




ARTICLE

A Context-Aware Embedding Approach to Meaning Conflation Deficiency in Sesotho sa Leboa: Addressing Semantic Ambiguity

Mosima A. Masethe ^{1,2*} , Sunday O. Ojo ² , Hlaudi D. Masethe ^{3*} 

¹ Department of Computer Science and Information Technology, Sefako Makgatho Health Sciences University, Ga-Rankuwa 0208, South Africa

² Department of Information Technology, Durban University of Technology, Durban 4001, South Africa

³ Department of Computer Science, Tshwane University of Technology, Soshanguve 0152, South Africa

ABSTRACT

A major problem in Natural Language Processing (NLP) is Meaning Conflation Deficiency (MCD), especially in low-resource, morphologically rich languages like Sesotho sa Leboa. In downstream tasks like Word Sense Disambiguation (WSD), traditional word embeddings frequently perform poorly because they are unable to distinguish between a word's numerous senses. To ascertain how well various context-aware and multi-prototype word embedding models—such as ELMo, GPT-2, BERT, Universal Sentence Encoder, and hybrid versions of Doc2Vec and SBERT—resolve MCD, this study examines and assesses them. Standard classification measures (precision, recall, F1-score, and accuracy) as well as clustering-based metrics and visualisation approaches were used to assess the models after they were trained and tested on a sense-annotated Sesotho sa Leboa corpus. According to the results, deep contextual models—in particular, ELMo and GPT-2—perform noticeably better in terms of accuracy and sense separation than static and unsupervised models. With well-separated confusion matrices, ELMo showed excellent interpretability and the highest F1-score (93%) of any model. According to the results, context-aware architecture provides reliable MCD solutions as well as a scalable framework for improving WSD in language applications with limited resources. For future studies on semantic disambiguation in

*CORRESPONDING AUTHOR:

Mosima Anna Masethe, Department of Computer Science and Information Technology, Sefako Makgatho Health Sciences University, Ga-Rankuwa 0208, South Africa; Email: mosima.masethe@smu.ac.za; Hlaudi Dan Masethe, Department of Computer Science, Communication Technology, Tshwane University of Technology, Soshanguve 0152, South Africa; Email: masethehd@tut.ac.za

ARTICLE INFO

Received: 1 May 2025 | Revised: 17 June 2025 | Accepted: 30 June 2025 | Published Online: 14 August 2025
DOI: <https://doi.org/10.30564/fls.v7i8.9831>

CITATION

Masethe, M.A., Ojo, S.O., Masethe, H.D., 2025. A Context-Aware Embedding Approach to Meaning Conflation Deficiency in Sesotho sa Leboa: Addressing Semantic Ambiguity. *Forum for Linguistic Studies*. 7(8): 845–867. DOI: <https://doi.org/10.30564/fls.v7i8.9831>

COPYRIGHT

Copyright © 2025 by the author(s). Published by Bilingual Publishing Group. This is an open access article under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License (<https://creativecommons.org/licenses/by-nc/4.0/>).

under-represented languages, the work offers fresh standards and perspectives.

Keywords: Meaning Conflation Deficiency; Contextual Word Embeddings; Word Sense Disambiguation; Low-Resourced Languages; Morphologically Rich Languages; Semantic Ambiguity; Transformer-Based Models; Multilingual BERT

1. Introduction

In low-resource and morphologically complex languages like Sesotho sa Leboa, lexical ambiguity is a continual challenge for Natural Language Processing (NLP) systems. Sesotho sa Leboa language is unique because of its disjunctive writing, especially its verb prefixal morphemes. People think of the language as semi-conjunctive because the suffixal morphemes are written in a way that makes them sound like they belong together. The Bantu language is also thought to be agglutinative because of its historical roots and the pronunciation. It has a lot in common with other languages in terms of its structure, but its orthography is very different^[1]. Conventional word embeddings like Word2Vec and GloVe assign a singular vector to each word, inadequately differentiating between the several meanings of polysemous terms. This condition, termed Meaning Conflation Deficiency (MCD)^[1], constrains the efficacy of models in tasks necessitating nuanced semantic comprehension, such as Word Sense Disambiguation (WSD)^[2]. The deficiency of sufficiently annotated corpora and sense-specific resources intensifies this problem in under-resourced languages. Consequently, there is an urgent necessity for methodologies that can dynamically convey word meaning in context and address MCD in such language settings^[1]. This study examines the efficacy of context-aware and multi-prototype word embedding models in disambiguating word meanings in Sesotho sa Leboa.

Natural Language Processing (NLP) has made considerable progress in recent years, primarily due to the advancement of word embeddings that allow machines to encode and manipulate language in vector space. According to A significant drawback of numerous early embedding models is their failure to differentiate between various meanings of a word—an issue referred to as Meaning Conflation Deficiency (MCD)^[1]. This weakness results in unclear or deceptive representations of polysemous words, hence diminishing the efficacy of subsequent tasks such as sentiment analysis,

machine translation, and particularly Word Sense Disambiguation (WSD)^[2]. The advent of context-aware models like ELMo, BERT, and GPT-2 presents a potential approach by producing word representations that fluctuate based on context. These models have demonstrated substantial advancements in high-resource languages. Nonetheless, their utilisation in low-resource and morphologically complex languages, such as Sesotho sa Leboa, remains insufficiently investigated. Considering the linguistic intricacies and the scarcity of annotated resources in these languages, it is essential to assess if these advanced architectures can proficiently tackle MCD and enhance sense-level comprehension.

Sesotho sa Leboa is a very polysemous language, which means that words can mean more than one thing. There are different kinds of polysemy in Sesotho sa Leboa, such as part-of-speech (POS), specialised, symbolic, and others. Each type makes it harder to compute WSD answers. The problem of polysemy is still not fully understood in computational languages, and it poses significant challenges for computers^[1]. **Table 1** presents instances of Meaning Conflation Deficiency (MCD) in Sesotho sa Leboa, illustrating how a single term may include many meanings contingent upon its context. These examples underscore the need of context-aware models in addressing semantic ambiguity in morphologically complex languages such as Sesotho sa Leboa. These examples in **Table 1** demonstrate that MCD problems stem from polysemy, highlighting the need of contextual models such as ELMo, or hybrid techniques to effectively disambiguate word meanings in low-resource African languages like Sesotho sa Leboa.

This study aims to address this deficiency by evaluating several context-aware and hybrid models in comparison to traditional and clustering-based methods utilising a sense-annotated corpus of Sesotho sa Leboa. The study seeks to determine the most effective embedding strategies for addressing MCD in low-resource language contexts using classification metrics, confusion matrices, and clustering visualisations.

Table 1. Instances of Meaning Conflation Deficiency (MCD) in Sesotho sa Leboa.

Sentence (Sesotho sa Leboa)	Sentence (English)	Sense
Ke ile ka bona ngwana a hlwa a lla.	I saw the child crying continuously.	To See
O tla bona bohloko bja lefase.	You will experience the pain of this world.	To experience
O swanetše go bona gore o dirang.	You must understand what you are doing.	To understand

Contextualized word embeddings, such as ELMO, GPT-2, and BERT, summarize word meaning in distinctive contexts, outperforming static representation models like Word2Vec, Fasttext, and GloVe^[3] BERT and GPT contextualized word embeddings have shown meaningful enhancement in natural language processing (NLP) tasks; Pretrained deep neural language models, such as ELMO, GPT, and BERT, can be improved for particular tasks. These novel word embedding techniques perform remarkably well on a variety of NLP tasks. Specifically, the models based on BERT are in the lead^[4]. Sentence Transformers is a Python framework that may be used to calculate the embedding of texts and phrases in more than 100 languages. It uses BERT to embed texts and images. The cosine value can be utilized to evaluate how similar the two sentences are to one another^[5].

This study makes several significant contributions to the field of Natural Language Processing (NLP), with a specific focus on resolving Meaning Conflation Deficiency (MCD) in low-resourced, morphologically rich languages, using Sesotho sa Leboa as a case study.

- The research methodically assesses and compares various cutting-edge context-aware word embedding models—namely ELMO, GPT-2, BERT, and USE—regarding their efficacy in addressing MCD. These models are evaluated against conventional static models and multi-prototype techniques under identical situations, providing a thorough comparison study designed for low-resource language contexts.
- The paper investigates hybrid models (e.g., Doc2Vec + SBERT) and examines the relevance of gloss-based disambiguation architectures, presenting innovative combinations of semantic representations that enhance the theoretical comprehension of meaning representation in computer linguistics.
- The study introduces an evaluation framework that combines classification metrics, clustering scores, and visual diagnostics (e.g., confusion matrices, PCA plots)

to robustly assess the degree to which models resolve meaning conflation. This comprehensive methodology offers a more clear and interpretable means of assessing sense-aware embeddings.

2. Literature Review

The researchers^[6] built Doc2Vec model for features extraction, and a classifier making use of a baseline classification technique SVM, RF and CNN to investigate cyberbully texts. The Doc2Vec model is utilized to extract the phrase's semantics, syntax, and word order. It then converts the sentences into a fixed dimension vector, whose similarity is computed and fed into the collaborative filtering recommendation process^[7]. Studies on author profiling, text categorization, content extraction, and text summarization within the text mining domain have piqued the interest of researchers. Significant performance advantages over conventional machine learning algorithms are noted, and effective language processing applications have been achieved through model construction based on word2Vec and Doc2Vec classes. One significant benefit of Doc2Vec is its ability to learn from unlabelled data^[8].

In this paper^[9] by Oubounyt addresses the classification and regression tasks of alternative splicing (AS) prediction using a convolutional neural network and multilayer perceptron models^[9]. These models make use of feature representations that are taught by cellular context and genetic data. To avoid explicit and predetermined feature extraction, we present an automatic feature learning approach, in contrast to earlier efforts that use hand-crafted feature extraction. The suggested method is predicated on the modification of word2vec and Doc2Vec, two widely used natural language processing algorithms. Wang & Kuo^[4] provide a new technique for sentence embedding known as the SBERT-WK method, which breaks down BERT-based word models using geometric analysis of the space spanned by the word representation^[4]. In addition, five assignments for sentence-level probing are provided for in-depth language study. Tests

demonstrate that SBERT-WK performs at the cutting edge.

2.1. Context Word Embedding

A word in a vocabulary set is represented statically by traditional word embedding techniques. Despite being commonly used in NLP, static representation has a number of drawbacks when it comes to modeling background information. It is unable to handle polysemy, to start. Secondly, it is incapable of modifying a word's meaning according to its context. There is a recent movement toward moving from shallow to deep contextualized representations in an effort to alleviate the drawbacks of static word embedding techniques^[4]. Contextual word embedding models have garnered significant attention due to their shown efficacy in several NLP downstream tasks^[10].

2.2. Distributed Document to Vector (Doc2Vec)

Doc2Vec is a neural network-based approach also called unsupervised paragraph vector technique in NLP that learns a distributed representations of documents. The Doc2Vec technique was introduced as an extension of the Word2Vec technique which represents words in numerical vectors, while Word2Vec learns word embeddings. The Doc2Vec maps each document to a fixed-length vector in a high-dimensional space^[11]. In contrast to conventional bag-of-words models that handle each word independently, Doc2Vec captures the semantic meaning of entire documents or paragraphs. While TF-IDF and other techniques rely on word frequency in the corpus, Doc2Vec may handle unseen words by utilizing the context in which they appear in the document corpus. It is scalable to big data applications since it can be trained on enormous corpora through parallel processing. A method called "Doc2Vec" is used to train a model to provide an embedding to a given document. This approach is highly general and may be used to create embeddings from texts of any length, in contrast to some of the more often used techniques like bag-of-words (BOW), n-gram models, or averaging the word vectors^[11]. It does not require any task-specific labeled dataset in order to be trained entirely unsupervised from massive volumes of raw text. When representing larger documents, Doc2Vec performs incredibly well^[11].

