computers

MDPI

Systematic Review

Context-Aware Embedding Techniques for Addressing Meaning Conflation Deficiency in Morphologically Rich Languages Word Embedding: A Systematic Review and Meta Analysis

Mosima Anna Masethe [1,2,*], Hlaudi Daniel Masethe [3] and Sunday O. Ojo [2,*]

[1] Department of Computer Science and Information Technology, School of Science and Technology, Sefako Makgatho Health Sciences University, Ga-Rankuwa 0208, South Africa
[2] Department of Information Technology, Faculty of Accounting and Informatics, Durban University of Technology, Durban 4001, South Africa
[3] Department of Computer Science, Faculty of Information Communication Technology, Tshwane University of Technology, Pretoria 0001, South Africa; masethehd@tut.ac.za
* Correspondence: mosima.masethe@smu.ac.za (M.A.M.); sundayo1@dut.ac.za (S.O.O.)

Abstract: This systematic literature review aims to evaluate and synthesize the effectiveness of various embedding techniques—word embeddings, contextual word embeddings, and context-aware embeddings—in addressing Meaning Conflation Deficiency (MCD). Using the PRISMA framework, this study assesses the current state of research and provides insights into the impact of these techniques on resolving meaning conflation issues. After a thorough literature search, 403 articles on the subject were found. A thorough screening and selection process resulted in the inclusion of 25 studies in the meta-analysis. The evaluation adhered to the PRISMA principles, guaranteeing a methodical and lucid process. To estimate effect sizes and evaluate heterogeneity and publication bias among the chosen papers, meta-analytic approaches were utilized such as the tau-squared (τ^2) which represents a statistical parameter used in random-effects, H-squared (H^2) is a statistic used to measure heterogeneity, and I-squared (I^2) quantify the degree of heterogeneity. The meta-analysis demonstrated a high degree of variation in effect sizes among the studies, with a τ^2 value of 8.8724. The significant degree of heterogeneity was further emphasized by the H^2 score of 8.10 and the I^2 value of 87.65%. A trim and fill analysis with a beta value of 5.95, a standard error of 4.767, a Z-value (or Z-score) of 1.25 which is a statistical term used to express the number of standard deviations a data point deviates from the established mean, and a *p*-value (probability value) of 0.2 was performed to account for publication bias which is one statistical tool that can be used to assess the importance of hypothesis test results. The results point to a sizable impact size, but the estimates are highly unclear, as evidenced by the huge standard error and non-significant *p*-value. The review concludes that although contextually aware embeddings have promise in treating Meaning Conflation Deficiency, there is a great deal of variability and uncertainty in the available data. The varied findings among studies are highlighted by the large τ^2, I^2, and H^2 values, and the trim and fill analysis show that changes in publication bias do not alter the impact size's non-significance. To generate more trustworthy insights, future research should concentrate on enhancing methodological consistency, investigating other embedding strategies, and extending analysis across various languages and contexts. Even though the results demonstrate a significant impact size in addressing MCD through sophisticated word embedding techniques, like context-aware embeddings, there is still a great deal of variability and uncertainty because of various factors, including the different languages studied, the sizes of the corpuses, and the embedding techniques used. These differences show how future research methods must be standardized to guarantee that study results can be compared to one another. The results emphasize how crucial it is to extend the linguistic scope to more morphologically rich and low-resource languages, where MCD is especially difficult. The creation of language-specific models for low-resource languages is one way to increase performance and consistency across Natural Language Processing (NLP) applications in a practical sense. By taking these actions, we can advance our understanding of MCD more thoroughly, which will ultimately improve the performance of NLP systems in a variety of language circumstances.

check for updates

Citation: Masethe, M.A.; Masethe, H.D.; Ojo, S.O. Context-Aware Embedding Techniques for Addressing Meaning Conflation Deficiency in Morphologically Rich Languages Word Embedding: A Systematic Review and Meta Analysis. *Computers* **2024**, *13*, 271. https://doi.org/10.3390/computers13100271

Academic Editor: Ming Liu

Received: 16 August 2024
Revised: 30 September 2024
Accepted: 7 October 2024
Published: 17 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Keywords: meaning conflation deficiency; context-aware word embedding; sense conflation; context conflation; entity conflation; contextual word embedding; word embedding; systematic review; meta-analysis; publication bias; PRISMA

1. Introduction

Meaning Conflation Deficiency (MCD) [1] is a serious problem when dealing with morphologically rich complex languages like Sesotho sa Leboa, which have complex derivation and inflection systems. A single root in these languages can give rise to numerous forms that express different semantic and grammatical nuances [1]. Because of this morphological richness, a root may have multiple meanings depending on its affixes and the grammatical context in which it is used. This can result in increased ambiguity. This can lead to misunderstandings or miscommunications because speakers and language processing systems may find it difficult to correctly interpret or produce the intended meanings [2].

Given their extensive use at the heart of many natural language processing (NLP) systems, MCD is one of the primary limiting aspects of word representations, which can result in faulty semantic interpretation of the input text and inevitably impair the performance. Nevertheless, their possible influence has hardly ever been examined in downstream NLP applications [2]. The conventional procedure involves treating the input text as a series of words and representing each word in the sequence with a dense distributional representation (word embedding), especially for neural models. It's important to note that this configuration ignores the possibility that a word could be polysemous, meaning it could have several (potentially unrelated) meanings [2].

The consequences of MCD are further exacerbated by the large degree of morphological variety, which makes it more difficult to map shape to meaning. Morphologically rich languages necessitate a more sophisticated comprehension of the interactions between various morphological constructions, in contrast to languages with simpler morphological structures, where form-meaning linkages are more obvious. In addition to influencing language production and understanding, this complexity presents difficulties for natural language processing models, which might find it difficult to deal with the variety of morphological patterns seen in these languages [2,3].

This paper intends to shed light on the specific ways that morphological richness exacerbates the challenges associated with meaning conflation by investigating the implications of MCD. In the end, researchers hope to advance understanding of language processing in a variety of linguistic contexts.

When working with morphologically rich languages (MRLs) [4], the elaborate word structures greatly enhance the complexity of language processing. Many inflectional, derivational, and compounding processes that change word forms based on grammatical elements including tense, gender, number, case, and mood are present in MRLs. On the other hand, morphologically simple languages (such as English) typically rely more on syntax and word order to convey meaning than on affixes or word form modifications. While polysemy is a problem in all languages, in MRLs it becomes more noticeable.

Sesotho sa Leboa (Northern Sotho) is a morphologically rich language. Its root word "bona" (to see) can be expressed as "ke bona" (I see), "re bone" (we saw), or "ba tla bona" (they shall see). Although the syntactic and semantic meanings of each of these forms differ, static models might interpret them as semantically equivalent. Inaccurate representations and subpar Word Sense Disambiguation (WSD) performance might result from a model that ignores the rich inflectional morphology of a word, which can collapse different senses or uses of the word. The difficulty to distinguish between these various incarnations of the same root term makes meaning ambiguity worse [4].

NLP models face special hurdles when morphology and meaning conflation deficiency interact. Multisense confusion occurs when different senses or meanings are combined into one word representation. This is especially problematic in languages where word forms

contain additional semantic and grammatical information. The difficulty here is not only in identifying these morphological forms but also in comprehending their relationships and contributions to the overall meaning of the word in context. A key issue for NLP in morphologically rich languages is the interplay between morphology and MCD. There is a higher chance of confusing various meanings of the same root word due to the intricacy that inflection, derivation, and compounding introduce.

Although word embeddings have shown to be successful in practice, there is a fundamental theoretical disadvantage that has long been acknowledged, many words are polysemous, meaning they have several senses or meanings [5]. Therefore, it appears dubious to represent every word as a single point in semantic space, regardless of how many interpretations it contains. This issue has been dubbed the "deficiency in meaning conflation". Polysemy leads to distortion in word embeddings, which is a practical consequence of the meaning conflation defect. For example, we would locate unrelated words left and wrong unnaturally close in the vector space because of their closeness to two different senses of the word. Word embedding models may have trouble determining which meaning of an ambiguous word is relevant in a particular situation [5].

Modeling the meanings of linguistic items in a machine-interpretable form, i.e., semantic representation, is one of the oldest, yet most active, areas of research in NLP. Confusion of all possible meanings of a word result from the main shortcoming of most semantic representation techniques, which typically treat word types as a single point in semantic space. This problem has been addressed in several recent research studies by learning distinct representations for individual meanings of words [6].

One significant flaw in most semantic representation approaches is that they typically treat a word type as a single point in the semantic space, conflating all possible meanings for the word. The solution to this problem, which involves teaching words to have unique representations for each of their meanings, has been the focus of numerous studies in recent years. Nevertheless, the generated sense representations are erroneous for rare word senses or are not linked to any sense inventory. Researchers [6] provide a method that addresses these issues by distinguishing between word representations based on the in-depth understanding that each word obtains from a semantic network [6]. Researchers contend that a series of tests on a cutting-edge neural network-based reverse dictionary system demonstrate that significant advancements can be achieved with a straightforward fix intended to overcome the meaning conflation deficiency [2]. Researchers proposed multi-prototype embedding or sense representation to overcome meaning conflation deficiency. Though, these representations are assessed using sense-centered tests like WSD or generic benchmarks like word similarity [2].

The most effective word, word sense, and concept modeling methods available today create dense vector representations that capture semantic commonalities in a low-dimensional space by utilizing massive corpora and information resources [7]. Nevertheless, most existing methods have a monolingual bias, and their efficacy varies with the volume of cross-linguistic data that is accessible. To overcome this problem, researchers developed conception, a unique method that maintains clear links between concepts while putting multilingualism at the center of language-independent vector representations of concepts. In multilingual and cross-lingual semantic word similarity and WSD, the method produces high coverage representations that surpass the state of the art and are especially resilient on low-resource languages [7]. The interest in sense embeddings is motivated by several applications, in addition to the limitations related to the meaning conflation deficiency. Sense embeddings have proven effective in word sense induction (WSI), and WSD has benefited from the application of knowledge-based sense embeddings [5].

Reverse dictionary system based on neural networks, with a simple adjustment is capable to address the meaning conflation deficiency substantial [2]. Returning a term given its explanation or description is known as concept lookup, conceptual dictionary, or reverse dictionary. For example, the system must return "Sickle" (Sesotho sa Leboa: Selepe) given "Sedirišwa seo se šomišwago go sega bjang (A tool used for cutting grass)". Because

of this, authors and translators may find a reverse dictionary system especially helpful in situations where they are unable to recollect a term quickly or are uncertain of how to explain a concept [2]. By enabling users to enter descriptions and automatically retrieve pertinent words, a neural network-based reverse dictionary for Sesotho sa Leboa could enhance language accessibility.

2. Related Work

This problem of Meaning Conflation Deficiency is particularly noticeable in low-resource, morphologically rich languages like Sesotho sa Leboa, where compounding, derivation, and inflection allow words to take on a variety of forms. Advanced word representation methods like context-aware embeddings and multi-prototype embeddings are needed to address MCD in such languages [8].

For a single word, multi-prototype embeddings [8] provide numerous vectors, each of which represents a different sense or meaning. By preventing a single, confused embedding for polysemous words, this directly addresses MCD. The embedding model groups words according to the various contexts they appear in, giving each group a unique vector. Each word is given many embeddings as a result, based on how it is understood in different situations. Based on context, the model chooses the best vector each time a word comes in a sentence [9].

The term "rata" has two meanings in Sesotho sa Leboa: "to love" and "to like". Given one vector to "rata", static embeddings would conflate the two meanings. However, separate vectors for rata (to love), as in "Ke rata mme" (I love my mother), and rata (to like), as in "Ke rata dipapadi" (I like sports), would be produced via multi-prototype embeddings [9]. The word "rata" would have a unique vector in each scenario that corresponded to its several meanings, enabling the algorithm to capture the minute variations between these senses [8].

Word representations in context-aware embeddings (e.g., BERT, mBERT, ELMo) are dynamically modified according to their surrounding context [10]. This method goes beyond static word vectors since the embedding of each word varies based on how the word is used in a phrase. Models such as BERT can fully comprehend a word by looking at its surrounding terms when reading text in both left-to-right and right-to-left directions. Context-aware embeddings, as opposed to static embeddings, produce distinct vectors for the same word based on the sentence-level context in which it appears.

Take the word "gola", which depending on the context can mean "grow" or "become large". In a statement such as "Ngwana o gola ka pela" (The child grows swiftly), the embedding of the word "gola" would indicate that it refers to the process of growing. However, in "Dilo di gola kudu" (Things are getting bigger), the word "gola" denotes getting bigger, and its embedding would support this interpretation. The word "gola" would have various representations in each sentence depending on the context if context-aware embeddings were used. This context-sensitive method guarantees that the model comprehends the correct meaning of "gola" each time it is used, which naturally resolves MCD [10].

Strong methods to alleviate MCD in morphologically rich but low-resource languages such as Sesotho sa Leboa are multi-prototype embeddings and context-aware embeddings. While context-aware embeddings dynamically modify word representations based on sentence-level context, multi-prototype embeddings tackle polysemy by producing numerous representations for a word in distinct senses. Combining these methods allows for richer language understanding and more accurate word sense disambiguation, especially in languages with complicated morphological systems. Although these methods have drawbacks such data sparsity and computational expense, they can greatly improve NLP systems for low-resource languages.

Cross-lingual word embeddings enable the learning of multilingual word representations without the need for direct bilingual signal, so bridging the gap between high-resource and low-resource languages. These algorithms, which map pre-trained embeddings into

a shared latent space, are projection-based. Orthogonal transformation—which presupposes that language vector spaces are isomorphic—is the foundation for most of these techniques. This requirement, however, is not always met, particularly in morphologically rich languages. A self-supervised technique is put forth by researchers [11] to improve the alignment of unsupervised multilingual word embeddings. Better alignment of cross-lingual embeddings is made possible by the suggested approach, which also ensures length- and center-invariance and shifts word vectors and their related translations closer to one another. The experimental results in a multilingual lexicon induction task outperform the state-of-the-art approaches [11].

