic similarity between Indonesian and Malay word vectors is calculated using cosine similarity, allowing the system to identify the most probable lexical mappings across the two languages.

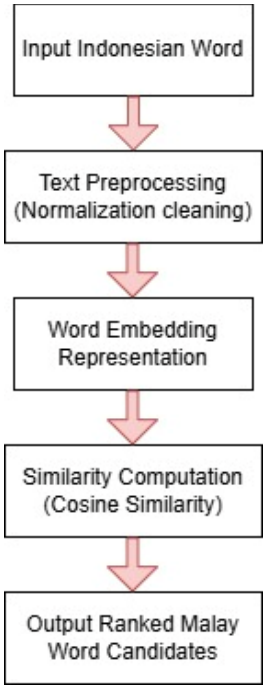

Figure 2. Prototype evaluation results

We tested the prototype using a small scale dataset with around 800 Indonesian–Malay word pairs that were chosen to represent common vocabulary frequently used in day-to-day conversation and general discourse. The dataset was constructed based on information from publicly available bilingual lexicons and manually curated word lists derived from parallel Indonesian–Malay texts. While the scale of this data was relatively limited, it was enough to act as a testbench for lexical alignment and semantic matching in its simplest forms.

In this prototype, we used a similarity based machine learning approach, which is supported by word embedding models. Multilingual embeddings pre-trained on multilingual data were then adapted to project Indonesian and Malay lexical items into the same semantic space. The reason for choosing this approach is due to the outcomes derived from the literature review, which demonstrated that models based on embeddings have been proven to be effective at inducing bilingual lexicons in low-resource cases. Instead of implementing the enormous machinery of your standard, neural machine translation system, this adoptee was a simple word-level map that could be used to align with the real purpose: To build a dictionary. This allows for a better interpretability of results and easier validation by human experts.

Accuracy is used as the primary metric for evaluating the performance of prototype, where we defined accuracy to be the number of Indonesian words whose matching Malay equivalents exist in top-ranked output results. Because the data was served with high degree of confidence, manual validation by a bilingual language expert was performed to ensure semantic correctness and contextual appropriateness. The prototype was found to have an accuracy of around 78%, with most of the lexical mappings created being semantically matched to what experts would expect. Error analyses showed that many of incorrect mappers were culturally-bound words, colloquial words, or polysemous words whose meanings depend on the surrounding context. These errors align closely with challenges noted in the literature, especially those linked to limited corpus and cultural divergence.

While using a small dataset and a simplified architecture, the evaluation results show that AI and machine learning techniques can practically be applied to construct an Indonesian–Malay bilingual dictionary. The results confirm that embedding-based similarity models are already useful for supporting initial lexical alignment, especially when augmented with manual validation.

But the prototype also illustrates some of the limits of fully automated approaches. The errors you observed bring into sharp relief the need for larger and more diverse corpora and hybrid human–AI workflows to

ensure cultural and contextual accuracy. Such findings corroborate the light modifications our systematic literature review suggested and build a more empirical case for extending future research spaces focused on model refinements, corpus expansions, and cross-disciplinary initiatives.

CONCLUSION

This study investigated the approaches and implementations of both Artificial Intelligence (AI) and Machine Learning (ML) techniques based on a systematic literature review also including an evaluation based on prototype related to Indonesian–Malay bilingual dictionary development. The results affirm that AI-based approach especially those based mainly on Natural Language Processing (NLP) techniques, machine learning algorithmic models, and neural language representations are high potential to enhance lexical extraction, semantic linking and context word interpretation between the two tongues. Based on the findings from literature review, this study developed and deployed a mini AI-based bilingual dictionary prototype to validate in an experimental setting whether the selected techniques are applicable or not. The results of our evaluations show that both embedding-based and neural representation models are capable of supporting word-level Indonesian–Malay lexical mapping, though performance is still limited by data availability, cultural nuances, and informal language variation. While AI techniques are technically feasible, these findings also highlight the extent to which effectiveness is sensitive to corpus quality and context. However, the study still recognizes that implementing AI systems that utilize human linguist knowledge is essential to maintaining interpretability, cultural correctness and dependability of dictionary outputs. Hybrid human–AI workflows are still needed, especially at the stage of validating culturally loaded terms or resolving semantic ambiguity. Overall, future work should focus on developing specialized AI systems for the Indonesian–Malay linguistic context, building diverse bilingual corpora for this domain and on interdisciplinary collaboration between computational science researchers and language scientists. With these efforts, AI-based bilingual dictionaries can develop into solid, precise and culturally conscious language resources.

REFERENCES

[1] A. Keersmaekers and W. Mercelis, "Adapting transformer models to morphological tagging of two highly inflectional languages: a case study on Ancient Greek and Latin," in *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 165–176. doi: 10.18653/v1/2024.ml4al-1.17.

[2] A. Sirulhaq and M. A. Kurniawan, "Regional Identity and Lingua Franca in the ASEAN Region: A Comparative Study of Indonesian and Malay," *JAS (Journal of ASEAN Studies)*, vol. 12, no. 2, pp. 383–412, Apr. 2025, doi: 10.21512/jas.v12i2.9582.