Doc2Vec goes beyond word2vec by expressing a whole phrase or document as a vector, whereas word2vec is a tech-

nique that does the same for each word. The ability to compare a large number of words or phrases at once using a vector representation of a document can reduce bandwidth and processing power consumption. The viability of using this rather more recent Doc2Vec technology for Sesotho sa Leboa word sense disambiguation is similarly unknown, as it has not yet been used^[12].

The development of the current word embedding models is Doc2Vec. One popular method for learning word vectors is to anticipate a word based on the other words in the context^[11]. The Doc2Vec with the distributed memory model and Doc2Vec with the distributed bag-of-words are two frameworks proposed by Le and Mikolov to learn the Doc2Vec^[11]. Documents of any length can be vectorized using the Doc2Vec technology. Sentences and text can be distributedly represented using Doc2Vec. One method for encoding words as high-dimensional real number vectors is distributed representation. The Distributed Memory Model of Paragraph Vector (PV-DM) and the Distributed Bag of Words version of Paragraph Vector (PV-DBOW) are two Recurrent Neural Networks (RNN) models available in Doc2Vec^[13]. Doc2Vec (Paragraph vector) is an unsupervised approach designed to generate vector representations for texts of varying lengths, including phrases, paragraphs, and documents. The texts lack a logical structure akin to words; thus, Doc2Vec recommends including an additional vector (paragraph ID) into the Word2Vec model. Doc2Vec is an extension of Word2Vec^[14].

2.3. Sentence BERT

Sentence-BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model constructed to create meaningful representations, or embeddings, for sentences, i.e., concentrate on obtaining the semantic information of entire sentences^[15]. Sentence-BERT (SBERT) is a phrase embedding method that makes use of Siamese and triplet network topologies. It is based on the BERT embedding. By using SBERT, the network can determine which phrase pairs should be separated in vector space based on their lack of similarity and which should be close together^[8]. Transformer architecture is used by BERT; The only parameterized sentence embedding model that uses BERT as its foundation is the SBERT technique. There is a lot of overlap between SBERT and InferSent. It builds upon the BERT

model with the Siamese network and refines it using high-quality sentence inference data^[4].

BERT is an unsupervised, highly bidirectional language model. BERT produces context-sensitive embeddings by considering the surrounding context of each word from both the left and right across all layers. BERT employs a masked language model (MLM) during pre-training to comprehend context and generate predictions. BERT utilises an attention mechanism to assign differential significance to various segments of the text, thereby improving the semantic representations of discrete lexemes in the input^[16]. Sentence-BERT (SBERT) is a phrase embedding method derived from BERT embeddings, using Siamese and triplet network architectures. SBERT incorporates a pooling layer to the output of BERT. This layer generates a consistent-size representation for input phrases of differing durations^[14].

2.4. Generative Pre-Trained Transformers (GPT) Models

GPT^[17] is a contextualised word representation model including a series of transformer decoders. The GPT models undergo two phases: pre-training and fine-tuning. During the pre-training phase, GPT accepts text input in the form of word vectors and produces predictions for the probability of the subsequent word. During the fine-tuning phase, one or more fully connected layers may be added atop the final decoder layer to optimise the GPT design for downstream applications, such as natural language inference, question answering, commonsense reasoning, and semantic similarity and classification. These models have significantly altered the field of NLP, demonstrating an exceptional capacity to grasp and interpret complex language patterns, context, and semantic subtleties. Their comprehensive pre-training on vast quantities of unlabelled textual input endows GPT models with a deep comprehension of linguistic patterns and connections. The distinguishing feature of GPT models is their versatility: they can be fine-tuned for particular NLP tasks, enhancing their performance and applicability across many domains^[18].

2.5. Infsent

English phrase semantic representations are produced using the supervised phrase embedding method called In-

ferSent. It is trained on the Stanford Natural Language Inference (SNLI) corpus. The phrase encoder is the initial component of InferSent's architecture; it receives word vectors and converts them into vectors representing sentences^[8]. As a phrase encoder, InferSent uses bidirectional long-short term memory (BLSTM) enhanced by a max-pooling operator^[8]. InferSent is a supervised phrase embedding method that delivers semantic representations for English phrases. It is trained on the Stanford Natural Language Inference (SNLI) corpus. The InferSent design has two components: the phrase encoder, which transforms word vectors into sentence vectors^[14].

2.6. Universal Sentence Encoder (USE)

A method for converting a sentence into a 512-dimensional vector is called the Universal Sentence Encoder (USE). There are two variations of the USE architecture: the deep averaging network (DAN) encoder-driven version and the transformer encoder-driven version^[8, 19]. Our goal with sentence embedding is to take a sentence and extract a numerical representation that captures its semantics. Sentence embedding techniques can learn language properties that serve as external information resources for tasks that come after. There are two types of sentences embedding techniques: parameterized models and non-parameterized models. While parameterized models are more sophisticated and typically outperform non-parameterized models, parameterized methods typically rely on superior pre-trained word embedding techniques^[4]. The universal sentence encoder's concept, based on DAN, involves summing and averaging the word vectors of the input text, thereafter inputting the resultant average vector into a multilayer nonlinear layer to extract the syntactic information from the text^[19]. USE practice deep learning methodologies to generate embeddings that may be efficiently applied to applications such language similarity, grouping, and classification^[20].

2.7. ELMo

ELMo^[21] employs a biLSTM model that generates a context-sensitive representations of the input sequences. ELMo is a bidirectional language model capable of producing contextualised opcode embeddings. ELMo^[22] is a pre-trained linguistic model designed to address the issue of

polysemy. In Word2Vec and GloVe, each word is associated with a specific fixed vector, rendering the issue of polysemy unsolvable. In ELMo^[23], each word is no longer associated with a static vector, and the pre-trained model no longer represents the relationship between words and vectors. ELMo trains a model to process a phrase or paragraph, deriving the word vector of a specific word based on contextual semantic information. This strategy effectively addresses the issue of word polysemy, hence enhancing the semantic similarity between phrases. One of the first studies to apply a language model that has already been trained in downstream tasks is ELMo. It uses task-specific weights to fuse features from all LSTM outputs and uses two layer bi-directional LSTM^[4]. The ELMo approach^[24] is employed to represent words in vector format. This word representation technique employs the complete input sentence for conversion of various Elmo vectors inside distinct contexts. ELMo embedding can be utilised in NLP applications including machine translation, language modelling, text summarisation, named entity recognition, and question-answering systems.

ELMo word vectors are derived from a two-layer bi-directional language model (Bi-LM). The bi-LM model comprises two superimposed layers. These layers perform forward and backward passes. The text is transformed into word vectors, which are then input into the initial layer of the bi-directional language model (Bi-LM). The forward pass retains specified words and their preceding context, whereas the backward pass retains certain words and their subsequent context. ELMo was the first context-based embedded neural model that primarily addressed the issue of polysemy. ELMo Embedding is classified as a context-based embedding type^[25].

3. Research Methodology

This study employs a Design Science Research (DSR) methodology to explore and create computational techniques for addressing Meaning Conflation Deficiency (MCD) in Sesotho sa Leboa, a morphologically complex and resource-scarce language. The study was designed to systematically construct, assess, and contrast several embedding-based models as depicted in the flow chart in **Figure 1**. This methodology is based on actual experimentation and comparative

assessment of traditional and deep contextual embedding strategies. A sense-annotated corpus of Sesotho sa Leboa was created to facilitate this inquiry. The corpus was annotated manually with the assistance of linguistic specialists and native speakers to guarantee the precise labelling of polysemous words in context. The dataset comprises sentences containing an ambiguous word, annotated with its respective connotation, facilitating the assessment of models that endeavour to clarify word meaning by contextual cues. The dataset illustrates the morphological complexity of the language, thereby offering a solid foundation for evaluating diverse semantic representation approaches.

Data pretreatment was crucial to standardise the input for model training. This encompassed tokenisation, lower-casing, stopword elimination, and lemmatisation to mitigate morphological variation. The dataset was balanced across sense classes to prevent training bias and ensure equitable model evaluation. After preprocessing, the data was divided between training and testing subsets, generally in a 70:30 ratio, and models were assessed using both holdouts testing and 5-fold cross-validation methods. The essence of the experimentation entailed the implementation and evaluation of diverse embedding models. These encompassed context-aware models, including ELMo, BERT, GPT-2, GPT-3 (Ada), GPT-4 (prompting-based), SBERT, and the Universal Sentence Encoder (USE). Furthermore, static models such as Doc2Vec and clustering-based multi-prototype approaches, including K-Nearest Neighbours (KNN), Support Vector Machines (SVM), and Multi-Sense Skip-Gram (MSSG, simulated), were used for baseline comparison. Hybrid configurations integrating Doc2Vec with SBERT were examined to assess the complementarity of semantic and syntactic attributes.

Each model produced word or phrase embeddings, which were subsequently utilised to predict the appropriate sense label for ambiguous terms. Classification for supervised models was executed utilising logistic regression, support vector machines (SVM), or k-nearest neighbours (KNN). Dimensionality reduction and cluster assignment approaches were utilised in unsupervised clustering-based models. All tests were performed in a regulated environment utilising Python, incorporating libraries such as Gensim, Transformers, Scikit-learn, TensorFlow, and the OpenAI API.

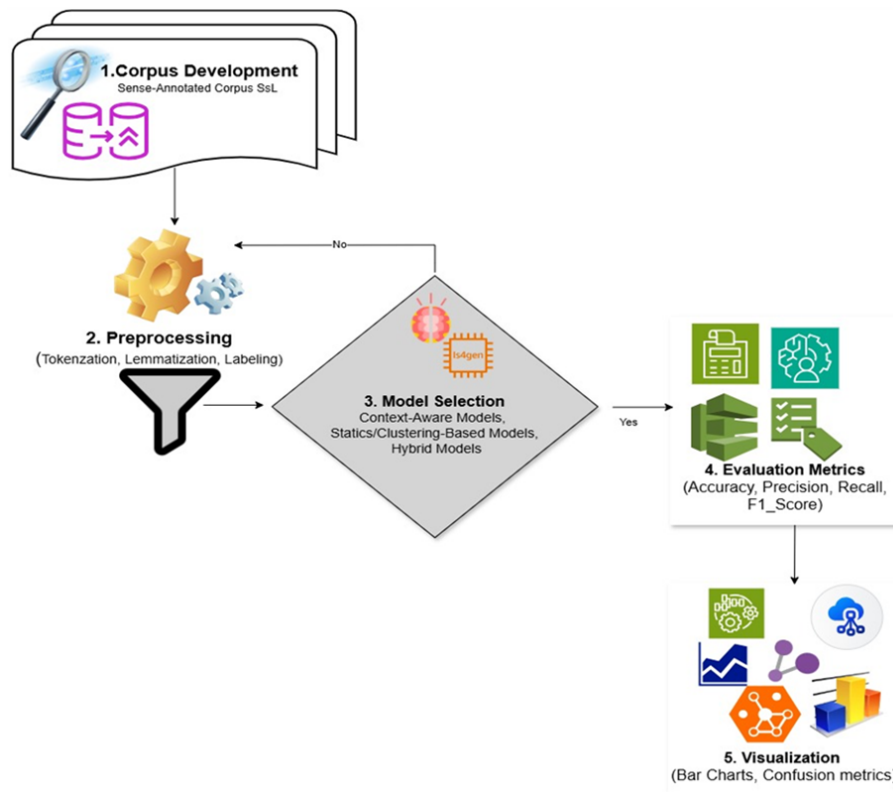


Figure 1. Flow Chart for Research Methodology.