2.1. Word Embedding (WE)

Word embeddings are dense vector maps where every word is mapped to a vector with a fixed size. By using patterns of word co-occurrence, these embeddings—which are acquired through extensive text corpora—seek to capture semantic links between words. Because WE are a static embedding, a word's representation remains constant regardless of its context. Context is not handled by WE; a word only has one fixed embedding. Usually produced by less complex models such as GloVe or Word2Vec [12]. One major flaw in the conventional methods for creating word embeddings is that they only create one vector representation for each word, failing to account for the possibility that ambiguous words could have several interpretations. Sensation representations have demonstrated remarkable performance on a variety of natural language processing tasks, surpassing word representations such as Word2Vec, GloVe, and FastText [13].

Static word embeddings, such as those produced by Word2Vec, GloVe, Wang2Vec, and FastText, are highly helpful in many applications, but they have a significant drawback: each word is linked to a single vector representation, neglecting the possibility that polysemous words could have more than one meaning. This restriction, which is a combination of several meanings in a single word, is known as Meaning Conflation Deficiency [7]. Many words have more than one sense or meaning, which is a basic argument against the idea of word embeddings. Words that have semantically unrelated meanings, like "*mafahla*" in Sesotho sa Leboa, are referred to as homophones. They are made up of multiple distinct entries in the lexicon that have the same spelling and phonological realization. Thus, it would seem illogical theoretically to capture the meanings of the term mafahla (lungs, twins) with a single mathematical representation, yet this is precisely what word embeddings accomplish. The meaning conflation deficiency is the name given to this disadvantage [5]. Word embeddings become distorted because of the meaning conflation defect. Comparing the vectors of two words is a typical way to determine how similar they are; consider the application of cosine distance in the word analogies challenge. Sadly, not much research has been done in low-resource languages like Sesotho sa Leboa to determine how much the lack of meaning conflation deficiency the effectiveness of NLP systems that use word embeddings [5].

2.1.1. Word2Vec

Word embeddings are created using the Word2Vec model, a shallow neural network model. The output vectors are arranged so that common contexts are near to one another in the vector space. This makes it possible to comprehend a word's meaning by considering its context. Word2Vec offers two word embedding models: the Skip-Gram and the Continuous Bag-of-Words (CBOW) models. Both models' concepts are derived from the Distributional Hypothesis, which claims that various terms that seem to have similar meanings in comparable settings [14]. The foundation of the CBOW model is the unigram, which is an n-gram—a continuous series of n terms—that consists of just one term. By averaging the context of every word in the selection window, the sequence can record the word's context. Skip-grams are used by the Skip-Gram model to define word context. This makes it possible to record the context of words that are farther apart in the sequence and

is used to anticipate the context given an input word [14]. The skip-gram objective [5] is expressed by the Equations (1) and (2):

$$J(c,\ o,\ u,\ v) = \log \vartheta\left(u_c^T v_o\right) + \sum_{i=1}^{k} E_{w_i} \sim p_n\left[\log\left(1 - \sigma\left(u_c^T v_{w_i}\right)\right)\right] \quad (1)$$

where input vector u∈R and output vector v∈R, i∈V, given c to be center word, o is a word in the context of c, k is the number of negative samples per true context word and the sigmoid function σ is given by

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

2.1.2. GloVe

Word embeddings are produced by the unsupervised machine learning algorithm GloVe. To achieve this, it first creates a term co-occurrence matrix, which includes the frequency at which each word occurs in relation to other words. Because there are a combinatorial number of contexts, this matrix has enormous dimensions [14].

2.1.3. FastText

The Word2Vec model is enhanced by the FastText model, which uses both CBOW and Skip-Gram. The first enhancement speaks to the disadvantages resulting from failing to consider the words' placement within the context window. The suggested method weights the context words by first learning the representations of word positions [14].

2.2. Contextual Word Embedding (CWE)

Contextual word embeddings are produced by models that alter a word's embedding according to the context in which it appears in a phrase. The words surrounding these embeddings cause them to change dynamically. frequently developed using sophisticated language models like BERT, GPT, or ELMo. Word representations in CWE are dynamic embeddings that change according to the context of a phrase. To handle context, CWE creates many embeddings for the same word depending on the words that surround it and is produced by more sophisticated models that employ transformer topologies and deep learning, such BERT or GPT [7]. Contextualized representations are constantly modified according to the context in which the target word occurs. This property is known as context sensitivity. Contextualized embeddings, or deep neural language models, are produced by the architectures ELMo, BERT, and GPT-2. These models can then be refined to produce models for a range of downstream NLP tasks. Phrase embeddings in these language models are constructed by summing these internal representations, which are referred to as contextualized word representations because they are functions of the complete input phrase [7].

Context-level learnt word embedding models are those that learn word representations solely by utilizing words that surround the chosen word. More recently learnt models at the context level exhibit computational efficiency. Research has mostly concentrated on the data used to train models, if inference computes a vector representation of a word through a single word. As previously indicated, there are a few issues with this method, chief among them being its incapacity to depict several interpretations of a single word. As a result, brand-new models known as contextualized word embeddings have been put forth. These models are more appropriate as their representations can alter depending on the situation [15]. To address the meaning conflation deficit issue, current WE approaches encode a word as a single vector at every time interval. With the goal of addressing not just what and when, but also how the meaning of a word changes, the researchers suggest a sense representation and tracking framework based on deep contextualized embeddings to handle this problem. The results demonstrate that the framework improves word change detection job significantly and is useful in capturing fine-grained word senses [16].

2.2.1. Embedding from Large Language Models (ELMo)

Uncontextualized word embeddings are computed using ELMo, which is solely character-based. Convolutional filters, two highway layers, and a linear projection are used to accomplish this. ELMo effectively defines two language models with BiLSTM networks using these representations: one model predicts the likelihood of a token for tokens that have already occurred and the other predicts the likelihood of tokens that will occur in the future [15]. The basic language model capability of ELMo is its capacity to anticipate the following word in a phrase. The model also begins to recognize language patterns after being trained on a sizable dataset. It is composed of a bidirectional LSTM language model with two layers that was initially trained on the Billion Word Benchmark dataset. It is constructed over a context-independent character CNN layer [13].The advent of contextualized models, such ELMo and BERT, has been one of the most important recent developments in NLP. In NLP, contextualized models, also known as contextualizes, are examples of successfully applied transfer learning. Contextualized models provide a corresponding series of contextualized word embeddings given an input word sequence, such as a phrase. Contextualized word embeddings, which capture some of the language phenomena that function beyond the word level, represent word meanings in context, whereas word embeddings represent the meaning of words in isolation [5].

2.2.2. T5

Researchers [17] use Text-to-Text Transfer Transformer (T5) and multilingual T5 (mT5), for contextual embeddings to be used in computing and avoid losing information by mapping the embeddings to the Word2Vec space and create a solution that is tailored to a Transformer model. This metric has previously been applied on contextualized embeddings by first mapping them to the Word2Vec embedding space. Transformer models are now being used in real-world applications more and more [17].

2.2.3. GPT

Model of Generative Pre-Training (GPT). GPT employs a unidirectional language model as opposed to a bidirectional one. The Transformer is a unique architecture that the GPT model employs for feature extraction in place of LSTM layers. This design has been shown to be highly effective in a variety of NLP tasks. As a result, GPT can handle long-term dependencies in the text by storing more organized memory, leading to a more comprehensive model [15].

2.2.4. ELECTRA

Scholars [18] contend that, in contrast to the conventional context-free approaches, there are few research using contextual word embedding techniques (i.e., transformers) that consider a word's context. To evaluate the effectiveness of various ensemble learning models in identifying both useful and non-informative disaster-related Twitter communications, the researchers [18] combines transformers with deep neural network techniques. When combined with Bi-LSTM, simpler transformer variations like ELECTRA and Talking-Heads Attention produced results that were both superior and comparable to the computationally expensive BERT, with F-scores ranging from 80% to 84%. Our results demonstrate the effective usage of the more recent and straightforward transformers at a lower computational cost [18].

2.2.5. BERT

Instead of using a one-way design like GPT, Bidirectional Encoder Representations for Transformers (BERT) uses a bidirectional transformer architecture. Additionally, they added two unsupervised tasks to BERT's pre-training [15]. BERT is a deep Transformer encoder that was trained on next-sentence prediction and a masked language model simultaneously [13]. Scholars [10] investigate the potential of the contextualized embeddings predicted by BERT to generate high-quality word vectors for such domains, namely knowl-

edge base completion, where the semantic qualities of nouns are the main focus. The study discover that a straightforward method of averaging the contextualized embeddings of masked word occurrences produces vectors that perform better than the static word vectors that BERT learnt and those derived from conventional word embedding models [10]. One way to conceptualize a contextualized word embedding model, like BERT [5], is as two parts in Equations (3) and (4): an embedding matrix $E \in R^{|V| x\ d}$, as well as a contextualizing function C that translates an input embedding sequence $x_1, x_2, \ldots, x_n \in R^d$ to output representations $y_1, y_2, \ldots, y_n \in R^d$

$$P(w|y) = \left[\text{softmax}\left(E_y + b\right)\right]_w \tag{3}$$

$$= \frac{e^{e_w^T y + b_w}}{\sum_{j \in V} e^{e_j^T y + b_j}} \tag{4}$$

The methods generate word-level representations that are unaffected by the context in which a word occurs; these so-called "static" representations frequently exhibit a significant bias in favor of a word's most common sense. Since language models like BERT, which are pretrained on vast volumes of text, have become widely available, contextualized embeddings have seen a sharp increase in popularity. A polysemous word's multiple features can be captured in a context using contextualized word embeddings [7].

Scholars investigated unsupervised sense representations, which, in contrast to conventional word embeddings, can elicit several meanings for a word by examining its contextual semantics inside a text. The researchers are looking into deep neural language models and sensory embeddings. In syntactic and semantic parallels tests, the results demonstrate that the sense embedding model (Sense2vec) outperformed conventional word embeddings, demonstrating that the created language resource can enhance the performance of NLP tasks in Portuguese. In the semantic textual similarity challenge, researchers also assessed the effectiveness of pre-trained deep neural language models (BERT and ELMo) in two transfer learning techniques: feature-based and fine-tuning. Tests show that the Multilingual and Portuguese BERT language models, after being fine-tuned, were able to outperform the ELMo model in terms of accuracy [13].

Academics [7] suggest a knowledge-based approach for learning word meaning representations that can provide useful assistance in tasks that come after. More precisely, the study suggests using sememes—the smallest semantic units of meaning—to extract semantic information from past human knowledge to create global sense context vectors and carry out dependable soft word sense disambiguation for polysemous terms. A contextual attention mechanism is added to the Skip-gram model architecture in this study to enable the learning of unique embeddings for each sense. Both the extrinsic experiment and additional analysis demonstrate that our sense embeddings can be used to reduce the negative effects of polysemy and enhance performance in a variety of real-world downstream tasks. The intrinsic experimental results outperform earlier work on word similarity tasks and capture the precise and different meanings of sensations [7]. Numerous studies have concentrated on modifying Neural Language Models (NLMs) to generate word embeddings that are more sense-specific considering the meaning conflation deficit problem with conventional word embeddings. We begin creating sense embeddings in this study by utilizing the methodology of recent works in contextual word embeddings, specifically context2vec [19]. The contextualized word embedding model is nevertheless susceptible to the meaning conflation defect regarding its word embeddings. It could be more difficult to distinguish the muddled meanings after this has happened [5].

2.3. Context Aware Word Embedding (CAWE)

Like contextual embeddings, context-aware word embeddings usually concentrate on adding more contextual information to enhance representation. This could involve knowledge of syntax, semantics, or a particular domain. Context-aware embeddings

frequently incorporate deeper contextual integration or more elements in addition to the basic context. incorporates more intricate or extensive contextual information to provide an additional degree of context awareness frequently incorporates extra contextual data beyond basic CWE or uses even more sophisticated models.

Sentential or documental contexts are dynamically represented for a word using context-aware word representation algorithms like context2vec or ELMo. Recently, deep neural language models have been trained for the Portuguese language, yielding state-of-the-art outcomes on downstream natural language processing tasks. Studies that rigorously assess these models are still required, nevertheless, for several applications. To compare deep neural language models' performance with traditional word embeddings for both Brazilian Portuguese and European Portuguese, researchers [20] plan to use the ASSIN dataset's semantic similarity tasks. The results of the experiments show that the ELMo language model outperformed all other pretrained models for the Portuguese language in terms of accuracy, and that vocabulary reduction on the dataset prior to training enhanced both the standalone performance of ELMo and its combination with traditional word embeddings [20]. Conventional approaches primarily see this work as a machine learning algorithm-based text categorization problem. But most of them mostly rely on the manually created elements that were taken from the text. The researchers jointly encode the global and local contextual information to present CADEN, a context-aware deep embedding network for financial text mining. Investigate capture and incorporate an attitude-aware user embedding to improve our model's performance [10].

Word embedding is associated with downstream activities, and an open topic is how to regularize word embeddings to obtain context-aware information. To address these problems, we first suggest context-aware dynamic word embedding, or CDWE, which may consider a word's context information in addition to its generic and domain-specific meanings. Based on CDWE, researchers [19] propose an attention-based convolution neural network for aspect term extraction (ATE) dubbed ADWE-CNN. By using an attention mechanism to assign varying importance to the respective embeddings, the network can adaptively capture the previous meanings of words [19].

To capture deep contextual representations, the researchers [21] suggested a supervised framework that uses pre-trained context-aware embedding from large language models (ELMo). This framework is further summarized by a bidirectional long short-term memory (Bi-LSTM) layer, which is used to learn long-range code dependency. The outcomes of the trial demonstrated that the suggested framework performed optimally in tasks that came after. Random forest and support vector machines beat four baseline systems on the datasets using feature representations produced by the proposed framework, demonstrating that the framework integrated with ELMo successfully captures vulnerable data flow patterns and facilitates detection job [21].

In discovering and labeling sensitive information in conversational data, the research [22,23] proposes a predictive context-aware model based on a Bidirectional Long Short Term Memory network with Conditional Random Fields (CRF) (BiLSTM + CRF) (multi-class sensitivity labelling). A discriminative model called Conditional Random Fields (CRFs) is primarily utilized for labeling and classifying sequential data. The best performing model architecture, according to the results, is the BiLSTM + CRF model with BERT embeddings and WordShape features (F1 score 96.73%) [22,23].