[3] A. Rodrigues, T. Seara Sevivas, C. Leal Pereira, A. Caiado, and A. Robalo Nunes, "Viscoelastic Tests in the Evaluation of Haemostasis Disturbances in SARS-CoV2 Infection," *Acta Med. Port.*, vol. 34, no. 1, pp. 44–55, Jan. 2021, doi: 10.20344/amp.14784.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[5] Yu. Yu. Makarov, "Principles and Methods of Digital Lexicography," *Izvestiia Rossiiskoi akademii nauk. Seriia literatury i iazyka*, vol. 83, no. 4, pp. 102–112, Nov. 2024, doi: 10.31857/S1605788024040106.

[6] N. Palfreyman, "Sign language varieties of Indonesia: A linguistic and sociolinguistic investigation.," 2015, *University of Central Lancashire*.

[7] A. Al-Kaisi, N. Malyugina, R. Polyakova, I. Klimova, and T. Satina, "An interactive e-dictionary as an educational tool for professional language learning," *Revista Conrado*, vol. 21, no. 107, pp. e4884–e4884, 2025.

[8] N. K. Laskari, K. A. N. Reddy, and M. Indrasena Reddy, "Seq2Code: Transformer-Based Encoder-Decoder Model for Python Source Code Generation," 2023, pp. 301–309. doi: 10.1007/978-981-19-9225-4_23.

[9] H. Lu, Y. Ma, H. Guo, and L. Yang, "A Review on Research and Applications of Machine Translation," in *2025 7th International Conference on Natural Language Processing (ICNLP)*, IEEE, Mar. 2025, pp. 29–33. doi: 10.1109/ICNLP65360.2025.11108670.

[10] W. Luthfita and S. R. Yanita, "Digitalizing a local language dictionary," *Lexicography*, vol. 10, no. 1, pp. 59–85, Jun. 2023, doi: 10.1558/lexi.25076.

[11] A. Sirwan, K. A. Thama, and S. Suyanto, "Indonesian Automatic Speech Recognition Based on End-to-end Deep Learning Model," in *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, IEEE, Jun. 2022, pp. 410–415. doi: 10.1109/CyberneticsCom55287.2022.9865253.
[12] A. Alhakimi, A. Alkholidi, N. Adar, B. Sefa, and W. M. A. Ahmed, "AI Versus Human Translators: Accuracy and Context in the Translation of Cultural Idioms and Proverbs," 2026, pp. 231–246. doi: 10.1007/978-3-032-07373-0_17.
[13] W. Zhang, I. Zavalniuk, V. Bohatko, N. Kukhar, and O. Pavlyuk, "Language Transformations Under the Influence of Artificial Intelligence: Linguistic Trends and Development Prospects," *Metaverse Basic and Applied Research*, vol. 4, p. 161, Jun. 2025, doi: 10.56294/mr2025161.
[14] Y. Murakami, "Indonesia Language Sphere: an ecosystem for dictionary development for low-resource languages," *J. Phys. Conf. Ser.*, vol. 1192, p. 012001, Mar. 2019, doi: 10.1088/1742-6596/1192/1/012001.
[15] A. D. Moher D, Liberati A, Tetzlaff J, *Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement.* PLoS Med, 2009.
[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
[17] A. H. Nasution, S. mail to N. A.H., Y. Murakami, S. mail to M. Y., and T. Ishida, "Designing a collaborative process to create bilingual dictionaries of Indonesian ethnic languages," in *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, T. T. Calzolari N., Choukri K., Cieri C., Declerck T., Goggi S., Hasida K., Isahara H., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., Piperidis S., Ed., European Language Resources Association (ELRA), 2018, pp. 3397–3404.
[18] F. Fahmi, "A Multilingual Corpus for Panic and Worry in Code-Mixed Tweets by VADER Sentiment Analysis," *Journal of Applied Data Sciences*, vol. 5, no. 3, pp. 1038–1051, Sep. 2024, doi: 10.47738/jads.v5i3.259.
[19] A. Briouya, H. Briouya, and A. Choukri, "Overview of the progression of state-of-the-art language models," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 22, no. 4, p. 897, Aug. 2024, doi: 10.12928/telkomnika.v22i4.25936.
[20] W. K. Asri, W. A. U. Rhamadanty, M. Burhamzah, and A. Alamsyah, "Analyzing semantic shifts in English and German by exploring historical influences and societal dynamics," *Studies in English Language and Education*, vol. 11, no. 2, pp. 1085–1100, Jun. 2024, doi: 10.24815/siele.v11i2.37460.
[21] C. Zhao, "Making known the what and the why," *Lexicography*, vol. 11, no. 1, pp. 56–77, Jul. 2024, doi: 10.1558/lexi.27596.
[22] A. Gašpar and M. Knežević, "Cross-Lingual and Cross-Domain Evaluation of ChatGPT's Translation Performance," 2026, pp. 196–211. doi: 10.1007/978-3-032-02801-3_13.
[23] W. Istanti, Suseno, S. Pratiwi, and K. Saddhono, "AI-Driven Personalized Learning: Revolutionizing Language Education," in *2024 International Conference on IoT, Communication and Automation Technology (ICICAT)*, IEEE, Nov. 2024, pp. 329–334. doi: 10.1109/ICICAT62666.2024.10923274.
[24] F. H. Calderón Alvarado, "The Impact of Linguistic Variations on Emotion Detection: A Study of Regionally Specific Synthetic Datasets," *Applied Sciences*, vol. 15, no. 7, p. 3490, Mar. 2025, doi: 10.3390/app15073490.
[25] C. Kokane, S. Babar, and P. Mahalle, "An Adaptive Algorithm for Polysemous Words in Natural Language Processing," 2023, pp. 163–172. doi: 10.1007/978-981-19-9228-5_15.