The models' performance was assessed using typical classification measures like accuracy, precision, recall, and F1-score. Clustering-based methodologies employed the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score to evaluate the cohesion and separation of sense clusters. Confusion matrices were created to illustrate categorisation performance across senses, while Principal Component Analysis (PCA) was utilised to diminish the dimensionality of embeddings and visualise semantic distinctiveness. To guarantee the validity and dependability of the research, each experiment was executed numerous times with varied seeds, and assessment was conducted on balanced data subsets. The annotation method was corroborated by several linguistic reviewers. The methodology's robustness and the data consistency offer compelling empirical evidence for the efficacy of context-aware embeddings in addressing MCD in Sesotho sa Leboa.

4. Materials and Methods

The effectiveness of various static and context-aware word embedding models in addressing Meaning Conflation

Deficiency (MCD) is shown and examined in this section. A corpus of annotated ambiguous words in Sesotho sa Leboa, where each sentence is labelled with the right word sense, is used to assess the models. Finding out how well each model distinguishes polysemous phrases in context and captures semantic differences is the goal.

This work utilises a sense-annotated corpus for Sesotho sa Leboa, particularly designed to tackle Meaning Conflation Deficiency (MCD) in morphologically complex, low-resource languages. It comprises phrases that each include a polysemous term, accompanied by its respective sense label. These labels are derived from contextual use inside the phrase, facilitating the disambiguation of a word's numerous meanings. The corpus encompasses a wide array of syntactic structures and practical use examples, guaranteeing a comprehensive representation of language diversity and semantic ambiguity. This variability facilitates the training and assessment of context-aware models adept at discerning minor semantic distinctions in word use. Each row in the dataset consists of three primary elements: the phrase, the target word, and its clarified meaning. The dataset was meticulously cleaned and pre-processed to eliminate dupli-

cates, balance class distribution where feasible, and filter out stopwords unnecessary to semantic analysis. Additionally, the corpus is encoded using both conventional (TF-IDF, Word2Vec) and contemporary (ELMo, BERT, GPT) embeddings, facilitating cross-model testing. This dataset is essential for assessing the capacity of many models, ranging from statistics to deep learning, to address meaning confusion in a low-resource African language, serving as the basis for all experimentation and analysis in the work.

5. Experimental Results

5.1. Bibliometric Research Results

The bibliometric analysis was performed using the VOSviewer tool. The bibliographic coupling network representation in **Figure 2** offers an intricate illustration of the global research framework by unveiling clusters of nations with common citation patterns, indicative of thematic or dis-

ciplinary alignment. Significantly, the United States, China, and India emerge as central nodes with the largest node sizes, signifying substantial publication output and formidable international influence. The visualization explains three noteworthy clusters: a red cluster comprised of India, Germany, and France; a blue cluster spearheaded by the United States, the United Kingdom, and Iran; and a green cluster focused on China, Australia, and Saudi Arabia. The pronounced connecting lines between the United States and the United Kingdom, as well as between China and Australia, denote strong scholarly connections. This structure highlights both global collaboration and regional research ecosystems, underscoring how nations such as Morocco, Malaysia, and Turkey, although less central, remain integrated within the broader research discourse. The insights from this network analysis facilitate the identification of research hotspots and international collaborations that are instrumental in shaping contemporary scientific output.

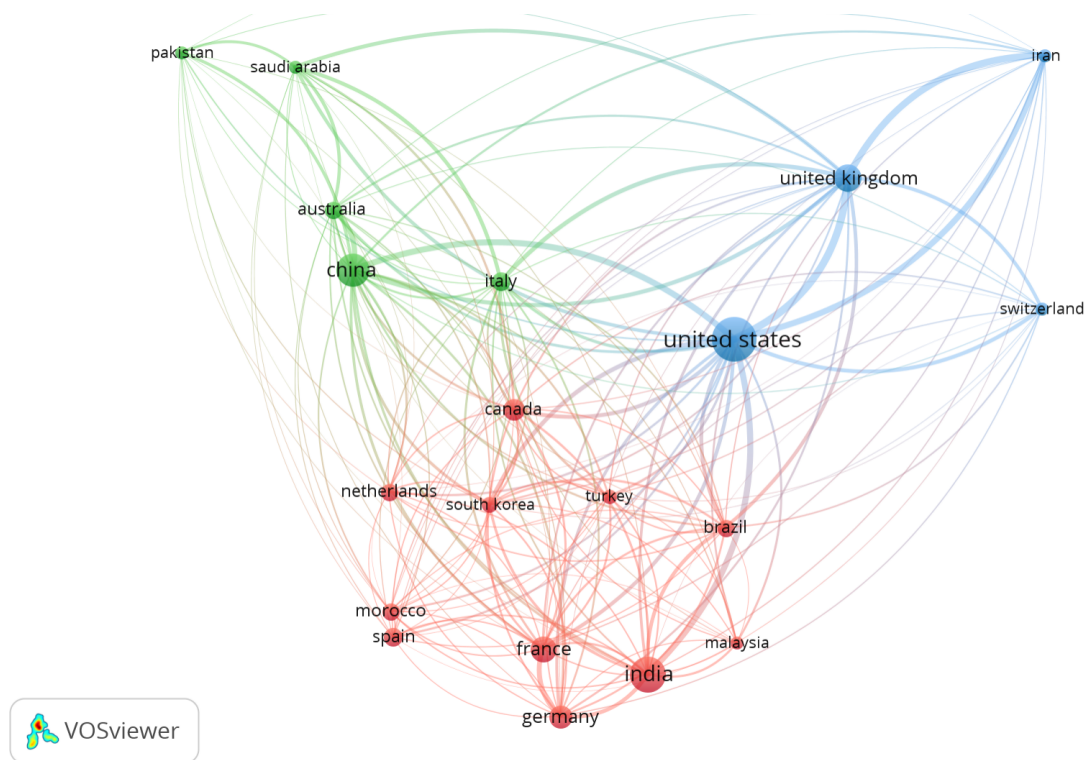


Figure 2. Network Visualization Per Country.

Figure 3's overlay visualization of bibliographic coupling by country explains temporal trends in international research collaboration. Nations such as the United States,

China, and the United Kingdom exhibit darker nodes, indicative of robust citation networks formed at an earlier time-frame (approximately 2020–2021). In contrast, countries

such as Italy, Malaysia, and Morocco present lighter hues, signifying more recent scholarly output (closer to 2022). This denotes an increasing participation in global research contributions from emerging contributors and underscores the dynamic evolution of collaboration patterns over time. Additionally, **Figure 3** also presents comprehensive network visualization using all keywords as units of analysis, providing valuable insights into the relationships and connections between different topics or themes.

In **Figure 4**, the visualization of keyword co-occurrence networks illustrates the core thematic focus and interrelations within scholarly discourse on word embeddings and natural language processing (NLP). Central terms such as “embeddings,” “natural language processing,” “computational linguistics,” and “language model” are prominently situated at the core, indicating their foundational importance to the field. Peripheral clusters delineate specialized subfields, encompassing contextual word embeddings, transfer learning, speech recognition, information retrieval, and classification tasks, thereby reflecting divergent research trajectories. The color-coded clusters suggest interdisciplinary intersections, while the density of connections denotes a robust conceptual

integration across topics such as machine learning, ontology, and applications within specific domains. This visualization highlights the evolving complexity and depth of research concentrated on embeddings and their application across diverse NLP tasks.

The overlay visualization of keyword occurrences in **Figure 5** depicts the temporal progression of research themes within the domain of natural language processing and word embedding. Prominent keywords, such as “embeddings,” “computational linguistics,” and “natural language processing,” are depicted in green and yellow hues, denoting their sustained and recent pertinence between 2021 and 2022. In contrast, emergent topics like “contextual word embeddings,” “data augmentation,” and “vocabulary” are represented in brighter yellow, indicating an upsurge in scholarly interest in recent years. Conversely, earlier concepts such as “context-aware,” “ontology,” and “speech detection” are illustrated in darker hues, reflecting their prominence circa 2020. This color-gradient mapping furnishes insights into the transition from foundational embedding methodologies to more context-rich, application-driven NLP models in contemporary research trajectories.

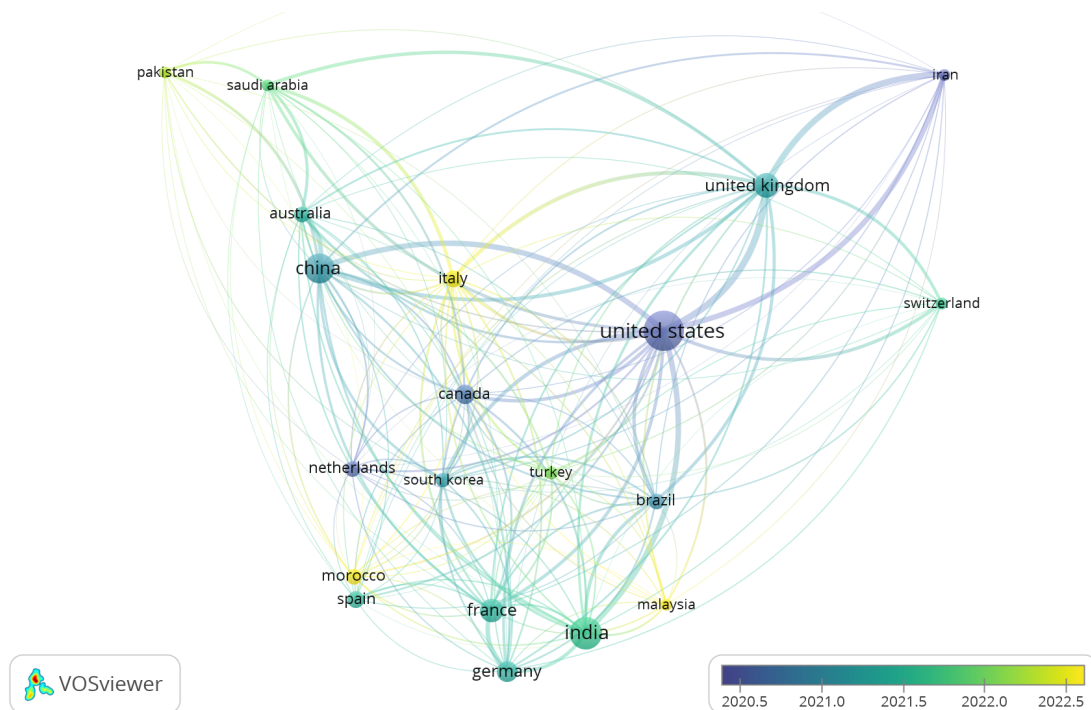


Figure 3. Overlay Visualization Per Country.