Embeddings that are contextualized, such as BERT and ELMo, consistently perform better than those that are not. The best model is the solo ELMo model using logistic regression, which achieves remarkable accuracy scores of 90% on Twitter and 94% on YouTube data. Word embeddings that deal with word meaning, like Word2vec and Glove, have begun to proliferate in the field. They produce a single vector (embedding) for every word, integrating all its various senses into one single vector, regardless of the context. Depending on where a word appears in the context, BERT and ELMo can produce distinct word embeddings for the same word. These models can greatly enhance the performance of a variety of NLP issues since they were trained on enormous volumes of data [14,24].

The Scholars [14,15,24] contend that the incorporation of fuzzy logic with Word2Vec, GloVe, and BERT embeddings yields good accuracy results across many datasets, indicating the effectiveness of the model in comprehending the complex language of sarcasm. The Bidirectional Encoder Representations from Transformers (BERT) model is a noteworthy advancement in pretrained models since it can efficiently refine smaller, domain-specific datasets, leading to better performance for certain objectives. The advantages of this strategy outweigh the significant computational resources required. Since homonyms are words with the same spelling but different meanings, ELMo's ability to understand context allows it to create distinct representations for them. This makes ELMo different from traditional models in that it considers the entire input phrase, leading to a more sophisticated understanding of words [24].

Bidirectional LSTM (BiLSTM), the brains behind ELMo, is trained utilizing a language modeling technique on copious volumes of text. By merging forward and backward LSTM components, BiLSTM effectively analyzes sentences and can predict word sequences in both directions. This is why a two-way language model, which leverages word embeddings and BiLSTM processing to enhance language understanding, is a helpful tool for numerous NLP applications [23,24]. The study [25] presents a model that combines word embedding, clustering, and classification algorithms to efficiently group Turkish project proposals. The suggested model extracts terms from project proposals' titles and abstracts in Turkish using FastText, BERT, and term frequency/inverse document frequency (TF/IDF) word-embedding approaches. The methods of classification and clustering were used to the extracted terms. A fresh dataset of more than 30,000 naturally occurring, non-trivial instances of noun-verb ambiguity is produced by the researchers [26]. When measured on the WSJ, taggers within 1% of one another have accuracy rates ranging from 57% to 75% for this challenge set. The accuracy of the strongest current tagger is increased to 89%, a 14% absolute (or 52% relative) improvement, by adding contextual word embeddings and tailored training data. In the downstream process, utilizing only this improved tagger results in a 28% decrease in error compared to the previous best learnt model for text-to-speech synthesis homograph disambiguation.

2.3.1. PolyLM

To circumvent word embeddings' "meaning conflation deficiency", some models have attempted to embed specific word senses. These approaches were formerly effective on word sense induction (WSI) tasks, but task-specific strategies that take use of contextualized embeddings have subsequently surpassed them [1]. However, contextualization and sense embeddings don't always have to conflict. The PolyLM method, which enables contextualization techniques to be used, is introduced by researchers [1]. It formulates the process of learning sense embeddings as a language modeling problem. Researchers [1] test PolyLM on WSI and find that, although having six times less parameters than the state-of-the-art specialized WSI method, it performs significantly better than earlier sense embedding algorithms [5]. To overcome the meaning conflation deficiency, researchers [5] created PolyLM, an end-to-end contextualized sense embedding model. PolyLM uses training on masked language modeling to simultaneously learn sense embeddings and a contextualizer. Word-sense related NLP jobs can be readily handled by PolyLM with little to no further training. PolyLM helps prevent meaning conflation deficiency by using solely sense embeddings rather than word embeddings [5]. Words in context evoke senses it induced. As a result, a text passage can be considered both a word sequence and a sequence of semes. The likelihood of the word w happening in a context c is equal to the sum of the probabilities of w's component occurring in the context, according to the first observation underpinning PolyLM as expressed in Equation (5):

$$P(w|c) = \sum_{s \in S_w} P(s|c) \quad (5)$$

2.3.2. XLNET

The sole pre-training goal of XLNet, which is built on a Transformer-XL architecture, is Permutation Language Modeling (PLM) [21]. PLM is motivated by the fact that it does not rely on masked tokens, bringing pre-training closer to the process of fine-tuning for activities that come after. Additionally, compared to BERT, it is trained on far bigger corpora, supplementing the BERT corpora with a substantial amount of web material from other sources. Sentence Piece, an open-source tokenization tool that is substantially like WordPiece, is used by XLNet in place of WordPiece [21]. Utilizing the Transformer-XL architecture, XLNET can learn in both directions by maximizing the likelihood across all factorization order variations [15].

2.3.3. DistilBERT

A linguistic model was put up by academics [22] to identify hate speech in Twitter data. Support Vector Machine (SVM) and Distil-BERT, a context-aware embedding model, have been developed for the classification of hate speech. The results demonstrate that using a context-aware embedding model improves accuracy. DistilBERT is a condensed form of the general-purpose BERT that is 40% smaller, 60% faster, and 97% capable of understanding language compared to the original BERT model. With distillation, this general-purpose language model is successfully trained [22].

2.3.4. RoBERTa

It is stated clearly that RoBERTa is an enhanced variant of BERT. As a result of discovering that the NSP pre-training target worsens performance in the described experimental environment, RoBERTa performs just MLM during pre-training. Additionally, it is trained using various hyperparameter selections (e.g., bigger batch sizes) that enhance performance on downstream tasks. Like XLNet, the models offered with RoBERTa are likewise trained on larger datasets primarily consisting of web content. RoBERTa uses byte-level BPE for tokenization, which makes it harder to retrieve embeddings for certain tokens (i.e., spacing needs to be encoded explicitly) [23].

2.3.5. GBCN

Researchers [24] contend that the BERT's input format is limited to a word sequence and is unable to give further contextual information. The work presents a novel approach called GBCN to address this issue. It enhances and controls the BERT representation for aspect-based sentiment analysis by using a gating mechanism with context-aware aspect embeddings. First, the input texts are fed into the context-aware embedding layer and BERT representation, respectively, producing a test F1 of 88.0 and 92.9 for the BERT representation and refined context-aware embeddings [24].

2.3.6. DeBERTa

Transformer serves as the foundational architecture for DeBERTa. DeBERTa presents a disentangled self-attention process in contrast to BERT [25]. A pre-trained language model from DeBERTa was used to generate the third model. Disentangled attention and improved decoding are two strategies used by the BERT-based transformer model DeBERTa to boost efficiency. They divided the location encoding and the token encoding into two distinct vectors within DeBERTa using a two-vector technique. As a result, the attention layers were able to distinguish and gain knowledge from every encoding vector independently. Additionally, they employed improved decoding, in which they provided the model both the absolute and relative placements of each word in the phrase. In many instances, DeBERTa performs better than BERT thanks to these advancements. In numerous Natural Language Processing jobs, it attains cutting-edge results [26].

2.3.7. ALBERT

As a more parameter-efficient variant of BERT, the study [23] suggested ALBERT, a lite BERT design with substantially fewer parameters than a regular BERT architecture. Even with modifications made to increase efficiency (such sharing of cross-layer parameters), ALBERT's design is comparable to that of BERT. In addition to increasing productivity, ALBERT enhances performance on downstream tasks by substituting the harder Sentence Order Prediction (SOP) target for NSP. XLNet and ALBERT both use SentencePiece tokenization, and both are trained on comparable corpora. It uses fewer parameters and exhibits benchmark performance like BERT. It is available in multiple configurations [23].

By combining WSD methods with the power of sequence-to-sequence (Seq2Seq) models, the researchers [27] developed a novel method for abstractive text summarization. To accomplish effective summarization, the study built a Seq2Seq model using LSTM units, embedding layers, and Time Distributed Dense layers. Its focus was on optimizing the vocabulary size, embedding dimensions, and LSTM units. Using the preprocessed dataset, the model is trained across multiple epochs to improve its summarization skills. The method produces more accurate and cogent summaries by demonstrating a notable improvement in catching the context and subtleties of the input texts [27]. Table 1 summarizes the performance from various authors in solving meaning conflation deficiency problem.

Table 1. Summary of Performance benchmarks of embeddings to solve MCD problem.

Author	Approach	Model	F1 Score (%)	Accuracy (%)
[26]	WE	FastText		89
[27]	CWE	ELMO		89
[23]	CAWE	DistilBERT	84	85
[24]	CAWE	GBCN	88	93
	CWE	ELECTRA	84	

3. Materials and Methods

Using a methodical approach to the literature selection was part of the process. This methodical selection of the literature made it possible to implement a structured search strategy with well-defined inclusion and exclusion criteria, encouraging objectivity and transparency. In the end, this improves reliability since possible biases in the literature selection may be reduced and freely debated. Using the PRISMA approach (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), a systematic literature review was carried out [28]. Our systematic literature review's (SLR) objective is to thoroughly locate, assess, and assemble relevant publications in the fields related to meaning conflation deficiency (MCD) for low-resource languages. The investigation was guided by the following sub-questions:

RQ1: What salient features of a less resource morphologically rich language, such as Sesotho sa Leboa, are required to develop a Context-Aware Word Embedding (CAWE) model?
RQ2: What are the existing word embedding methods?

Our systematic review used the PRISMA guidelines, which provided a structured approach to ensure comprehensive coverage and a rigorous analysis of the literature. We developed clear inclusion and exclusion criteria to pick the studies based on their relevance and to minimize bias [29].

3.1. Research Strategy

For this work, the researchers meticulously examined the IEEE Xplore, Springer, Web of Science, Google scholar, and Scopus databases as illustrated in Table 2 to conduct a thorough systematic literature review and to locate a greater range of content that may not be listed in traditional databases. A unique search string was used to individually target each database: "Meaning conflation deficiency" OR "context-aware word embedding" OR "sense conflation" OR "context conflation" OR "entity conflation" OR "contextual word embedding".

Table 2. Summary of database search results for WSD.

Database	Results	Search Phrase	Notes
Scopus	269	"Meaning conflation deficiency" OR "context-aware word embedding" OR "sense conflation" OR "context conflation" OR "entity conflation" OR "contextual word embedding"	extensive database with a broad scope.
Springer	10	"Meaning conflation deficiency" OR "context-aware word embedding" OR "sense conflation" OR "context conflation" OR "entity conflation" OR "contextual word embedding"	focused on information technology and computers.
IEEE Xplore	35	"Meaning conflation deficiency" OR "context-aware word embedding" OR "sense conflation" OR "context conflation" OR "entity conflation" OR "contextual word embedding"	abundant in publications on engineering and technology.
Google Scholar	56	"Meaning conflation deficiency" OR "context-aware word embedding" OR "sense conflation" OR "context conflation" OR "entity conflation" OR "contextual word embedding"	offers a quick and easy method for searching academic publications in general.

First, a thorough literature search is conducted in a variety of databases and sources to find studies that may be pertinent to the meaning conflation insufficiency. As seen in Figure 1, the study in this instance found 403 articles during this phase. Twenty-five publications were found to be eligible for inclusion in the meta-analysis following screening. The process of selecting and including studies is represented by the PRISMA flow diagram shown in Figure 1. A flow diagram of the different phases of our systematic review is depicted in Figure 1.

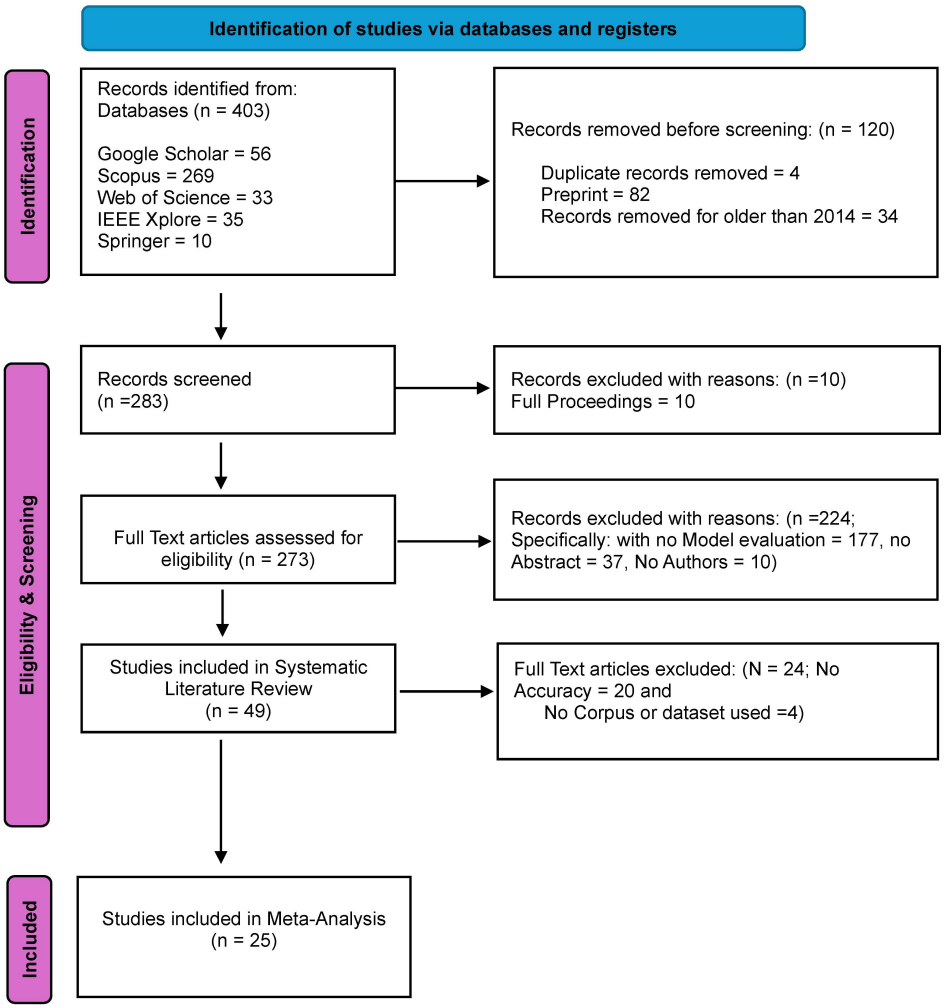


Figure 1. PRISMA flow diagram for the CAWE research [30].

3.2. Inclusion Criteria and Exclusion Criteria

The inclusion and exclusion criteria are crucial in choosing which studies are deemed relevant when doing a review or selection of papers on Meaning Conflation Deficiency (MCD). The inclusion criteria that were utilized to choose relevant papers were based on the type of study. Although the keywords were selected with the intended result clearly in mind, they did not give a clear picture of what was anticipated. When the search parameters were used, hundreds of relevant articles were discovered in the initial phase of the process. The keywords were present in the titles, abstracts, text, and reference lists. Together with the selected key words, we also defined specific inclusion and exclusion criteria, as Table 3 illustrates. Furthermore, to guarantee comprehensive coverage, studies that address MCD in a wide range of language families—including low-resource languages like Sesotho sa Leboa, Amharic, and Xhosa, and high-resource languages like English and Chinese—were included. A particular emphasis was given on research that look at morphologically rich languages like Finnish, Turkish, and Bantu languages.

Table 3. Inclusion and exclusion criteria.

Criteria	Decision
The predetermined keywords appear throughout the document, or at the very least in the title, keywords, and abstract sections.	Inclusion
Publications released in 2014 and after	Inclusion
Research article written in the English language.	Inclusion
Research articles without MCD-selected approaches, which are WE, CWE and CAWE	Exclusion
Research articles without accuracy	Exclusion
Research articles without a corpus or dataset	Exclusion
Articles not written in English, reports published prior to 2024, case reports and series, editorial letters, commentary, opinions, conference abstracts, and dissertations	Exclusion

Included are studies that use static word embeddings such as Word2Vec, GloVe, or FastText because these embeddings are known to display Meaning Conflation Deficiency because they cannot properly manage polysemy (many meanings). Articles that use transformer-based models, ELMo, BERT, mBERT, XLM-R, or other context-aware embeddings are included to evaluate their capacity to dynamically modify word meanings according to context, thus addressing MCD. Included are studies that mix static and context-aware embeddings (i.e., hybrid models that combine the best features of both worlds) since they offer valuable insights on how integrated techniques might minimize MCD. Research addressing the fundamental problem of MCD is prioritized. WSD approaches research is included because, by determining the proper sense, these methods immediately resolve MCD.

To guarantee the validity and academic rigor of the results, only research included have been published in peer-reviewed journals or conference proceedings. F1 score, precision, recall, and other metrics unique to WSD tasks (such the SemEval WSD benchmark) are important measures of a model's effectiveness in addressing MCD. The most recent developments have been captured by reviewing research that has been published in the last five to seven years, with a focus on transformer-based embeddings and how they help with MCD.

Research on languages with abundant resources may be disregarded if they don't offer techniques or insights that may be applied in environments with less resources. Research lacking precise assessment criteria have been omitted. This keeps theories and unproven methods out of the mix and concentrates on those that have been proven effective in real-world scenarios.

For doing an exhaustive and important assessment of studies on MCD, inclusion and exclusion criteria are essential. The review offers thorough awareness into the state of the field by emphasizing works on low-resource and morphologically rich languages, concentrating on embedding-based approaches, and requiring explicit assessments of success. On the other hand, excluding studies that don't have empirical support, concentrate just

on high-resource languages, or don't specifically address MCD contributes to keeping the review rigorous and relevant

3.3. Data Synthesis and Statistical Analysis

An Excel spreadsheet was used to enter the data to prepare it for statistical analysis. These were then put into two statistical analysis applications, STATA and RStudio. The outcomes of all primary investigations were calculated using the overall pooled effect size and the effect sizes of each primary study that was included in the data collected. Our investigation was based on the random-effects model [31]. A summary of the research topic was obtained by processing and analyzing the abstracts of the gathered literature. A manual check was conducted to achieve a more accurate inclusion. The study screened unneeded articles and manually extracted related articles because not all the articles were aligned with the research objectives. After several rounds of screening and extraction, 403 articles were found.

An accurate way to quantify a phenomenon's size is to use its effect size. Comparing the strength of findings from various research is helpful when conducting a meta-analysis. Researchers can assess how well embedding approaches work to solve MCD by looking at the effect magnitude. The study uses Equation (6) to compute Cohen's d [32] to determine the difference between two means:

$$d = \frac{M_1 - M_2}{SD_{pooled}} \tag{6}$$

where M_1 and M_2 are means of the groups and SD_{pooled} is the pooled standard deviation.

4. Results

A search of many databases, including Google Scholar, IEEE Xplore, Springer Link, Web of Science, and Scopus, produced 403 publications. A total of 120 articles were removed due to duplicate content, and other articles were removed because they were not in English. A total of 248 papers were excluded because the model evaluation was not completed, the authors were not identified in some instances, or the abstracts were not completed in others. All ten proceedings, including the pieces devoid of abstracts, were excluded as well. A final meta-analysis with twenty-five studies was carried out.

The diagram in Figure 2 is a proposed WSD Taxonomy based on the review of the literature. The move from static word embeddings to contextual embeddings is shown in Figure 2, which signifies a significant advancement in the comprehension and representation of word meaning by natural language processing (NLP) models. Word2Vec, GloVe, and FastText [13] are examples of static word embeddings, which are early methods of expressing words as fixed vectors in a continuous space. These embeddings, which learned word distributional patterns from vast corpora, captured syntactic and semantic links between words. These techniques lacked context-specific subtleties, such as MCD, which assigned a word only one vector representation independent of context, and Context Insensitivity, which did not modify embeddings based on surrounding words [13].

The shortcomings of static embeddings were addressed with the introduction of contextual word embeddings [7], which resulted in notable improvements. Context-aware representations were first presented by transformers and models such as ELMo (Embeddings from Language Models), BERT (Bidirectional Encoder Representations from Transformers), and GPT. This was done through context sensitivity in which contextual word embeddings (CWE) are sensitive to the words that surround a given word, in contrast to static embeddings and dynamic representation where pre-trained language models provide contextual embeddings by producing a dynamic vector for every word in each sentence. Now, depending on how a term is used, it can have several different representations, improving semantic disambiguation [7].

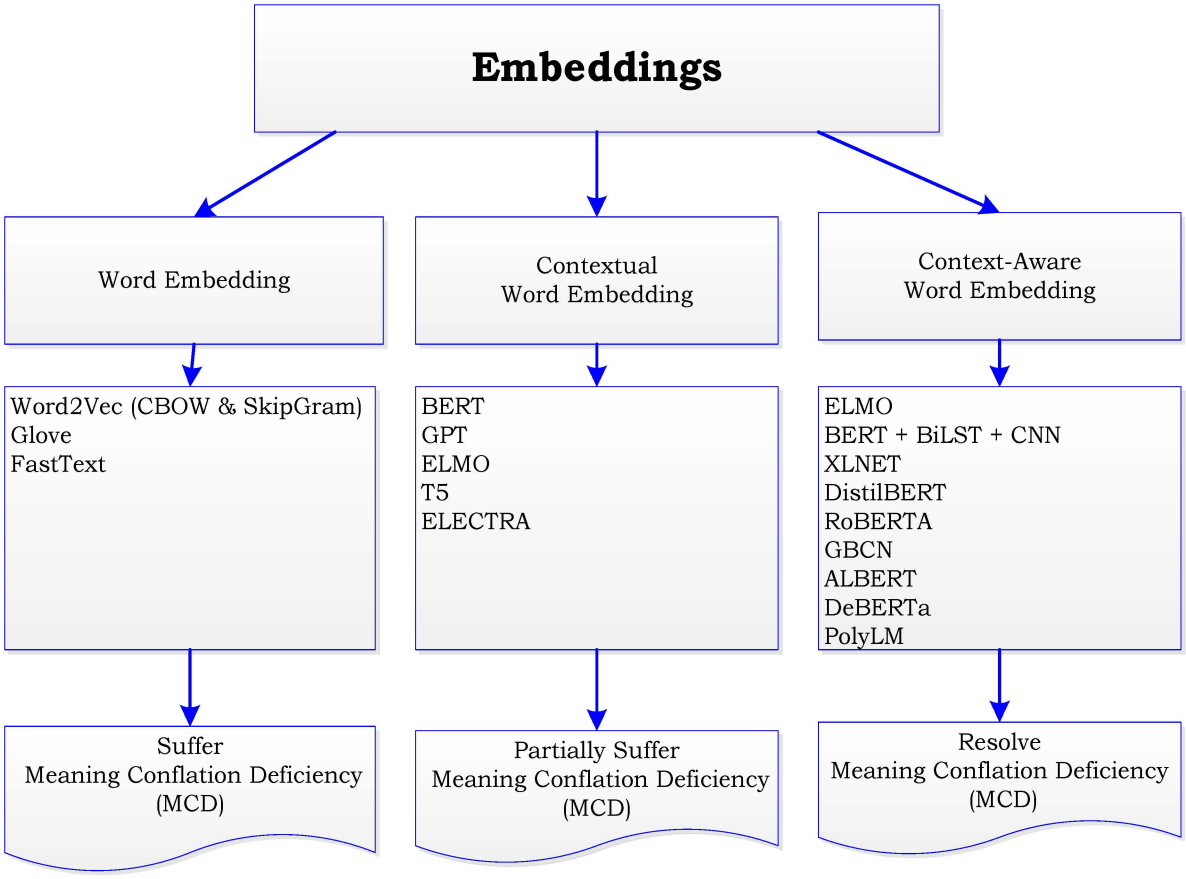


Figure 2. Proposed Embeddings Taxonomy.

The concepts of contextual word embeddings are expanded upon by context-aware word embeddings (CAWE) [20], which emphasize capturing deeper linguistic nuances and dependencies not only at the word level but also at the sentence and document levels. Contextual embeddings became more than just individual words as NLP developed; they now encompass phrases, sentences, and even entire texts. These embeddings can provide a deeper grasp of the overall context in addition to just deciphering individual words. Context-aware models—like BERT, GPT, RoBERTa, and T5—are adjusted for tasks after undergoing pre-training on sizable corpora. These models are quite adaptable as they learn broad language patterns that may be used to many NLP applications. Transformers employ self-attention methods to record dependencies across complete phrases, which power context-aware embeddings [20].

4.1. Meta-Analysis Summary

Table 4 presents the variation in effect sizes between studies represented by tau-squared, which measures the variation in effect sizes between research projects that isn't brought on by mistakes in sampling. With a τ^2 value of 8.8724, which is rather high, there is a significant amount of variation in the impact sizes among the studies. This implies that there exist notable variations in the way research reports meaning conflation deficiency, maybe resulting from variations in methodology, study populations, or definitions of MDC. I^2 calculates the percentage of overall effect size variability attributable to between-study variability as opposed to random variation. It has a percentage as its expression. With a high I^2 value of 87.65%, study differences, not random error, account for a significant amount of the diversity in effect sizes. This implies a great deal of variety in the outcomes, emphasizing the necessity for additional research into the causes of this variability. The ratio of overall variability (observed) to within-study variability (anticipated if true differences were absent) is used to compute H-squared, another metric for measuring between-study variability. It

shows how different the studies are from one another. When compared to within-study variability, an H^2 score of 8.10 indicates that there is a significant amount of variability between studies. This suggests that the outcomes of the studies are not readily comparable and further supports the existence of significant variability. Considerable variation among the studies is shown by the combination of high τ^2, I^2, and H^2 values. This variation may result from variations in Meaning Conflation Deficiency definitions, measurement methods, demographic characteristics, or study designs. The large degree of variation indicates that although MDC is a useful concept, its measurement and interpretation are not agreed upon in a single approach. Conclusions drawn from the interpretation of meta-analytic data should take this diversity into account.

Table 4. Meta-Analysis Summary on CAWE: random-effects model (DerSimonian–Laird).

Meta-Analysis Summary: Random-Effects Model Method: DerSimonian–Laird					
Heterogeneity: tau^2 = 8.8724; I^2 (%) = 87.65; H^2 = 8.10					
Study (*n* = 25); Effect Size; [95% CI]; Weight					
(Elkahky et al., 2018)	[33]	−10.339	−12.330	−8.349	4.11
(Aksoy et al., 2023)	[34]	−7.913	−9.991	−5.836	4.07
(Elagbry et al., 2021)	[35]	−10.091	−12.809	−7.373	3.77
(Sinha et al., 2015)	[36]	−18.286	−20.296	−16.275	4.10
(Younas et al., 2023)	[29]	−9.194	−11.333	−7.055	4.04
(Chaimae et al., 2020)	[37]	−13.569	−15.624	−11.515	4.08
(Kuling et al., 2022)	[38]	−10.463	−12.443	−8.483	4.11
(Gani et al., 2023)	[39]	−7.983	−10.097	−5.869	4.06
(van_Lookeren_Campagne et al., 2022)	[40]	−15.562	−17.726	−13.397	4.03
(Gedela et al., 2024)	[41]	−13.050	−15.060	−11.039	4.10
(Gao et al., 2019)	[42]	−4.353	−7.230	−1.476	3.69
(Iqbal et al., 2023)	[43]	−9.102	−11.092	−7.112	4.11
(Zhao, 2022)	[44]	−15.555	−17.760	−13.350	4.01
(Roman et al., 2021)	[45]	−8.931	−11.381	−6.481	3.90
(Elkaref et al., 2022)	[46]	−9.390	−11.568	−7.213	4.03
(Liu et al., 2018)	[47]	−11.298	−13.787	−8.809	3.88
(Hailu et al., 2023)	[48]	−9.261	−11.272	−7.251	4.10
(Harnmetta et al., 2022)	[49]	−6.309	−8.343	−4.275	4.09
(Walker et al., 2019)	[50]	−7.386	−9.464	−5.308	4.07
(Szarkowska et al., 2021)	[51]	−12.655	−14.698	−10.611	4.09
(Agarwal et al., 2021)	[52]	−8.036	−10.836	−5.235	3.73
(Kumar & Albuquerque, 2021)	[53]	−10.371	−12.881	−7.862	3.87
(Karnysheva & Schwarz, 2020)	[54]	−14.840	−17.350	−12.331	3.87
(Tran et al., 2022)	[55]	−9.603	−11.851	−7.354	3.99
(Alibadi et al., 2019)	[56]	−11.563	−13.629	−9.498	4.08
Theta		**−10.631**	**−11.881**	**−9.381**	
Test of theta = 0; z = −16.66			**Test of homogeneity: Q = chi2(31) = 194.41 Prob > |z| = 0.0000 Prob > Q = 0.000**		

The visual summary of effect sizes and associated confidence intervals (CIs) from several studies, as well as the total pooled effect size, is provided in Figure 3, which is an essential tool in meta-analysis. The forest plot offers important insights into the variability and distribution of study results when working with high heterogeneity measurements, such as a τ^2 value of 8.8724, an I^2 value of 87.65%, and a H^2 score of 8.10. The 95% confidence intervals for each effect magnitude are shown as horizontal lines that extend from them. The accuracy of the effect size estimate is indicated by the length of these lines. The total effect size from all of the included studies is summarized by the diamond at the bottom of the plot. The pooled effect size's 95% confidence interval is represented by the diamond's width. Each study's weight in the meta-analysis, which is frequently represented by the squares'

sizes, shows how much of an impact each study has on the total pooled effect size. This quantifies the variance between studies. A τ^2 score of 8.8724 suggests significant variation in the impact sizes of the studies. A high τ^2 indicates that there is a considerable difference in effect sizes between studies compared to what would be predicted by chance alone. This statistic shows the proportion of overall study variation attributable to heterogeneity as opposed to sampling error. An I^2 score of 87.65% is quite high, suggesting that variation within studies rather than chance error accounts for a significant amount of the range in effect sizes. The ratio of the overall variability to the variability predicted by sampling error alone is measured here. An H^2 value of 8.10 indicates significant heterogeneity across the studies, meaning that the variability seen is significantly higher than what would be predicted by chance.