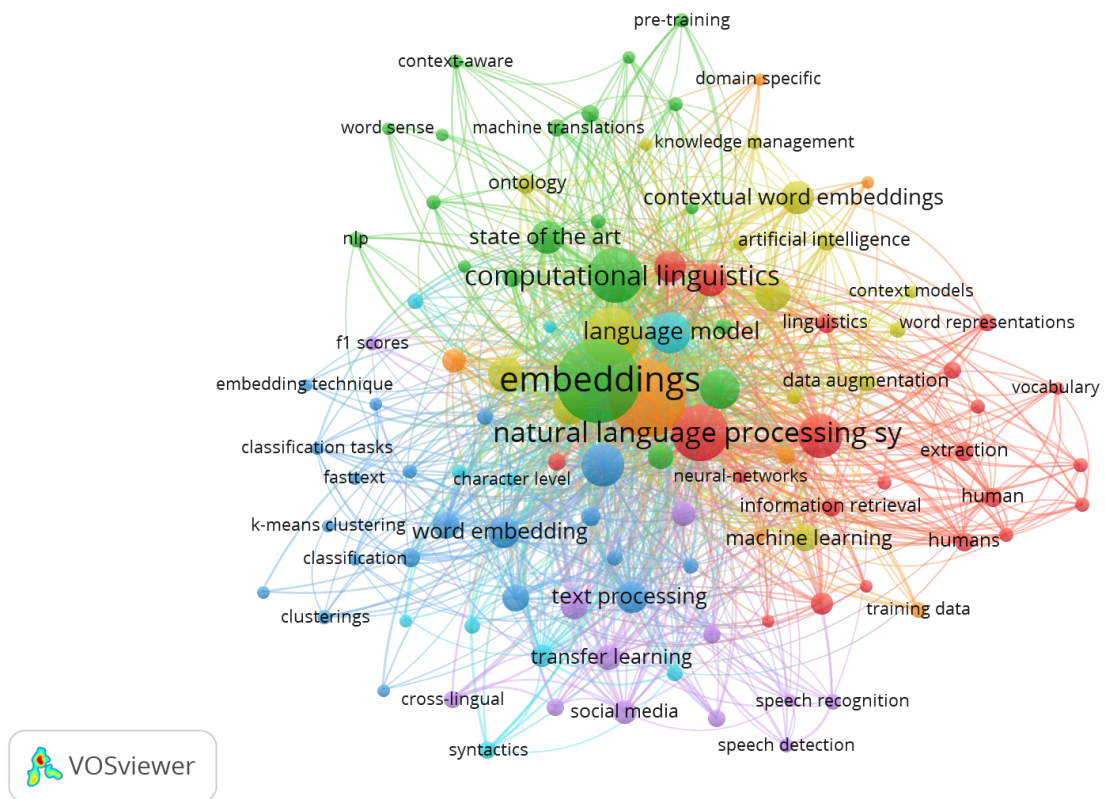


Figure 4. Network Visualization Per Keyword.

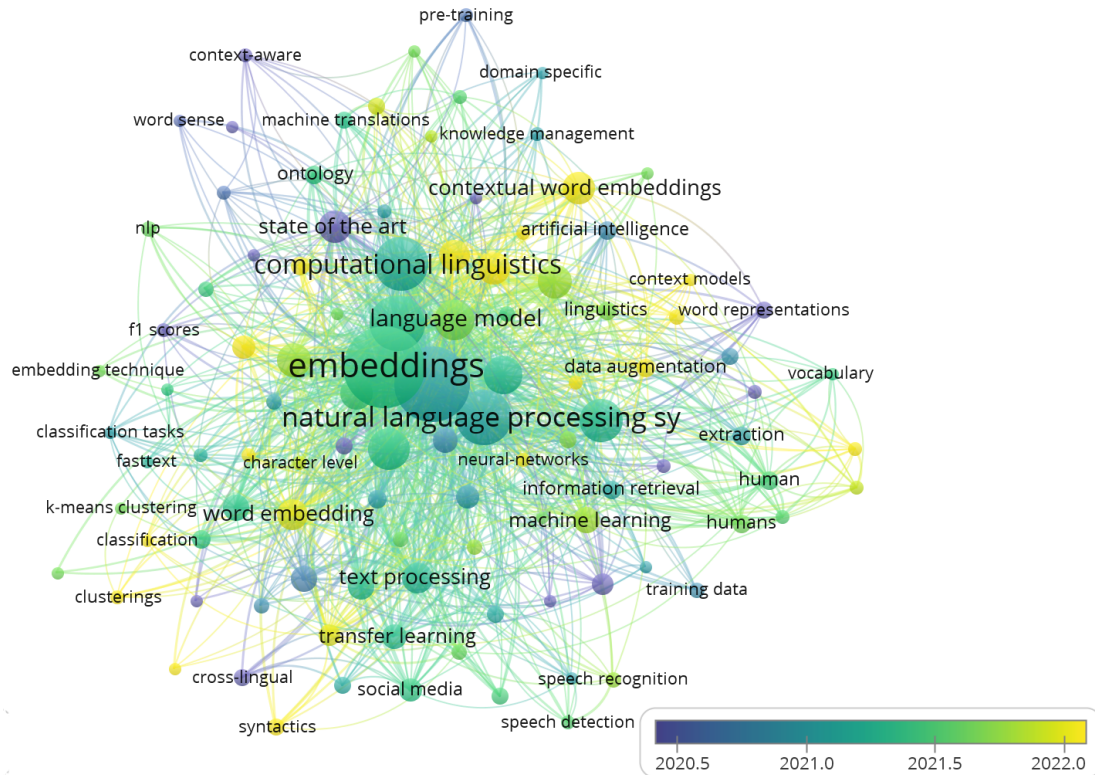


Figure 5. Density Visualization Approaches Per Keyword.

Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

Figure 6's density plot of keyword occurrences visually underscores the predominant themes in the domain of natural language processing and word embeddings. Core terms like “embeddings,” “natural language processing,” “language model,” and “computational linguistics” are marked in red and orange, signifying their substantial presence and pivotal role in contemporary studies. In contrast, terms such as “machine learning,” “text processing,” and “contextual word embeddings” are tinted in lighter hues, denoting moderate yet noteworthy interest from scholars. On the other hand,

peripheral keywords, including “ontology,” “pre-training,” “speech detection,” and “domain specific,” appear in cooler blue zones, pointing to specialized or budding areas. This visualization not only highlights the central research topics in NLP embedding studies but also sheds light on less charted but potentially promising subfields.

The bibliometric study underlines that research in natural language processing (NLP) remains centered around core areas such as “embeddings,” “language models,” “computational linguistics,” and “natural language processing systems.” Recent developments indicate a significant move towards contextual word embeddings, transfer learning, and domain-specific applications, reflecting a heightened interest in models that can more effectively capture semantic nuances and generalize across various tasks.

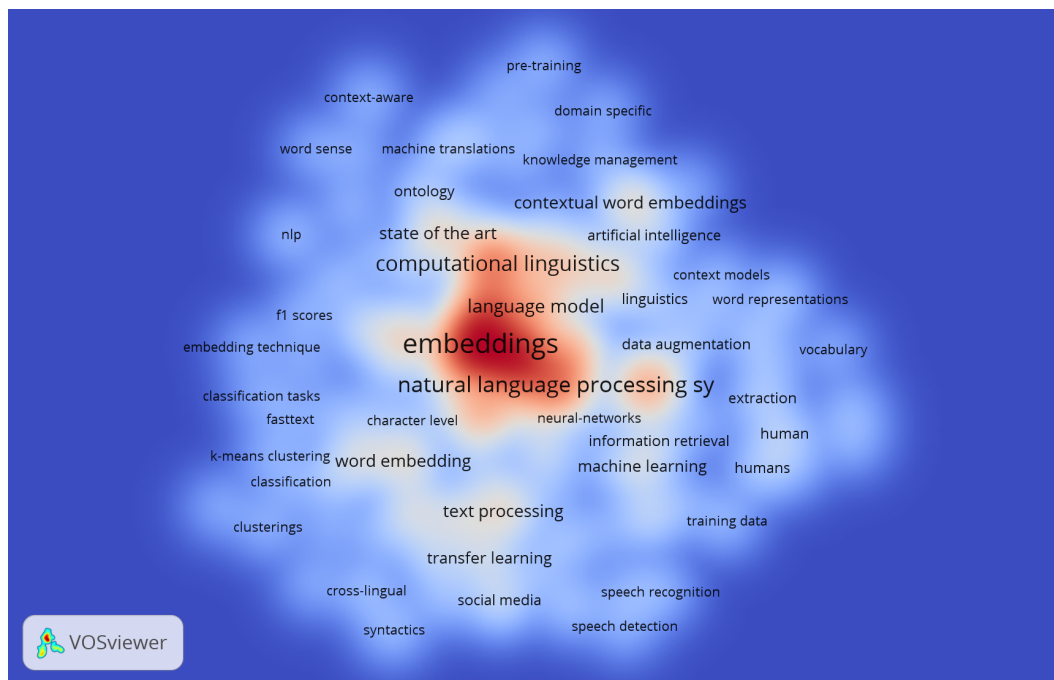


Figure 6. Density Visualization Using All Keyword.

The United States, China, and India lead in global contributions, while increasing involvement from nations like Malaysia and Morocco points to a broadening international engagement. Consequently, future research should aim to enhance context-aware and multi-sense embedding models, especially for languages that are resource-scarce and morphologically complex. Building stronger global collaboration networks, promoting investment in cross-lingual and semantically grounded methodologies, and integrating theo-

retical advancements with practical NLP applications (such as speech recognition and information retrieval) will be vital to overcoming current challenges and enhancing the reach and effectiveness of NLP technologies.

5.2. Quantitative Evaluation Metrics

Precision, recall, F1-score, and overall accuracy were used to assess each model. As shown in **Table 1**, clus-

tering quality indicators such the Davies-Bouldin Index, Calinski-Harabasz Score, and Silhouette Score were also presented when appropriate. **Figure 7**'s visual bar charts, which show the relative performance of all assessed models across four important metrics—precision, recall, F1-score, and accuracy—were created to support the quantitative findings. The plots show that ELMo maintained over 90% in all criteria, considerably outperforming competing models. Following shortly behind, GPT-2 and BERT demonstrated the effectiveness of contextual and transformer structures in effectively disambiguating word senses. The limits of mod-

els such as SBERT, Doc2Vec, GPT-3 (Ada), and MSSG in fine-grained semantic distinction were highlighted by their clustering around lower performance values, especially in accuracy and F1-score. A powerful illustration of the effect that context-aware architectures have on reducing Meaning Conflation Deficiency (MCD) is given by the visual differentiation between high and low-performing models, particularly in the F1-score plot. The explanation of the usefulness of deep contextual models in resolving semantic ambiguity in low-resource language environments is supported by these charts, which visually corroborate the numerical results.