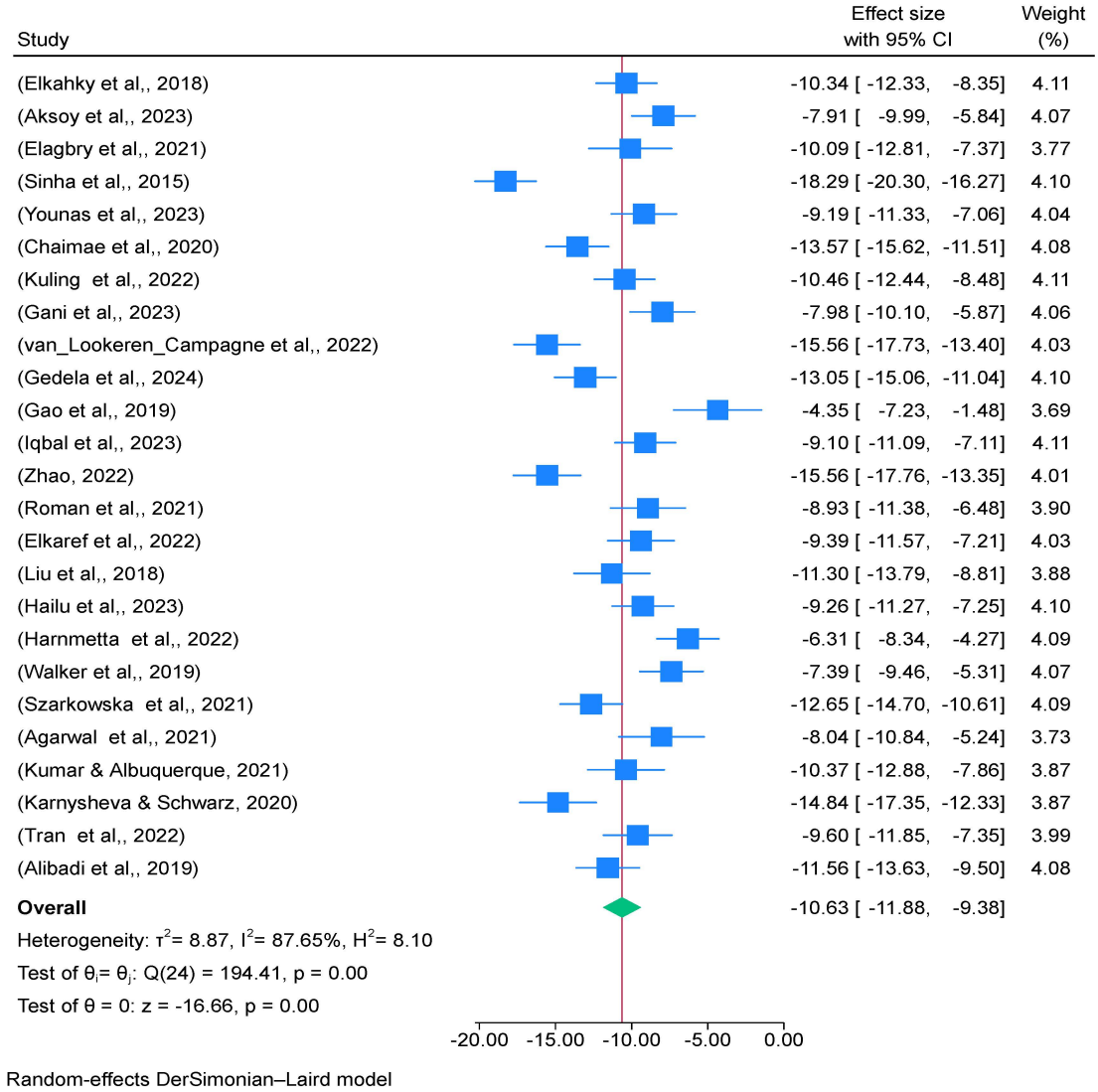


Figure 3. Forest plot for distribution of effect size of meta-analysis summary of CAWE in Table 3 [29,33–56].

Low-resource languages, due to a lack of large, annotated corpora, often face more difficulty in generating high-quality embeddings. Studies on these languages may report poorer performance in resolving MCD, contributing to higher variability compared to studies on high-resource languages, where large datasets and pre-trained models are more readily available.

With an I^2 of 87.65%, the high heterogeneity indicates a substantial diversity throughout the investigation. There are differences in the morphological complexity of different

languages. Word embeddings may face additional difficulties in morphologically rich languages (MRLs), such as Sesotho sa Leboa, Finnish, or Turkish, due to the abundance of word forms (inflections, derivations) that need to be distinguished. Low-resource language studies frequently encounter additional difficulties because of their smaller datasets, scarcity of pre-trained models, and lack of annotated corpora. Conversely, robust pre-trained models like BERT and GPT, as well as larger datasets, are advantageous for high-resource languages like English. How successfully models learn to distinguish between different word senses depends on the quality of the corpus, which includes annotations for tasks such as WSD. Research employing varied, well-annotated datasets will demonstrate greater efficacy in MCD mitigation.

The quality of the embeddings generated can vary significantly depending on the size of the corpus used in each investigation. Models can learn richer, more varied word representations from larger corpora, which can more successfully reduce MCD. Variability may arise from studies that use tiny, domain-specific, or less varied corpora, particularly for low-resource languages. The level of noise, the quality of the annotations, and the corpus's domain-specificity all affect how well models perform. Better WSD outcomes are produced by high-quality, sense-tagged corpora (such as SemCor for English), but inconsistent MCD resolution is introduced by unannotated or low-quality datasets.

The choice of embedding methodologies (static vs. contextual), corpus size and quality, and diversity of languages investigated account for a substantial part of the considerable variation observed in MCD studies. The variance between research is also influenced by other elements like training objectives, evaluation techniques, fine-tuning strategies, and model size. To address this heterogeneity and enable more accurate comparisons between outcomes, methods must be carefully considered and standardized when feasible.

The output of language modeling in Figure 4 is H^2 Score: 4.74, I^2 Score: 78.92%, and τ^2 Value: 4.69. Effect size variability is noteworthy yet mild. Significant variation between trials is indicated by a high I^2 value. A moderate degree of between-study variance is indicated by τ^2. H^2 indicates significant heterogeneity since the observed variance is noticeably higher than would be predicted by chance. The high I^2 score indicates that the outcomes of language modeling research differ significantly. The τ^2 value nevertheless shows substantial variability even though it is not as high as some other subgroups. This can be the result of variations in the techniques, data sources, or evaluation metrics used in various studies. The output for Machine Translation in Figure 4 is τ^2 Value: 28.41, I^2 Value: 94.55%, H^2 Score: 18.34. Very high variability in effect sizes. Extremely high I^2 value suggests that nearly all the variability is due to heterogeneity. The very high τ^2 indicates substantial between-study variance, and the H^2 score indicates that the variability is much higher than expected by chance. Machine translation shows the highest level of heterogeneity among the subgroups. This could be due to diverse language pairs, different translation models, or varying data quality. The high τ^2 and I^2 values suggest that there are significant differences between study results.

The output of Named Entity Recognition (NER) in Figure 4 is τ^2 Value: 0.00, I^2 Value: 0.00%, H^2 Score: 0.00. No variability observed in effect sizes. The absence of variability and heterogeneity indicates that studies are highly consistent with each other. NER shows no evidence of heterogeneity, implying that the effect sizes across studies are very similar. This could be due to the uniformity of datasets and methods used in NER studies. The output of Question Answering in Figure 4 is τ^2 Value: 0.00, I^2 Value: 0.00%, H^2 Score: 0.00. No variability observed in effect sizes. Similar to NER, the absence of variability indicates consistency across studies. The lack of heterogeneity in question answering suggests that studies in this area are quite homogeneous. The uniformity could be due to standard evaluation frameworks or similar approaches used in the studies.

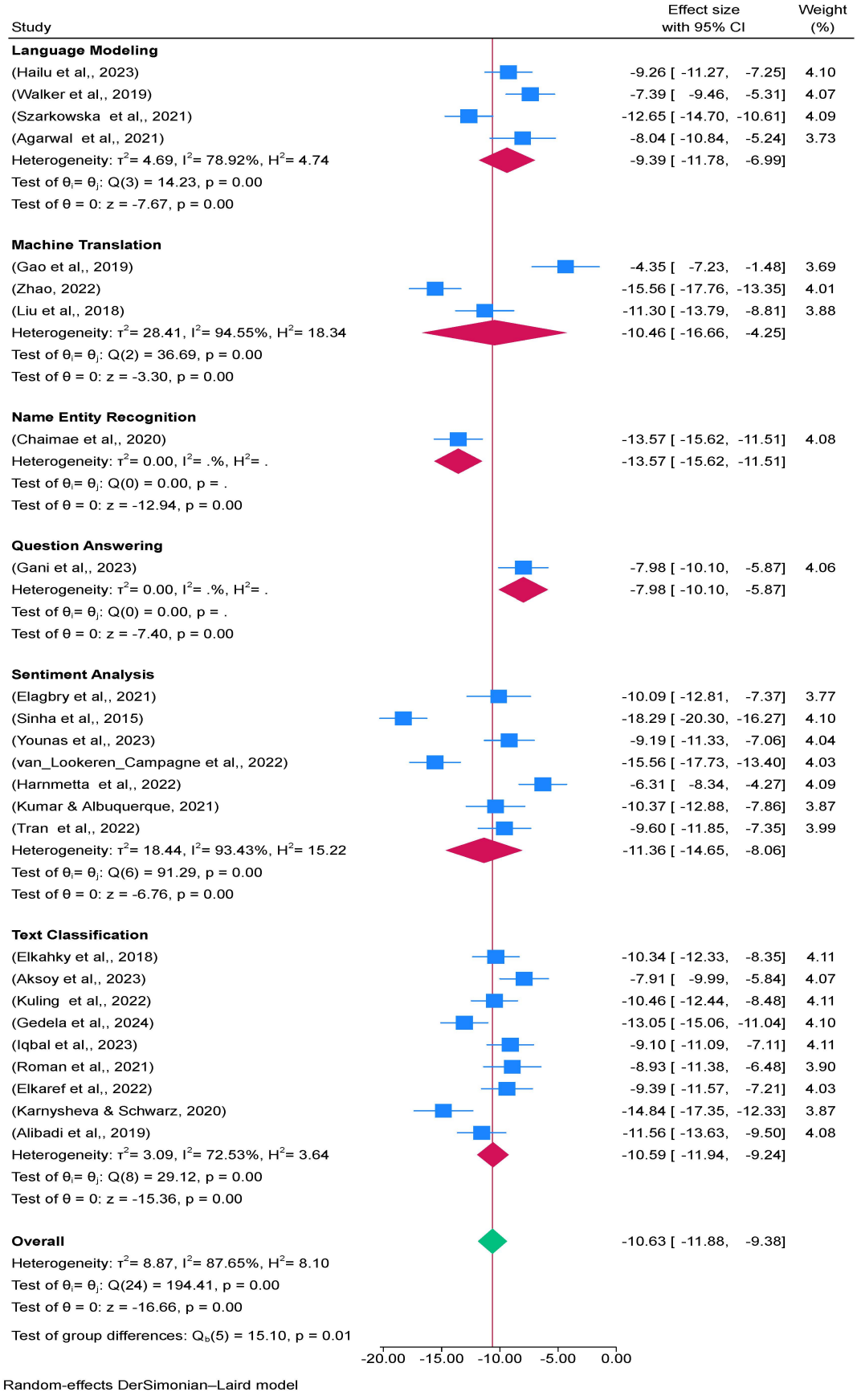


Figure 4. Forest plot with subgroups for distribution of effect size of meta-analysis summary of CAWE in Table 3 [29,33–56].

Figure 4 shows the results of the sentiment analysis: τ^2 Value: 18.44, I^2 Value: 93.44%, and H^2 Score: 15.22. high degree of variation in effect sizes. Significant variability between trials is indicated by very high I^2 and τ^2 values. The observed variation is significantly more than would be predicted by chance, as indicated by the high H^2 score. Significant heterogeneity is also revealed via sentiment analysis. This could be explained by the various datasets, evaluation metrics, and sentiment analysis techniques applied in various studies. Figure 4 displays the results of Text Classification: τ^2 Value: 3.09, I^2 Value: 72.53%, and H^2 Score: 3.64. Variability in effect sizes is moderate. A high I^2 value, however smaller than in some other groupings, indicates significant heterogeneity. The moderate between-study variance is indicated by the τ^2 value. Studies on text categorization exhibit moderate to high variability. This could be the result of the studies' various text kinds, categorization tasks, or algorithms. Variability may be influenced by variations in methods, data sources, or assessment measures, according to high heterogeneity in particular regions. Understanding these differences may be aided by performing subgroup analysis and looking into the sources of heterogeneity further. Strong and comparable results across research are suggested by consistency in NER and Question Answering, which may be the consequence of established procedures or assessment frameworks.

A diagnostic tool used in meta-analyses to evaluate the correlation between effect sizes and accuracy across many studies is a Galbraith plot, often referred to as a Galbraith plot of standardized effect sizes, as shown in Figure 5. It assists in locating studies that might be outliers or that might be disproportionately affecting the overall meta-analysis results. A lower standard error and a more trustworthy effect size estimate are associated with higher precision. The graphic displays lines that show the effect sizes' 95% confidence intervals. These lines aid in the identification of studies whose impact sizes deviate from the pooled effect size's predicted range.

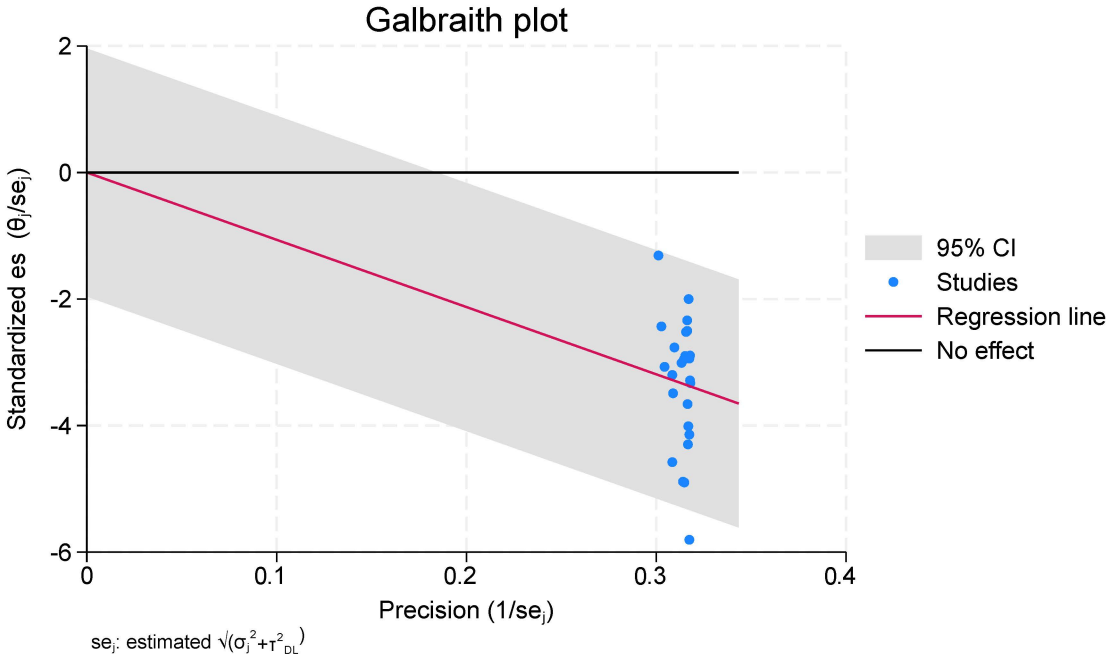


Figure 5. Galbraith Plot.

4.2. Publication Bias and Meta-Regression

In order to make sure that all study findings—rather than simply those that demonstrate successful or good outcomes—are available for evaluation, it is imperative to address publication bias. Complex and occasionally nuanced problems with how language models reflect multiple meanings of words are involved in MCD, especially when it comes to NLP and word embeddings. Due to linguistic complexity or a lack of training data, research on low-resource or morphologically rich languages, as Sesotho sa Leboa, may encounter

greater difficulties. There may be a bias in favor of high-resource languages where models perform better if research that neglect to address MCD in these languages have a lower chance of being published.

The preference for reporting success stories in languages with greater resources (such as English) may cause low-resource language-specific solutions to receive less attention, which would limit the diversity of research and innovation in solving MCD in various linguistic situations. To ensure a full and balanced understanding of the problem, particularly for underrepresented language families and methodological approaches, it is imperative to address publication bias in MCD studies. Studies with negative, null, or ambiguous results should be encouraged to be submitted to journals and conferences since they can offer important information about which treatments are ineffective for treating MCD. These findings can lead to the avoidance of needless replication of unsuccessful studies and pave the way for novel concepts or improved versions of current techniques.

Research on MCD can be severely impacted by publication bias, especially when low-resource or morphologically rich languages are involved. The research community has to embrace strategies like pre-registration, boost transparency, encourage the publication of unfavorable or equivocal results, and provide incentives for studies on underrepresented languages in order to overcome this problem. The field can get a fairer understanding of MCD and create more complete and efficient models for a larger range of languages, including Sesotho sa Leboa, by minimizing publication bias.

In meta-regression, the chi-square test determines if the moderator significantly accounts for variation in effect sizes. There is a larger correlation between the moderator and the effect sizes when the chi-square values are higher. Table 5's chi-square values of 169.11 indicate that the moderator has a considerable impact on the effect sizes. It suggests that a significant portion of the variation in the effect sizes can be explained by the moderator. Table 6's chi-square values of 128.71 show a significant, albeit weaker, correlation between the moderator and effect sizes. A visual aid for assessing publication bias and analyzing the correlation between effect sizes and precision in meta-analyses is the funnel plot in Figure 6. The effect sizes from several research is plotted versus the standard errors. These lines, which together constitute the funnel shape, show the region in which, in the absence of bias, effect sizes should fall 95% of the time. A high chi-square value indicates a large variation in the impact sizes with moderator. If the moderator has a significant impact, you may observe deviations from the typical funnel shape in the funnel plot. The study's observed and predicted frequencies differ significantly, as indicated by the Chi-Square score of 169.11. The fact that the value of 169.11 is so high often indicates that the observed and predicted data differ significantly, indicating that factors other than chance are likely impacting the outcome.

Table 5. Meta-Regression per publication year.

Parameter	Coefficient	Std. Err.	z	$p > \|z\|$	[95% Conf. Interval]	
Year	0.4624064	0.3034431	1.52	0.128	−0.1323312	1.057144
Constant	−945.1428	613.2503	−1.54	0.123	−1106.102	256.8058

Test of residual homogeneity: Q_res = chi2(23) = 169.11, Prob > Q_res = 0.0000.

Table 6. Meta-Regression per Dataset.

Parameter	Coefficient	Std. Err.	z	$p > \|z\|$	[95% Conf. Interval]	
Corpus	-1.05×10^{-7}	3.29×10^{-8}	−3.20	0.001	-1.69×10^{-7}	-4.07×10^{-8}
Constant	−10.2262	0.5565426	−18.37	0.000	−11.317	−9.135397

Test of residual homogeneity Q_res = chi2(23) = 128.71 Prob > Q_res = 0.0000.

In meta-regression, a bubble plot (see Figures 7 and 8) is utilized to investigate the relationship between effect sizes and study variables or moderators. A third variable, like the study weight or sample size, is represented by the size of the bubble. The large chi-square values in the bubble graphic indicate a strong correlation between the moderator

and effect sizes. A distinct pattern or trend can be seen in Figures 7 and 8, suggesting that the moderator has a major impact on the effect sizes. With a Chi-Square score of 128.71, study may conclude that there is a large discrepancy between the observed and anticipated frequencies, indicating a considerable deviation from the null hypothesis. In general, this high score indicates that it is improbable that the discrepancies are the result of pure chance.

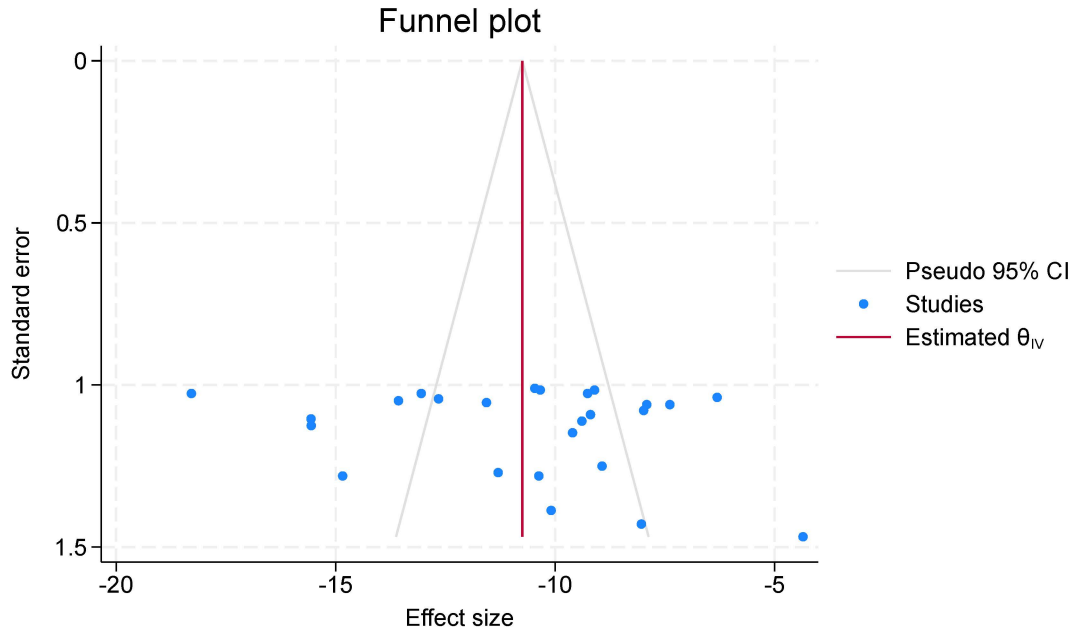


Figure 6. Funnel Plot.

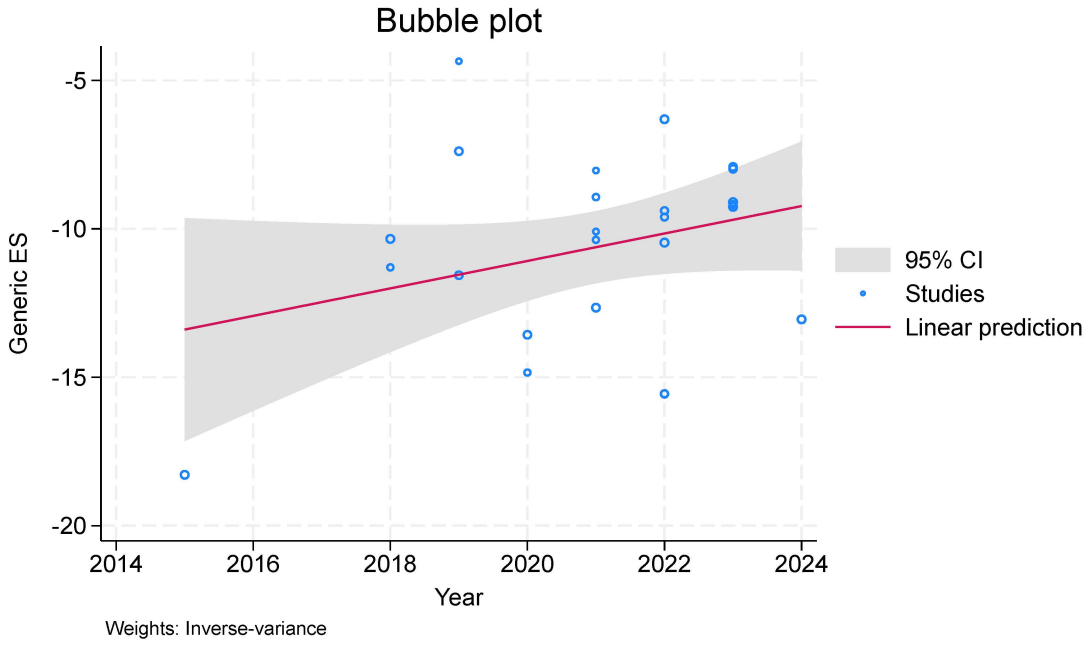


Figure 7. Bubble Plot per Publication year.

In meta-analyses, Trim and Fill Analysis in Table 7 is employed to evaluate publication bias. It estimates what the overall effect size might be if these missing studies were included, taking into consideration the chance that certain studies might be missing from the meta-analysis—a common result of publication bias. The estimated impact size, after correcting for publication bias, is 5.95 beta values (β). It suggests a significant effect size. The beta value of 4.767 has a rather high standard error (SE), indicating a significant degree of

uncertainty surrounding the beta estimate. The Z-value is not statistically significant at typical thresholds, as indicated by the *p*-value of 0.2. This indicates that the null hypothesis—that the impact magnitude is zero—cannot be rejected due to insufficient.

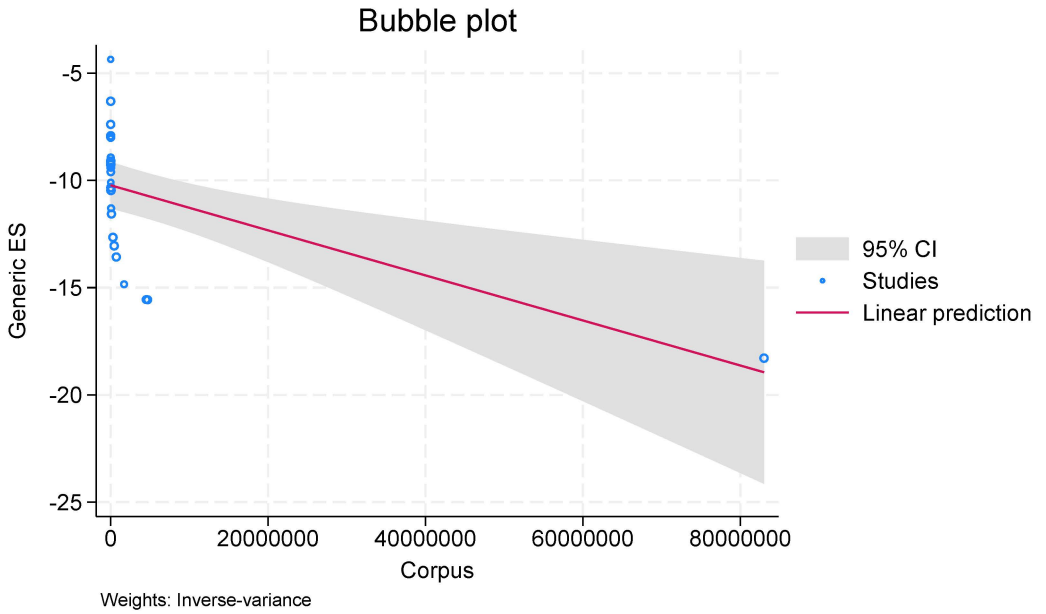


Figure 8. Bubble Plot for Dataset.

Table 7. Meta Trimfill Analysis of Public Bias.

Studies (n = 28)	Coefficient	[95% Conf. Interval]	
Observed (n = 28)	−10.747	−11.185	−10.310
Observed + Imputed (25 + 3)	−10.747	−11.185	−10.310

H0: β1 = 0; no small-study effects: β1 = 5.94 SE of β1 = 4.767 z = 1.25 Prob > |z| = 0.2124.