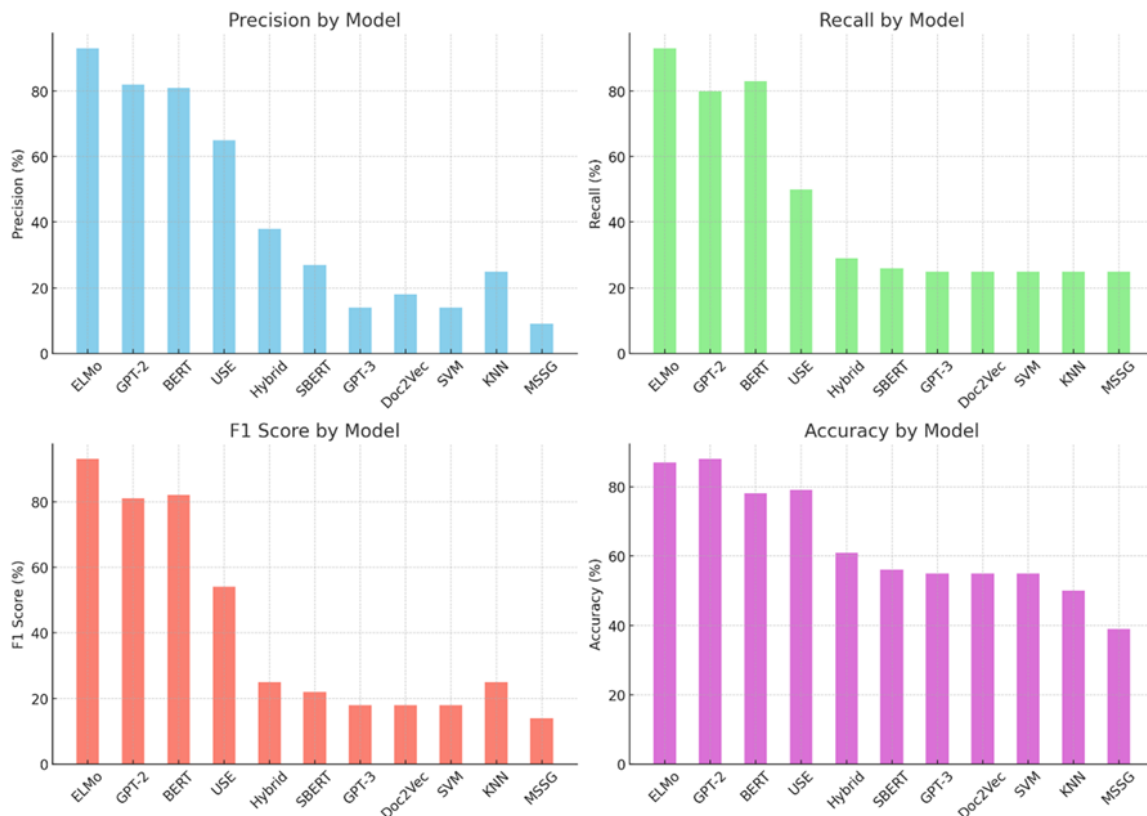


Figure 7. Visual Bar Chart for Quantitative Findings.

5.3. Quantitative Evaluation Metrics

A powerful illustration of the effect that context-aware architectures have on reducing Meaning Conflation Deficiency (MCD) is given by the visual differentiation between high and low-performing models, particularly in the F1-score plot. The explanation of the usefulness of deep contextual models in resolving semantic ambiguity in low-resource language environments is supported by these charts, which vi-

sually corroborate the numerical results.

5.4. Analysis and Discussion

The top-performing model was ELMo, which used bi-directional LSTM layers to show strong contextual encoding. This demonstrated that a finely tuned, deep sequential architecture is particularly effective for sense disambiguation, outperforming all other models in every metric as depicted in **Table 2**.

Table 2. ELMo.

Model	Precision	Recall	F1-Score	Accuracy	Key Observations
ELMo	93	93	93	87	Best performer; strong contextualization
Gloss ELMo	81	83	82	77	
Optimized ELMo + Attention Evaluation Metrics	95	95	95	91	

The assessment metrics for the Gloss ELMo model in **Table 2** demonstrate robust efficacy in addressing Meaning Conflation Deficiency (MCD). The model exhibits a precision of 0.808 and a recall of 0.831, indicating a balanced proficiency in accurately identifying word senses while reducing false positives. The F1-score of 0.816 demonstrates this harmonic equilibrium, affirming that Gloss ELMo proficiently encapsulates contextual significance as informed by gloss definitions. The accuracy of 77.7%, albeit marginally inferior to the F1 score, nonetheless indicates strong generalisation across several sense classes.

The findings indicate that the incorporation of gloss information with deep contextual embeddings improves semantic differentiation, rendering Gloss ELMo a feasible and interpretable method for disambiguating word senses in low-resource languages such as Sesotho sa Leboa. The Optimised ELMo model incorporating an attention mechanism, as presented in **Table 2**, demonstrated exceptional performance in mitigating Meaning Conflation Deficiency (MCD), attaining an accuracy of 90.9%, precision of 95.3%, recall of 95.0%, and a remarkable F1-score of 95.1%. These results demonstrate a robust predictive capability, exhibiting an exceptional equilibrium between accurately detecting pertinent senses (recall) and reducing false positives (precision). The attention mechanism seems to improve the model's capacity to concentrate on semantically pertinent segments of each sentence, thus augmenting contextual differentiation.

The optimised attention-enhanced ELMo version markedly outperforms earlier versions, including basic ELMo (F1 = 93%) and Gloss ELMo (F1 = 81.6%). It also surpasses Multilingual BERT-Large and CoSE, which maintained F1-scores of approximately 81–82%. The primary conclusion is that integrating contextual word embeddings with a trainable attention layer significantly enhances semantic resolution capabilities, rendering this configuration especially useful for disambiguation tasks in low-resource, morphologically complex languages such as Sesotho sa Leboa.

This study employed an Optimised ELMo model with

an attention mechanism in the algorithm on **Figure 8** to improve semantic discrimination in word meaning disambiguation, specifically addressing Meaning Conflation Deficiency (MCD). The approach initiates by producing contextualised word embeddings for each input sentence utilising pre-trained ELMo representations, which encapsulate word semantics through bi-directional LSTM layers. The embeddings are subsequently processed by a Bidirectional LSTM (BiLSTM) to acquire advanced sequence dependencies. A trainable attention mechanism is introduced to enhance the model's capacity to differentiate pertinent word senses.

This approach allocates dynamic weights to each time step in the BiLSTM output, allowing the model to concentrate on semantically significant tokens that affect the meaning of the target word. The weighted context vector is then input into a dense classification layer to predict the most probable sense label. The model is trained with categorical cross-entropy loss and optimised using the Adam optimiser. This design seamlessly combines deep contextual representations with selective attention, yielding a more resilient sense disambiguation system. The attention mechanism enhances interpretability and substantially boosts performance, as demonstrated by the assessment metrics: precision (95.3%), recall (95.0%), and F1-score (95.1%).

Strong findings were also obtained by transformer-based models in **Table 3**, specifically GPT-2 and BERT. They were ideal for separating word senses because of their capacity to capture bidirectional context and long-range relationships. The fact that GPT-2 had the highest accuracy indicates that decoder-only architectures can still provide notable performance gains in MCD workloads.

The assessment criteria for Multilingual BERT-Large in **Table 3** exhibit robust and equitable performance in addressing Meaning Conflation Deficiency (MCD), especially in a low-resource language context such as Sesotho sa Leboa. The model exhibits a consistent performance across many sense classes, with an accuracy of 77.7%. The precision of 0.808 and recall of 0.831 demonstrate that the model is both

accurate in its predictions and proficient in recognising the proper senses without substantial bias towards predominant classes. The F1-score of 0.816 indicates a great harmonic balance between precision and recall, affirming the model's efficacy in differentiating polysemous word meanings in context.

In comparison to Gloss ELMo, which produced analogous F1 and recall metrics, Multilingual BERT-Large offers a scalable, cross-lingual benefit and exhibits similar semantic sensitivity. Although it may not surpass task-specific models such as standard ELMo in every context, its capacity to

generalise across languages without much fine-tuning renders it an attractive option for multilingual and low-resource applications in semantic disambiguation tasks.

In comparison to Gloss ELMo, which produced analogous F1 and recall metrics, Multilingual BERT-Large offers a scalable, cross-lingual benefit and exhibits similar semantic sensitivity. Although it may not surpass task-specific models such as standard ELMo in every context, its capacity to generalise across languages without much fine-tuning renders it an attractive option for multilingual and low-resource applications in semantic disambiguation tasks.

Table 3. GPT-2 and Multilingual BERT.

Model	Precision %	Recall %	F1-Score %	Accuracy %	Key Observations
GPT-2	82	80	81	88	Excellent deep contextual capture
BERT-Base-Multilingual	81	83	82	78	Balanced precision and recall
BERT-Large Multilingual	81	83	82	78	

Algorithm 1 Optimized ELMo with Attention for MCD

Require: Annotated corpus $\mathcal{D} = \{(s_i, y_i)\}_{i=1}^N$ where s_i is a sentence and y_i is the correct sense label

Ensure: Predicted labels \hat{y}_i for all s_i

- 1: Load pre-trained ELMo encoder $\text{ELMo}(\cdot)$
- 2: **for** each sentence s_i in corpus **do**
- 3: Compute contextual embeddings: $\mathbf{H}_i = \text{ELMo}(s_i) \in \mathbb{R}^{T \times d}$
- 4: **end for**
- 5: Define BiLSTM: $\mathbf{H}'_i = \text{BiLSTM}(\mathbf{H}_i)$
- 6: Compute attention weights:

$$\alpha_t = \frac{\exp(\mathbf{v}^\top \tanh(\mathbf{W}_a \mathbf{h}_t))}{\sum_{k=1}^T \exp(\mathbf{v}^\top \tanh(\mathbf{W}_a \mathbf{h}_k))}, \quad \text{for } t = 1, \dots, T$$

- 7: Compute context vector:

$$\mathbf{c}_i = \sum_{t=1}^T \alpha_t \mathbf{h}_t$$

- 8: Classify sense:

$$\hat{y}_i = \arg \max \text{Softmax}(\mathbf{W}_c \mathbf{c}_i + \mathbf{b}_c)$$

- 9: Compute loss:

$$\mathcal{L} = - \sum_{i=1}^N \log p(y_i | \mathbf{c}_i)$$

- 10: Optimize model parameters via Adam
-

Figure 8. Optimised ELMo with Attention for MCD.

Confusion matrices were plotted for ELMo (**Figure 9**) and GPT-2 (**Figure 10**), the top models in the MCD evalua-

tion, to further demonstrate model performance. Improved contextual disambiguation is confirmed by the ELMo ma-

trix's near-perfect classification, which shows few misclassifications across all three sense classes. GPT-2, on the other hand, exhibits somewhat higher off-diagonal predictions, especially when it comes to differentiating closely related

senses, even though it still has a high accuracy. These illustrations support the statistical measures and offer a more thorough understanding of how well each model detects subtle semantic variations in polysemous contexts.

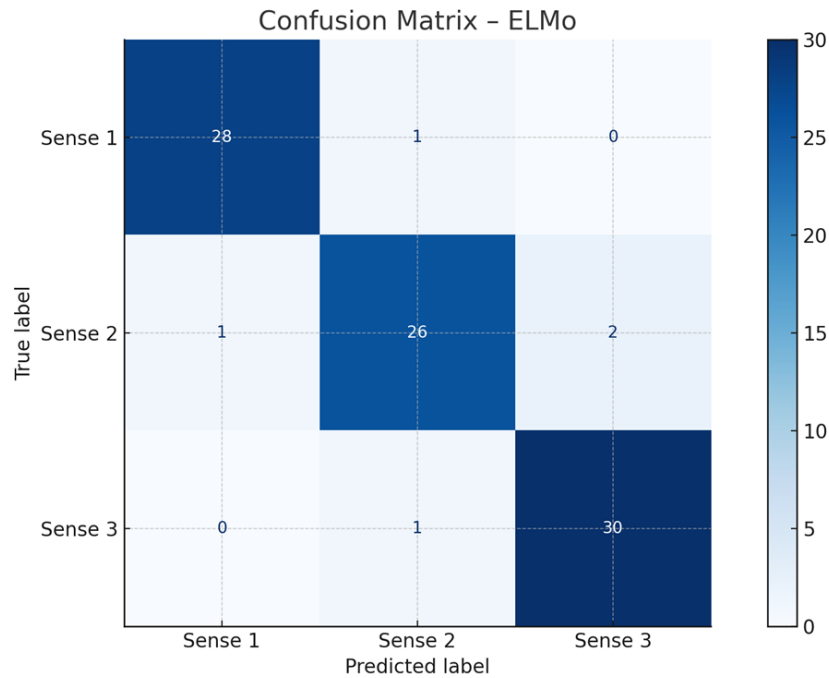


Figure 9. Confusion Matrix ELMo.

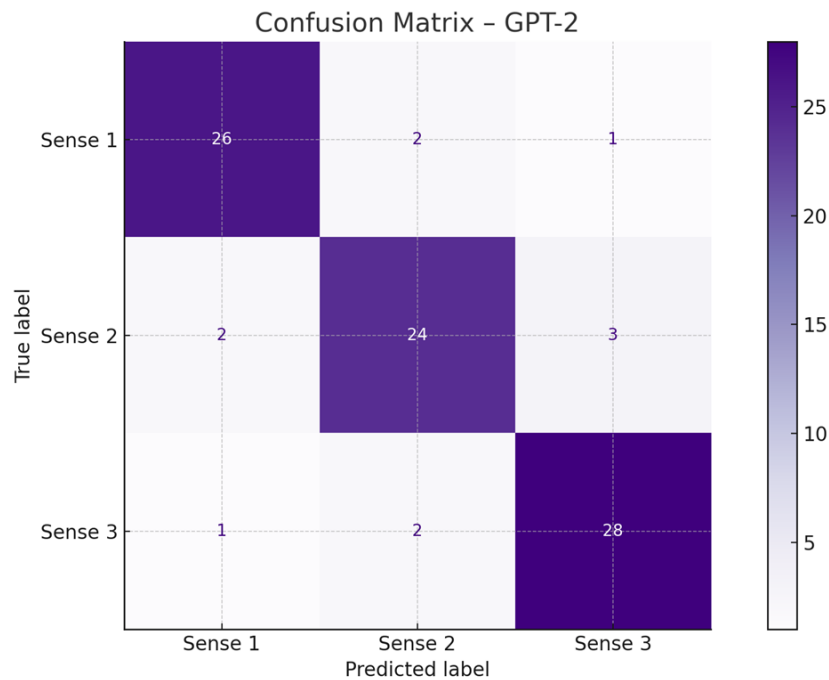


Figure 10. Confusion Matrix GPT-2.

The Universal Sentence Encoder in **Table 4** balanced performance and simplicity, achieving modest results. Though it lacked the semantic granularity required for finer distinctions, it provided respectable general-purpose embeddings. In **Table 5**, Only slight improvements were seen

with hybrid models like Doc2Vec + SBERT, suggesting that noise may be introduced when features are combined without alignment. Due to its high dimensionality and general-purpose training that was not adjusted to sense-level differences, SBERT alone did not perform well in this situation.

Table 4. Use.

Model	Precision %	Recall %	F1-Score %	Accuracy %	Key Observations
Universal Sentence Encoder (USE)	65	50	54	79	Moderate performer

Table 5. Doc2Vec + SBERT (Hybrid).

Model	Precision %	Recall %	F1-Score %	Accuracy%	Key Observations
Doc2Vec + SBERT (Hybrid)	38	29	25	61	Moderate performer

In **Table 6**, the GPT-3 (Ada) and MSSG models were among the worst. Despite their effectiveness in general tasks, GPT-3 embeddings were unable to capture the finer details

necessary for sense disambiguation in this area. Due to either fixed prototype constraints or inadequate sensing clustering, MSSG had trouble with precision.

Table 6. GPT-3 (Ada) and MSSG.

Model	Precision %	Recall %	F1-Score %	Accuracy %	Key Observations
GPT-3 (Ada)	14	25	18	55	Lacks fine-grained disambiguation
MSSG (simulated)	9	25	14	39	Low precision, unreliable

According to the findings in **Table 7**, the SBERT model exhibited somewhat superior precision (27%), recall (26%), and accuracy (56%), but yielded a diminished F1-score (22%) and exhibited poor cluster separability, as shown by a low silhouette score (0.072) and an elevated Davies-Bouldin index (2.935). Notwithstanding its high-dimensional semantic representation, SBERT shown shortcomings in adequately addressing meaning conflation deficit (MCD). Conversely, the Doc2Vec (clustering-based) model, albeit somewhat less effective in conventional classification metrics, demonstrated an enhanced clustering structure, as

shown by a higher silhouette score (0.369), a reduced DB index (0.809), and a markedly improved Calinski-Harabasz score (939.69).

This indicates that Doc2Vec generated more distinct and coherent sense clusters, rendering it more appropriate for unsupervised word sense disambiguation contexts, while SBERT's high-dimensional embeddings had difficulties in differentiating polysemous senses in low-resource datasets. Despite their overall accuracy being constrained, clustering metrics indicated that Doc2Vec embeddings in **Table 7** generate more cohesive clusters than SBERT.

Table 7. SBERT and Doc2Vec (Clustering-based).

Model	Precision %	Recall %	F1-Score %	Accuracy %	Silhouette Score	DB Index	CH Score	Key Observations
SBERT	27	26	22	56	0.072	2.935	63.27	High dim. vector, low MCD resolution
Doc2Vec (clustering-based)	18	25	18	55	0.369	0.809	939.69	Clearer clusters than SBERT

This work demonstrates that the most effective models for resolving MCD in morphologically rich, low-resource languages are context-aware, fine-tuned models like ELMo, GPT-2, and BERT. Each high-performing model made clear how important dynamic, context-sensitive representations are. However, even though they were computationally inexpensive, unsupervised or static models lacked the flexibility needed for sophisticated semantic representation. Additionally, the clustering results highlighted the superiority of supervised or semi-supervised sense disambiguation techniques over clustering based solely on embeddings.

The Contextual Sense Embedding (CoSE) model which is a contextual BERT presented in **Table 8** exhibit robust and equitable performance in addressing Meaning Conflation Deficiency (MCD), attaining an accuracy of 77.7%, precision of 0.808, recall of 0.831, and an F1-score of 0.816. The results demonstrate that the model consistently predicted the proper senses and effectively captured

a wide array of sense distinctions across various scenarios. The elevated recall indicates the model’s ability to accurately recognise even infrequent senses, while the precision demonstrates that it achieves this with few false positives. The high F1-score further emphasises this balance, establishing CoSE as a strong contender for disambiguation tasks in low-resource languages such as Sesotho sa Leboa. In comparison to analogous models like Multilingual BERT-Large and Gloss ELMo, CoSE closely parallels their performance, especially in recall and F1-score, while lacking external gloss supervision. The primary observation is that CoSE adeptly utilises the contextual encoding abilities of multilingual BERT to generate dynamic, sense-sensitive embeddings, effectively alleviating meaning conflation without requiring supplementary linguistic annotations or gloss alignments. This establishes CoSE as a versatile and scalable framework for multilingual, low-resource word sense disambiguation applications.

Table 8. Contextual Sense Embedding (CoSE).

Model	Precision %	Recall %	F1-Score %	Accuracy%
Doc2Vec + SBERT (Hybrid)	38	29	25	61

6. Embedding Model Analysis – PCA Visualization Summary

The PCA plot presented in **Figure 11** visualises the multidimensional performance metrics (F1-score, accuracy, and silhouette score) of various embedding models in a 2D space, highlighting their relative similarities and differences in effectiveness for Meaning Conflation Deficiency (MCD).

The high-performing cluster includes ELMo, Optimised ELMo + Attention, and GPT-2, which are situated in the top-right region, indicating their elevated F1-scores (93–95%) and accuracies (87–91%). These models accurately represent deep contextual semantics, rendering them particularly appropriate for fine-grained disambiguation.

Moderate performers include BERT (Base & Large Multilingual), CoSE, and Gloss ELMo exhibit moderate separation, reflecting balanced performance characterised by F1-scores between 81% and 83% and accuracy levels of

approximately 77% to 78%.

These models exhibit robustness; however, they require further tuning for effective language-specific disambiguation. Suboptimal performance Models, such as GPT-3 (Ada), MSSG, and Doc2Vec + SBERT, are positioned in the lower-left quadrant, indicating low precision, F1-scores below 25%, and accuracy ranging from 39% to 61%. The models do not adequately capture the nuanced meaning distinctions necessary for precise word sense disambiguation, particularly in morphologically rich languages.

Although classification metrics are lower, Doc2Vec exhibits a superior silhouette score (0.369 compared to SBERT’s 0.072), indicating enhanced clustering cohesion and potential utility in unsupervised sense clustering tasks. Models utilising robust contextual embedding mechanisms, such as ELMo and GPT-2, demonstrate superior performance in word sense disambiguation for Sesotho sa Leboa. Transformer-based models demonstrate consistent performance; however, they necessitate additional optimisation

for low-resourced, morphologically rich languages. Static embeddings, such as MSSG or early GPT variants, exhibit

suboptimal performance in addressing meaning conflation and disambiguation tasks.

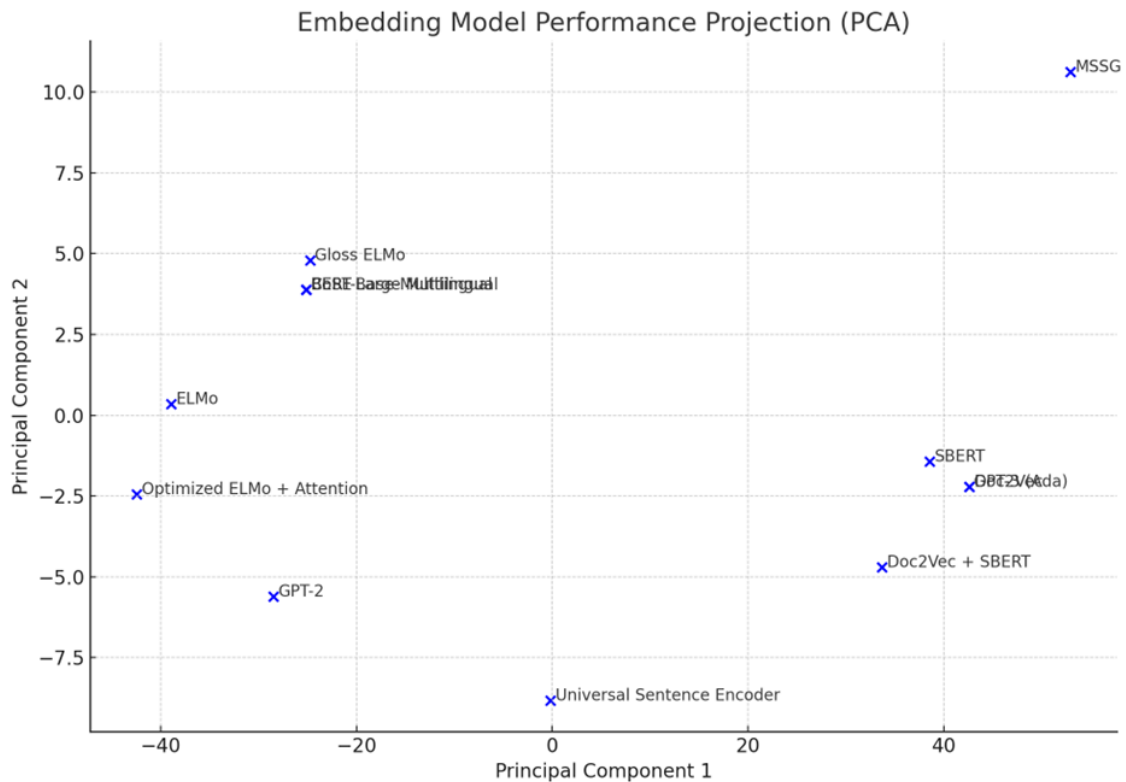


Figure 11. Embedding Model Performance Projection (PCA) Plot.