Although the provided beta value of 5.95 is high, the high standard error of 4.767 suggests a significant degree of uncertainty. It is suggested that the effect size is not statistically significant by the Z-value of 1.25 and the *p*-value of 0.2. This suggests that the effect magnitude is not much different from zero even after accounting for possible publication bias. Although the high standard error indicates that this estimate is not very exact, the large beta value raises the possibility of a significant effect. The *p*-value shows that there is insufficient evidence to support a significant effect. This may be because there is not enough research or because effect sizes vary a lot.

4.3. Descriptive Statistics of Primary Studies

A dataset of articles pertinent to meaning conflation deficiency for low-resource languages from 2014 to early 2024 was acquired to help the research. Academic journals, research articles, and conference proceedings from IEEE Xplore, Scopus, Google Scholar, Springer, and Web of Science were used to collect the data. The metadata, timestamp, accuracy, meaning conflation deficit, contextual word embeddings, word embeddings, context-aware word embedding approaches, and year of publication were all included, the researchers made sure. The publishing patterns from 2015 to 2024 are displayed in Figure 9. From 2018 to 2022, over 100 articles show a significant increase in interest in meaning conflation insufficiency. However, in other years there was a decline in publications. ACM Transactions on Asian and Low-Resource Languages Information Processing journal has increased, as seen in Figure 10. It should be noted that data collected as of early 2024 was used in the research. Consequently, there's a strong possibility that additional publications will be made in journals that are released in the latter months of 2024.

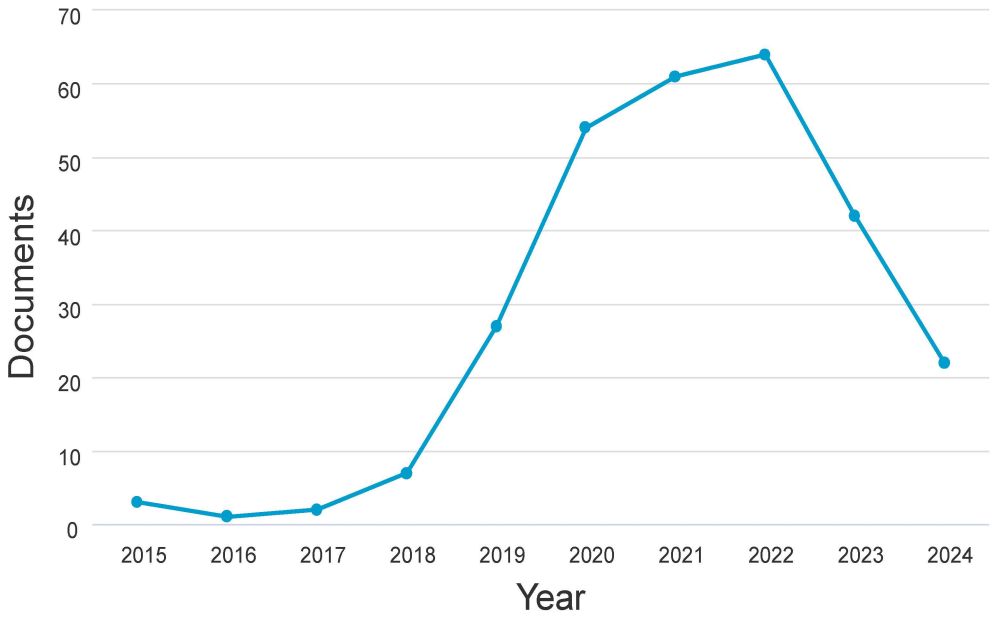


Figure 9. Publications by Year.

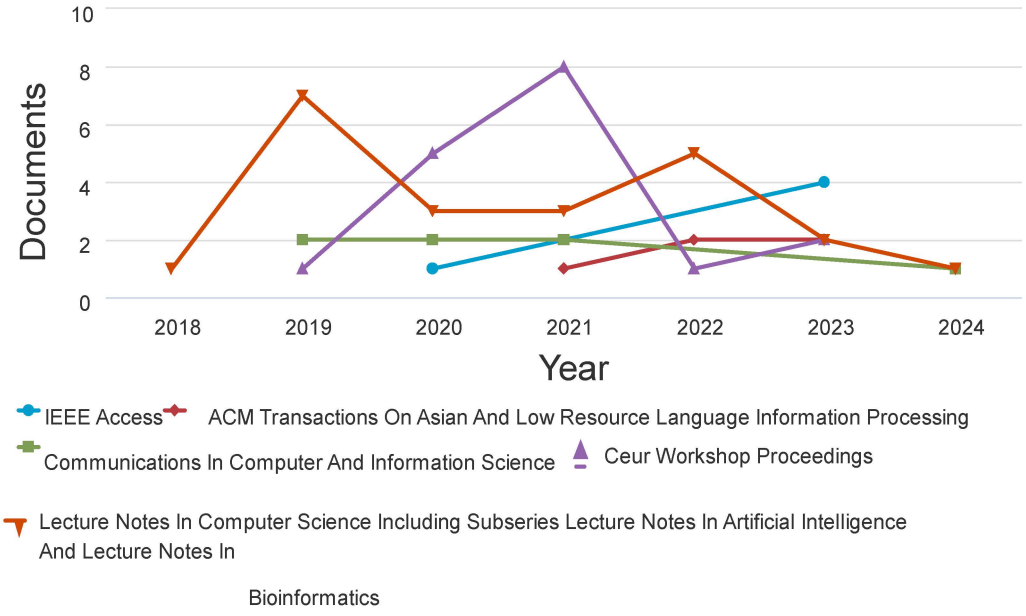


Figure 10. Publication per Year by Source.

As seen in Figure 11, the researchers also carefully examined or analyzed subjects within text data from various nations, offering insights into the speech and thematic patterns unique to each nation. The graphic presentation shows the advancements made in the study of meaning conflation deficit for languages with limited resources. According to this data, the United States is currently in the lead, followed by China and India.

Based on citation metrics, Figure 12 shows bibliometric study that investigates the network of significant writers in the field of network visualization. The study finds important contributors, patterns, and cooperative networks in this field of study by examining citation data. designed to show the connections between writers according to the number of citations they have received. Citation relationships between authors are represented by edges, which are nodes.

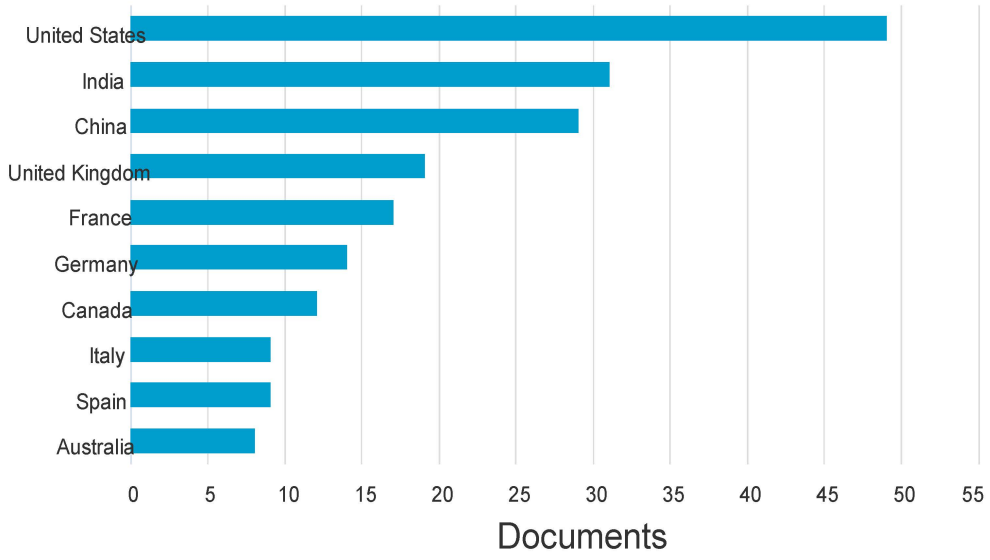


Figure 11. MCD by Country.

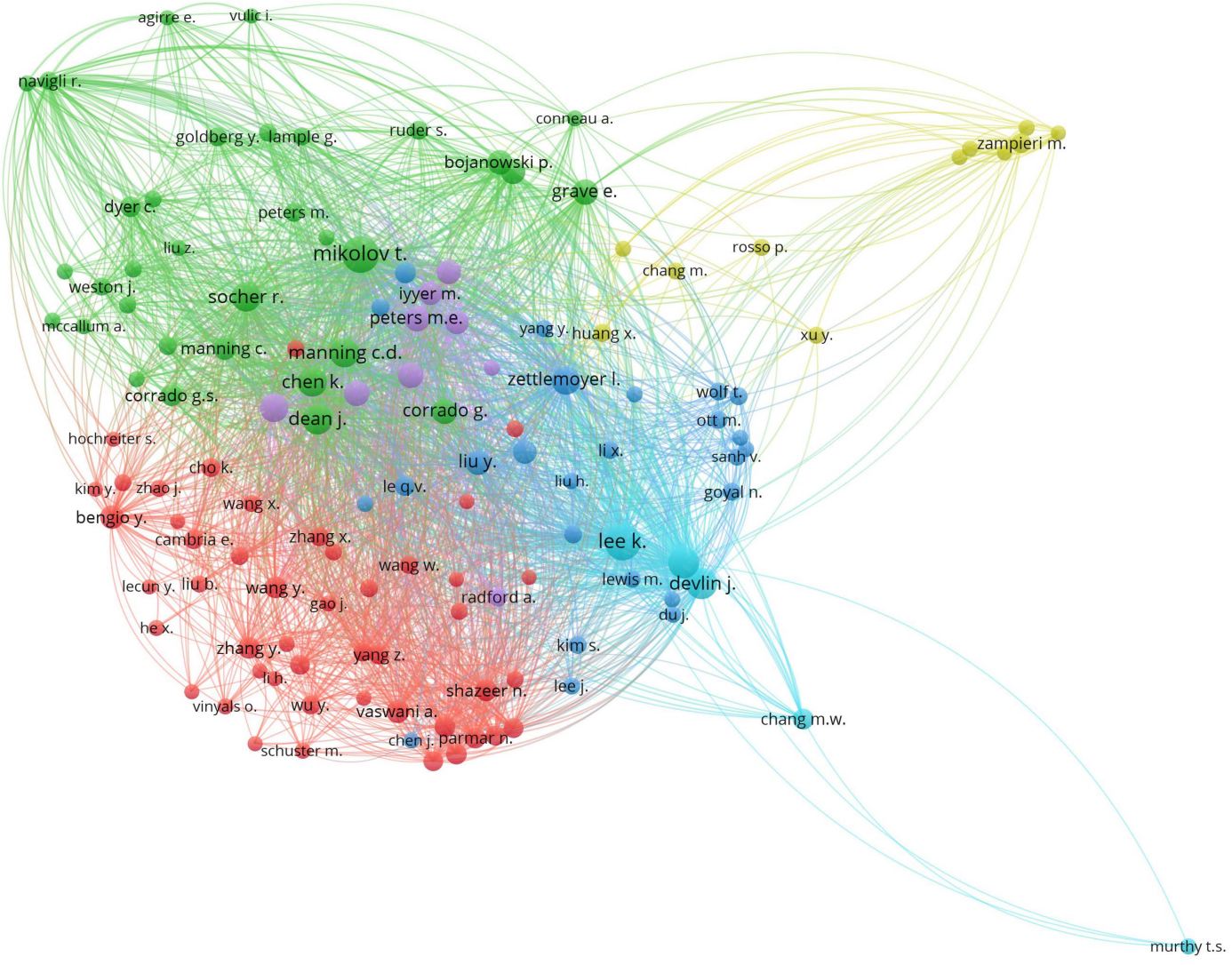


Figure 12. Bibliometric Analysis on Network Visualization per cited Authors.

Given their success in capturing the subtleties of meaning in many situations, contextual word embeddings and contextual-aware models—which are shown in Figure 13—have attracted a lot of interest. These models, which include BERT, GPT, and their derivatives, help solve the issue of word sense disambiguation and meaning conflation by dynamically adjusting the embeddings of words based on their surrounding context. On the other hand, hybrid models, which incorporate more features or blend many embedding types, may not have as many publications, but they can still be quite effective. Their goal is to utilize the advantages of several methods to tackle certain problems related to meaning resolution and word embeddings.

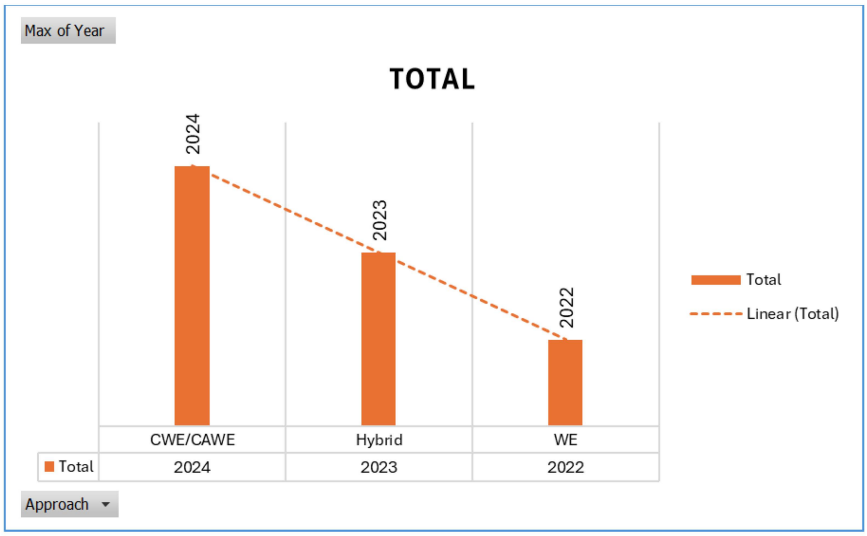


Figure 13. Approaches per year.

An intriguing pattern can be seen in Figure 14, where the emphasis on sentiment analysis and text categorization is in line with their immediate relevance and wide range of applications. The surge in research activity that occurred in 2021 and 2022 may have been caused by the quick development of new models and methods at that time. Research on meaning conflation insufficiency may have declined starting in 2023 for a variety of reasons, including a change in research goals, the development of current techniques, or a focus on other newly developing NLP difficulties. It might also mean that scholars are investigating novel avenues or uses for word embeddings, which have gained popularity lately.