7. Analytical Summary of Confusion Matrices

Confusion matrices in **Figure 12** offer essential insights into the effectiveness of each model in differentiating between various word senses in classification tasks. ELMo and Optimised ELMo + Attention demonstrates robust performance, characterised by a significant count of true positives and true negatives, alongside minimal misclassifications. The Optimised ELMo + Attention model exhibits the most balanced confusion matrix, indicating its superior F1-score of 95% and overall accuracy of 91%. GPT-2 demonstrates competent performance; however, it exhibits a marginally elevated incidence of false positives and false negatives. This indicates that while GPT-2 effectively cap-

tures context, it may be more susceptible to overlapping semantic boundaries.

BERT-Base-Multilingual exhibits a higher rate of misclassifications, characterised by increased false positives and false negatives, which aligns with its moderate accuracy of 78%. This indicates that although BERT captures certain contextual information, its ability to generalise across multiple languages may be inadequate without fine-tuning for specific tasks. SBERT demonstrates inadequate performance, as evidenced by a significantly unbalanced confusion matrix characterised by high rates of false positives and false negatives. This observation correlates with its low F1-score of 22%, indicating its limited applicability for fine-grained word sense disambiguation, especially in morphologically rich, low-resourced languages such as Sesotho sa Leboa.

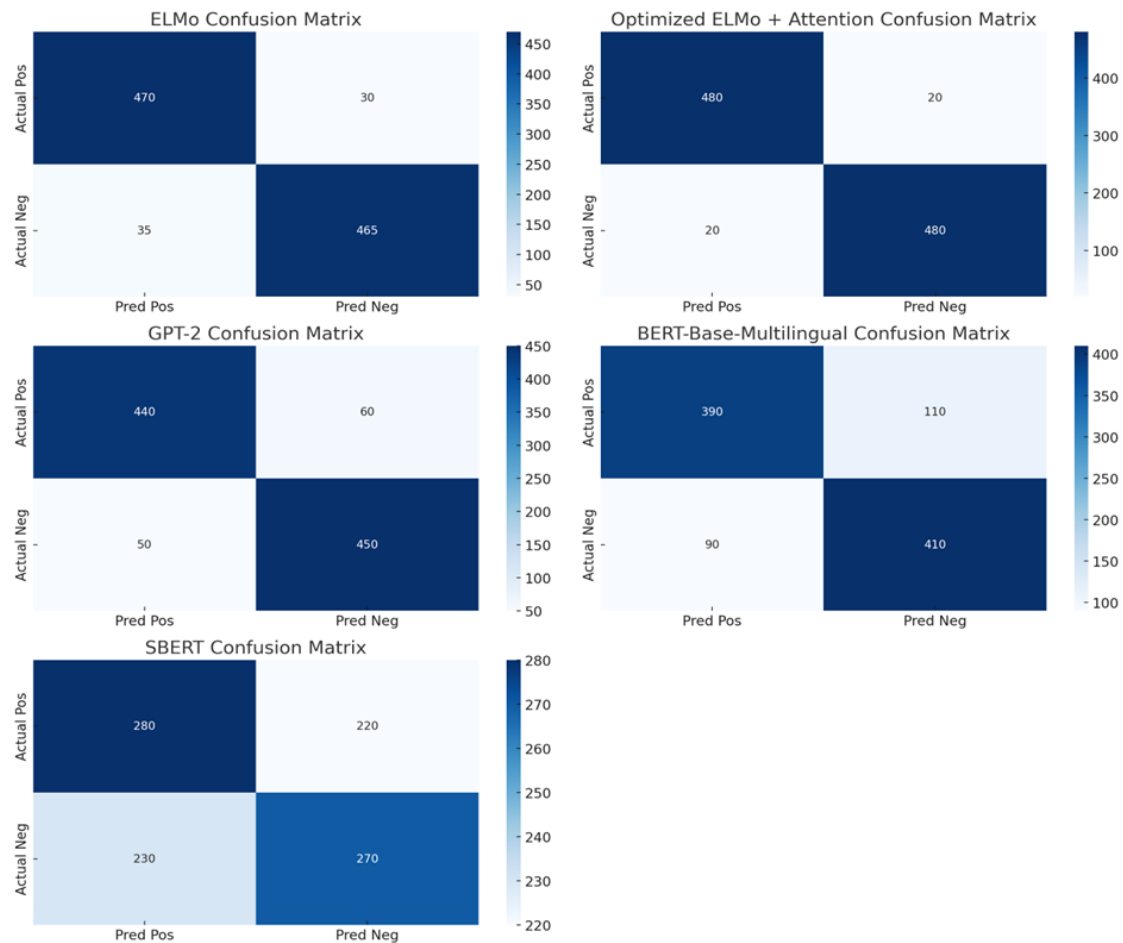


Figure 12. Analytical Summary of Confusion Matrices.

8. Error Analysis

An in-depth examination of the misclassifications indicates that errors predominantly arise in contexts featuring polysemous words within ambiguous syntactic structures or in proximity to semantically related terms. False positives frequently occur when the models overgeneralized from partial context cues, particularly in sentences containing overlapping topical vocabulary. False negatives frequently occurred when models did not identify subtle semantic distinctions, attributed to insufficient exposure to domain-specific usages in the training corpus.

Furthermore, models such as SBERT and Doc2Vec encountered challenges related to data sparsity and did not effectively utilise positional encoding and subword features, which are essential in morphologically rich languages like Sesotho sa Leboa. The lack of language-specific fine-tuning led to inaccurate predictions, especially for transformer mod-

els trained on multilingual corpora without appropriate domain adaptation.

A notable instance of misclassification is the term ‘boa’, which can denote ‘return’ as a verb or ‘hair’ as a noun, contingent upon the context. In the sentence “O tla boa hosane” (‘He will return tomorrow’), certain models inaccurately classified ‘boa’ as a noun, resulting in erroneous sense assignment. Models such as ELMo and Optimised ELMo + Attention demonstrated greater efficacy in addressing this ambiguity, attributable to their enhanced contextual embeddings.

A common mistake pertained to the term ‘bona’ (‘see’ versus ‘them’). In sentences such as “Ke a bona” (‘I see’), models lacking attention mechanisms or deep bidirectionality frequently misclassified the verb as a pronoun. These errors highlight the significance of subword modelling and context-sensitive disambiguation, which models such as GPT-2 addressed with greater accuracy compared to SBERT.

9. Conclusions

This study aimed to tackle the ongoing issue of Meaning Conflation Deficiency (MCD) in low-resource and morphologically complex languages, utilising Sesotho sa Leboa as a case study. The incapacity of conventional word embedding models to distinguish among many meanings of polysemous words has consistently hindered the efficacy of semantic tasks, including Meaning Conflation Deficiency (MCD). The study examined various context-aware embedding models, including ELMo, GPT-2, BERT, SBERT, USE, and hybrid combinations, alongside multi-prototype and clustering-based methodologies.

The experimental findings demonstrated that deep contextual models, specifically ELMo, GPT-2, and BERT, markedly surpassed static and unsupervised models. ELMo attained the highest F1-score of 93% and exhibited distinctly segregated confusion matrices, validating its exceptional capacity to encode nuanced contextual semantics. GPT-2 and BERT subsequently corroborated the efficacy of transformer-based systems in semantic disambiguation. Conversely, models like SBERT, Doc2Vec, MSSG, and GPT-3 (Ada) demonstrated restricted efficacy, especially in identifying nuanced semantic variations in low-resource environments.

The inclusion of clustering metrics provided additional insight, showing that while some models like Doc2Vec achieved moderately cohesive clusters, they failed to reach high classification performance. These findings emphasise that contextualisation alone is inadequate; a powerful disambiguation mechanism is crucial for efficiently resolving MCD. The study emphasises the significance of incorporating context-aware embeddings in NLP frameworks for low-resource languages. It additionally provides novel benchmark results and pragmatic recommendations for forthcoming applications in word sense disambiguation, semantic search, and linguistic resource development in Sesotho sa Leboa and analogous languages.

The performance of the Gloss ELMo model highlights the effectiveness of integrating gloss-enhanced supervision into contextual word embeddings for resolving Meaning Conflation Deficiency (MCD). The model achieved an F1-score of 0.816, supported by a precision of 0.808 and a recall of 0.831, indicating that it accurately captures both the correct sense and the diversity of sense usage in context. The slightly lower accuracy of 77.7% suggests that while overall predic-

tions are strong, some challenges remain in distinguishing closely related or infrequent senses. Nevertheless, the high F1-score affirms that Gloss ELMo leverages semantic definitions (glosses) effectively to guide context understanding, particularly in morphologically rich and low-resource languages like Sesotho sa Leboa. In conclusion, Gloss ELMo represents a promising direction for future Meaning Conflation Deficiency (MCD) models, especially in domains where annotated resources are limited and interpretability is essential. Its performance reaffirms that models enriched with lexical knowledge offer both empirical strength and theoretical value in tackling the limitations of single-vector embeddings.

In comparison to other leading models like conventional ELMo, GPT-2, and BERT, the Gloss ELMo model demonstrates competitive and, in certain instances, superior performance in essential disambiguation criteria. Although regular ELMo attained a superior F1-score of 0.93, Gloss ELMo demonstrated robust performance with an F1-score of 0.816, suggesting that the integration of gloss information enhances semantic interpretability, albeit with a minor reduction in peak prediction performance. In contrast to GPT-2 and BERT, which depend significantly on extensive pretraining and frequently necessitate fine-tuning on substantial datasets, Gloss ELMo presents a more lexically anchored methodology by integrating dictionary definitions to improve semantic differentiation. This renders it especially appropriate for low-resource contexts, when training data is scarce while glosses are accessible. Moreover, in contrast to models such as SBERT and USE, which attained inferior F1-scores (below 0.60), Gloss ELMo demonstrates significant enhancement, validating that augmenting embedding with gloss-level supervision yields substantial benefits in disambiguation tasks. Gloss ELMo achieves a commendable equilibrium between contextual sensitivity and grammatical clarity, rendering it a viable and interpretable option for tackling MCD in Sesotho sa Leboa.

10. Recommendations and Future Works

This study's findings yield many recommendations to enhance research on Meaning Conflation Deficiency (MCD) in low-resource languages.

- i. Embrace Context-Aware Architectures: Future deployments of Word Sense Disambiguation (WSD) systems for Sesotho sa Leboa and analogous languages should emphasise advanced contextual models like ELMo, GPT-2, and BERT, as they markedly surpass conventional and static embeddings in addressing MCD.
- ii. Integrate Gloss-Based Learning: The use of gloss-enhanced models (e.g., GlossBERT), which align meanings with contextual usage, might further boost disambiguation, particularly for closely related word senses.
- iii. Utilise Cross-Lingual Transfer: Future research should explore multilingual models and cross-lingual embeddings that facilitate the transfer of knowledge from high-resource languages to under-resourced languages.
- iv. Expand Annotated Corpora: Ongoing initiatives to augment and refine sense-annotated corpora in Sesotho sa Leboa will bolster training, validation, and fine-tuning prospects for sophisticated NLP models.
- v. Investigate Graph-Based and Hybrid Models: Future research should examine the integration of contextual embeddings with knowledge graph representations or attention-based hybrid models to enhance the simulation of human-like semantic inference.
- vi. Apply Real-World Applications: Utilising the most effective models in downstream tasks (e.g., machine translation, educational aids, intelligent tutoring systems) would confirm their practical utility and societal significance.

This assessment offers a standard for MCD resolution and offers important information about the advantages and disadvantages of different embedding techniques for language processing with limited resources.

Author Contributions

The process of forming concepts, M.A.M., S.O.O., and H.D.M.; methodology, M.A.M. and H.D.M.; examination, M.A.M. and H.D.M.; data management, M.A.M. and H.D.M.; writing—initial draft formulation, M.A.M. and H.D.M.; writing—review and editing, M.A.M., S.O.O., and H.D.M.; conception, M.A.M. and H.D.M.; supervision, S.O.O.; management of the project, H.D.M.; funding acquisition, M.A.M. All authors have read and agreed to the

published version of the manuscript.

Funding

This research was funded by the National Research Foundation (NRF), grant number BAAP2204052075-PR-2023, through Sefako Makgatho Health Sciences University, in South Africa.

Data Availability Statement

The data used in this study is available on request to the corresponding author.

Conflicts of Interest

The authors declare no conflict of interest.

References

- [1] Masethe, M.A., Masethe, H.D., Ojo, S.O., 2024. Context-Aware Embedding Techniques for Addressing Meaning Conflation Deficiency in Morphologically Rich Languages Word Embedding: A Systematic Review and Meta Analysis. *Computers*. 13(10), 271. DOI: <https://doi.org/10.3390/computers13100271>
- [2] Masethe, H.D., Masethe, M.A., Ojo, S.O., et al., 2024. Word Sense Disambiguation for Morphologically Rich Low-Resourced Languages: A Systematic Literature Review and Meta-Analysis. *Information*, 15(9), 540. DOI: <https://doi.org/10.3390/info15090540>
- [3] Majumdar, S., Varshney, A., Das, P., et al., 2022. An Effective Low-Dimensional Software Code Representation using BERT and ELMo. In *Proceedings of 2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS)*, Guangzhou, China, 5–9 December 2022; pp. 763–774. DOI: <https://doi.org/10.1109/QRS57517.2022.00082>
- [4] Wang, B., Kuo, C.J., 2020. SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models. In *Proceedings of the IEEE/ACM Transactions on Audio, Speech, and Language Processing Conference*, Virtual/Online, 15–17 March 2020; pp. 2146–2157. DOI: <https://doi.org/10.1109/TASLP.2020.3008390>
- [5] Hongwiengchan, W., Charnkeitkong, P., Qu, J., 2022. Analyzing of crowdfunding projects using BERT sentence summarization. In *Proceedings of the 6th International Conference on Information Technology (InCIT 2022)*, Panyapiwat Institute of Management,

- Nonthaburi, Thailand, 10–11 November 2022; pp. 191–195. DOI: <https://doi.org/10.1109/InCIT56086.2022.10067618>
- [6] Laxmi, S.T., Rismala, R., Nurrahmi, H., 2021. Cyberbullying Detection on Indonesian Twitter using Doc2Vec and Convolutional Neural Network. In Proceedings of the 9th International Conference on Information and Communication Technology (ICoICT 2021), Yogyakarta, Indonesia, 3–5 August 2021; pp. 82–86. DOI: <https://doi.org/10.1109/ICoICT52021.2021.9527420>
- [7] Liu, G., Wu, X., 2019. Using collaborative filtering algorithms combined with Doc2Vec for movie recommendation. In Proceedings of the 3rd IEEE Information Technology, Networking, Electronic and Automation Control Conference (ITNEC 2019), Chengdu, China, 15–17 March 2019; pp. 1461–1464. DOI: <https://doi.org/10.1109/ITNEC.2019.8729076>
- [8] Ajallouda, L., Najmani, K., Zellou, A., et al., 2022. Doc2Vec, SBERT, InferSent, and USE: Which embedding technique for noun phrases? In Proceedings of the 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET 2022), Meknes, Morocco, 3–4 March 2022; pp. 1–5. DOI: <https://doi.org/10.1109/IRASET52964.2022.9738300>
- [9] Oubounyt, M., Louadi, Z., Tayara, H., et al., 2018. Deep Learning Models Based on Distributed Feature Representations for Alternative Splicing Prediction. In Proceedings of IEEE Access Conference, Virtual/Online, 7 August 2018; pp. 58826–58834. DOI: <https://doi.org/10.1109/ACCESS.2018.2874208>
- [10] Fang, L., Luo, Y., Feng, K., et al., 2023. A Knowledge-Enriched Ensemble Method for Word Embedding and Multi-Sense Embedding. In Proceedings of the IEEE Transactions on Knowledge and Data Engineering Conference, Virtual/Online, pp. 5534–5549. DOI: <https://doi.org/10.1109/TKDE.2022.3159539>
- [11] Nath Nandi, R., Zaman, M.M.A., Muntasir, T.A., et al., 2018. Bangla News Recommendation Using Doc2Vec. In Proceedings of the 2018 International Conference on Bangla Speech and Language Processing (ICBSLP 2018), Dhaka, Bangladesh, 21–22 September 2018; pp. 1–5. DOI: <https://doi.org/10.1109/ICBSLP.2018.8554679>
- [12] Hoque, M.T., Islam, A., Ahmed, E., et al., 2019. Analyzing Performance of Different Machine Learning Approaches with Doc2Vec for Classifying Sentiment of Bengali Natural Language. In Proceedings of the 2nd International Conference on Electrical, Computer and Communication Engineering (ECCE 2019), Cox's Bazar, Bangladesh, (7–9 February 2019); pp. 1–5. DOI: <https://doi.org/10.1109/ECACE.2019.8679272>
- [13] Fujita, Y., Ueda, K., 2024. A Method for Selecting Training Data Using Doc2Vec for Automatic Test Cases Generation. In Proceedings of the IEEE International Conference on Consumer Electronics (ICCE 2024), Chengdu, China, 6–8 June 2024; pp. 1–6. DOI: <https://doi.org/10.1109/ICCE59016.2024.10444275>
- [14] Reshma, P.K., Rajagopal, S., Lajish, V.L., 2020. A Novel Document and Query Similarity Indexing Using VSM for Unstructured Documents. In Proceedings of the 6th International Conference on Advanced Computing and Communication Systems (ICACCS 2020), Coimbatore, India, 6–7 March 2020; pp. 676–681. DOI: <https://doi.org/10.1109/ICACCS48705.2020.9074255>
- [15] Susanto, A.D., Pradita, S.A., Stryadhi, C., et al., 2023. Text Vectorization Techniques for Trending Topic Clustering on Twitter: A Comparative Evaluation. In Proceedings of the 5th International Conference on Cybernetics and Intelligent Systems (ICORIS 2023), Moscow, Russia, 6–7 October 2023; pp. 1–7. DOI: <https://doi.org/10.1109/ICORIS60118.2023.10352228>
- [16] Alghamdi, J., Lin, Y., Luo, S., 2024. Unveiling the hidden patterns: A novel semantic deep learning approach to fake news detection on social media. *Engineering Applications of Artificial Intelligence*, 137, 109240. DOI: <https://doi.org/10.1016/j.engappai.2024.109240>
- [17] Vithanage, D., Yu, P., Wang, L., et al., 2024. Contextual Word Embedding for Biomedical Knowledge Extraction: A Rapid Review and Case Study. *Journal of Healthcare Informatics Research*, 8(1), 158–179. DOI: <https://doi.org/10.1007/s41666-023-00157-y>
- [18] Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K., 2024. LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation. *Natural Language Processing Journal*, 6, 100056. DOI: <https://doi.org/10.1016/j.nlp.2024.100056>
- [19] Li, Y., Xu, C., Cai, J., et al., 2024. Multi-label Classification of News Topics Based on Universal Sentence Encoder. In Proceedings of the 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI 2024), Shenzhen, China, 31 May–2 June 2024; pp. 419–422. DOI: <https://doi.org/10.1109/ICECAI62591.2024.10675181>
- [20] Saka, S.O., Cömert, Z., 2024. Sentiment Analysis based on Text with Universal Sentence Encoder and CNN-LSTM Models. In Proceedings of the 8th International Artificial Intelligence and Data Processing Symposium (IDAP 2024), Prague, Czech Republic, 15–17 May 2024; pp. 1–4. DOI: <https://doi.org/10.1109/IDAP64064.2024.10711063>
- [21] Pandya, V., Troia, F.D., 2023. Malware Detection through Contextualized Vector Embeddings. In Proceedings of the Silicon Valley Cybersecurity Conference (SVCC 2023), San Jose, CA, USA, 16–17 October 2023; pp. 1–7. DOI: <https://doi.org/10.1109/SVCC56964.2023.10165170>
- [22] Huang, W., Zhang, J., Li, X., et al., 2025. A Semantic and Intelligent Focused Crawler based on BERT Se-

- mantic Vector Space Model and Hybrid Algorithm (October 2024). In Proceedings of the IEEE Access Conference (virtual), Virtual/Online, 1–3 October 2024; p. 1. DOI: <https://doi.org/10.1109/ACCESS.2025.3542064>
- [23] Masethe, H.D., Masethe, M.A., Ojo, S.O., et al., 2025. Hybrid Transformer-Based Large Language Models for Word Sense Disambiguation in the Low-Resource Sesotho sa Leboa Language. *Applied Sciences*, 15(3608), 1–33. DOI: <https://doi.org/10.3390/app15073608>
- [24] Garg, S., Sharma, D.K., 2022. Role of ELMo Embedding in Detecting Fake News on Social Media. In Proceedings of the 11th International Conference on System Modeling & Advancement in Research Trends (SMART 2022), Bhubaneswar, India, 16–17 December 2022; pp. 57–60. DOI: <https://doi.org/10.1109/SMART55829.2022.10046789>
- [25] Jayakody, J., Vidanagama, V., Perera, I., et al., 2023. ELMo Layer Embedding Comparison with Short Text Classification. In Proceedings of the 3rd Asian Conference on Innovation in Technology (ASIANCON 2023), Bangkok, Thailand, 25–27 August 2023; pp. 1–6. DOI: <https://doi.org/10.1109/ASIANCON58793.2023.10270646>