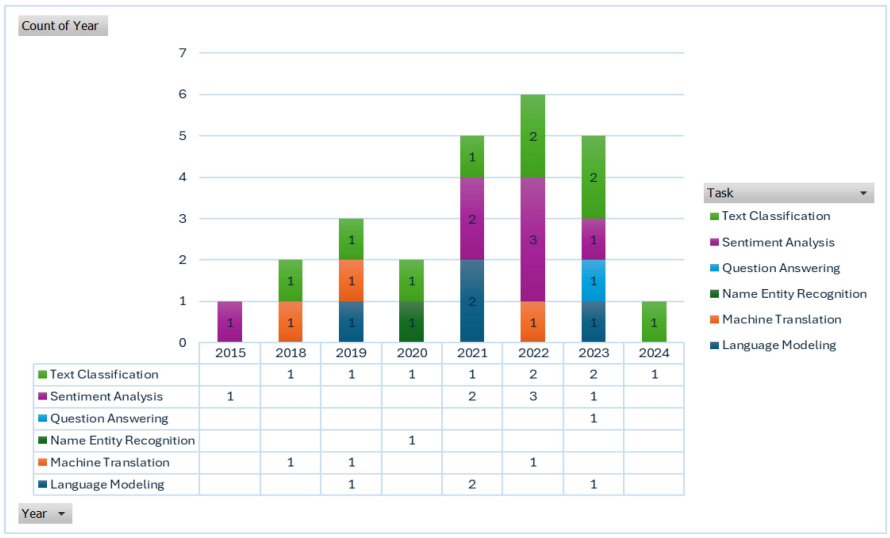


Figure 14. Models per year.

Higher citation rates in ACM publications, as shown in Figure 15, indicate that the computer science and engineering community place a high value on research addressing meaning conflation deficit. Given that ACM is renowned for emphasizing cutting-edge computer research, this suggests that the study has a great deal of relevance to current advancements in the field. Elsevier and MDPI publications are well-known for serving a wider range of academic readers.

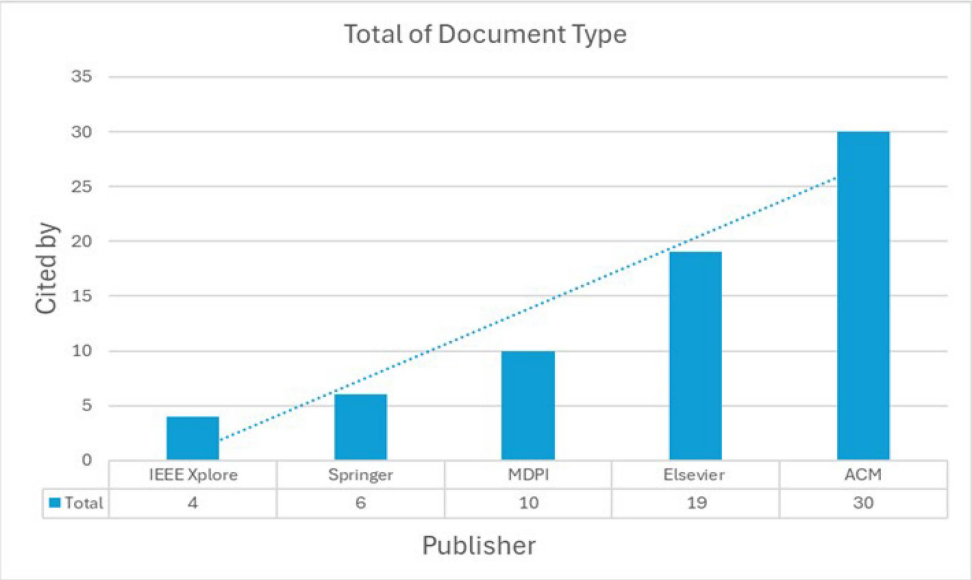


Figure 15. Publisher per citations.

5. Conclusions

In summary, there is a lot of diversity and uncertainty in the available findings, even though context-aware embeddings show promise for treating Meaning Conflation Deficiency. To provide more trustworthy and broadly applicable results, future research should try to overcome these problems by enhanced study designs, investigation of different strategies, and more comprehensive analysis. The estimated variance of the genuine impact sizes across studies is represented by the τ^2 value of 8.8724. This high number suggests that there is a great deal of variation in the effect sizes, indicating that the results of the studies that were part of the meta-analysis differ significantly. The entire variation among trials owing to heterogeneity rather than chance is reflected in the I^2 value of 87.65%. This high I^2 score indicates significant heterogeneity between the research, indicating that study methods may have an impact on the effect sizes, which vary greatly. The degree of variability in effect sizes above and beyond what would be predicted by sampling error alone is indicated by the H^2 score of 8.10. An elevated H^2 score suggests a significant degree of variability in effect sizes, hence supporting the conclusions drawn from τ^2 and I^2.

Even though the beta value of 5.95 is high, the effect size estimate appears to be unclear due to the high standard error. Although the variability and uncertainty raise doubts about the dependability of the effect, the high β value indicates that there may be a significant one. Given the comparatively significant standard error of 4.767, it may be concluded that the beta estimate's precision is low. The large confidence intervals and lack of statistical significance are caused by this. When expressed as β/SE, the Z-value of 1.25 denotes that the effect size estimate is not statistically different from zero. This bolsters the results' lack of importance. It appears that the effect size is not statistically significant based on the *p*-value of 0.2.

Several important conclusions may be drawn from the comprehensive literature assessment on applying context-aware embeddings to resolve meaning confusion deficit:

- Excessive τ^2, I^2, and H^2 values in the meta-analysis suggest significant variability in the effect sizes. This indicates that there is considerable variation in the efficacy of

contextually aware embeddings in treating Meaning Conflation Deficiency among research studies.

- Although there may have been publication bias, the effect magnitude is still not statistically significant, according to the trim and fill analysis. This implies that while contextually aware embeddings may be useful, there is uncertainty and a lack of consistency in their effects on Meaning Conflation Deficiency.
- The uncertainty surrounding the effect size estimations is highlighted by the substantial standard error and non-significant *p*-value. To get more accurate and trustworthy projections, more study is required.

Understanding of the sources of study heterogeneity should be the main goal of future research. To find the elements causing the variation in effect sizes, this may entail investigating various study parameters, approaches, and environments. Examine different models and methods to deal with Meaning Conflation Deficiency. While context-aware embeddings show promise, more sophisticated neural network architectures or hybrid models that combine many embedding types may provide superior outcomes.

Meaning Conflation Deficiency (MCD) and word embeddings research has various major implications for natural language processing (NLP) systems, especially when studying morphologically rich and low-resource languages. When used, these discoveries may influence language models' design and enhance their practical uses. Due to the confusion of word meanings, existing NLP systems frequently have trouble handling WSD, particularly in morphologically rich and low-resource languages. This has an impact on activities that come next, like semantic analysis, machine translation, information retrieval, and question answering. These results can be used to combine multi-prototype embeddings, which assign distinct representations to words dependent on their senses in context, and context-aware word embeddings into NLP systems. This would drastically lower MCD.

MCD-caused ambiguities are a common problem for machine translation systems, especially for low-resource and morphologically rich languages. When a word with several meanings is misconstrued by the model, this results in erroneous translations. Using these findings, multi-sense embeddings—which dynamically modify word meaning based on surrounding context—could be incorporated into neural machine translation (NMT) systems to better manage polysemy and homophones. Translations would become more precise and appropriate for the context as a result.

Polysemous words are currently difficult for conversational agents and chatbots to grasp, especially when they are employed in casual or colloquial situations. This causes misunderstandings and irrelevant responses. Conversational AI may comprehend the intended meaning of words based on the discussion context by utilizing advances in multi-prototype embeddings, leading to more natural and context-aware interactions. NLP systems will become more reliable, context-aware, and able to handle a variety of linguistic phenomena by tackling these real-world issues. This will result in more precise and significant language exchanges.

Author Contributions: Conceptualization, M.A.M. and H.D.M.; methodology M.A.M. and H.D.M.; formal analysis M.A.M. and H.D.M.; data curation, M.A.M. and H.D.M.; writing—original draft preparation, M.A.M. and H.D.M.; writing—review and editing M.A.M. and H.D.M.; visualization, M.A.M. and H.D.M.; supervision, S.O.O.; project administration, 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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The material used in this study is publicly available from the sourced databases.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ansell, A.; Bravo-Marquez, F.; Pfahringer, B. PolyLM: Learning about polysemy through language modeling. In Proceedings of the EACL 2021—16th Conference of the European Chapter of the Association for Computational Linguistics, Kyiv, Ukraine, 19–23 April 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 563–574. [CrossRef]
2. Pilehvar, M.T. On the Importance of Distinguishing Word Meaning Representations: A Case Study on Reverse Dictionary Mapping. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 2151–2156. [CrossRef]
3. Cer, D.; Yang, Y.; Kong, S.; Hua, N.; Limtiaco, N.; John, R.S.; Constant, N.; Guajardo-Céspedes, M.; Yuan, S.; Tar, C.; et al. Universal sentence encoder for English. In Proceedings of the EMNLP 2018—Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2–4 November 2018; pp. 169–174. [CrossRef]
4. Masethe, H.D.; Masethe, M.A.; Ojo, S.O.; Giunchiglia, F.; Owolawi, P.A. Word Sense Disambiguation for Morphologically Rich Low-Resourced Languages: A Systematic Literature Review and Meta-Analysis. *Information* **2024**, *15*, 540. [CrossRef]
5. Ansell, A. Contextualised Approaches to Embedding Word Senses. The University of Waikato. 2020. Available online: http://researchcommons.waikato.ac.nz/ (accessed on 21 August 2024).
6. Pilehvar, M.T.; Collier, N. De-conflated semantic representations. In Proceedings of the EMNLP 2016—Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 1680–1690. [CrossRef]
7. Zhang, T.; Ye, W.; Xi, X.; Zhang, L.; Zhang, S.; Zhao, W. Leveraging human prior knowledge to learn sense representations. *Front. Artif. Intell. Appl.* **2020**, *325*, 2306–2313. [CrossRef]
8. Yang, X.; Mao, K. Learning multi-prototype word embedding from single-prototype word embedding with integrated knowledge. *Expert Syst. Appl.* **2016**, *56*, 291–299. [CrossRef]
9. Won, H.; Lee, H.; Kang, S. Multi-prototype Morpheme Embedding for Text Classification. In Proceedings of the SMA 2020: The 9th International Conference on Smart Media and Applications, Jeju, Republic of Korea, 17–19 September 2020; pp. 295–300. [CrossRef]
10. Li, N.; Bouraoui, Z.; Camacho-Collados, J.; Espinosa-anke, L.; Gu, Q.; Schockaert, S. Modelling General Properties of Nouns by Selectively Averaging Contextualised Embeddings. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), Montreal, QC, Canada, 19–25 August 2021. [CrossRef]
11. Biesialska, M.; Costa-Jussà, M.R. Refinement of unsupervised cross-lingualword embeddings. *Front. Artif. Intell. Appl.* **2020**, *325*, 1978–1981. [CrossRef]
12. da Silva, J.R.; Caseli, H.d.M. Generating Sense Embeddings for Syntactic and Semantic Analogy for Portuguese. *arXiv* **2020**. [CrossRef]
13. Rodrigues da Silva, J.; Caseli, H.d.M. Sense representations for Portuguese: Experiments with sense embeddings and deep neural language models. *Lang. Resour. Eval.* **2021**, *55*, 901–924. [CrossRef]
14. Ilie, V.I.; Truica, C.O.; Apostol, E.S.; Paschke, A. Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings. *IEEE Access* **2021**, *9*, 162122–162146. [CrossRef]
15. Vusak, E.; Kuzina, V.; Jovic, A. A Survey of Word Embedding Algorithms for Textual Data Information Extraction. In Proceedings of the 2021 44th International Convention on Information, Communication and Electronic Technology, MIPRO 2021, Opatija, Croatia, 27 September–1 October 2021; pp. 181–186. [CrossRef]
16. Hu, R.; Li, S.; Liang, S. Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3899–3908. [CrossRef]
17. Katsarou, S.; Rodríguez-gálvez, B.; Shanahan, J. Measuring Gender Bias in Contextualized Embeddings †. *Comput. Sci. Math. Forum* **2022**, *3*, 3. [CrossRef]
18. Balakrishnan, V.; Shi, Z.; Law, C.L.; Lim, R.; Teh, L.L.; Fan, Y.; Periasamy, J. A Comprehensive Analysis of Transformer-Deep Neural Network Models in Twitter Disaster Detection. *Mathematics* **2022**, *10*, 4664. [CrossRef]
19. Loureiro, D.; Jorge, A.M. LIAAD at SemDeep-5 challenge: Word-in-Context (WiC). In Proceedings of the 5th Workshop on Semantic Deep Learning, SemDeep 2019, Macau, China, 12 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 51–55. Available online: https://aclanthology.org/W19-5801/ (accessed on 10 September 2024).
20. Li, X.; Lei, Y.; Ji, S. BERT- and BiLSTM- Based Sentiment Analysis of Online. *Futur. Internet* **2022**, *14*, 332. [CrossRef]
21. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–11. Available online: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html (accessed on 10 September 2024).
22. Kavatagi, S.; Rachh, R. A Context Aware Embedding for the Detection of Hate Speech in Social Media Networks. In Proceedings of the 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), IEEE, Pune, India, 29–30 October 2021; pp. 1–4. [CrossRef]
23. Loureiro, D.; Mário Jorge, A.; Camacho-Collados, J. LMMS reloaded: Transformer-based sense embeddings for disambiguation and beyond. *Artif. Intell.* **2022**, *305*, 103661. [CrossRef]
24. Li, X.; Fu, X.; Xu, G.; Yang, Y.; Wang, J.; Jin, L.; Liu, Q.; Xiang, T. Enhancing BERT Representation With Context-Aware Embedding for Aspect-Based Sentiment Analysis. *IEEE Access* **2020**, *8*, 46868–46876. [CrossRef]

25. Liu, J.; Zhang, Z.; Lu, X. Aspect Sentiment Classification via Local Context- Focused Syntax Based on DeBERTa. In Proceedings of the 2024 4th International Conference on Computer Communication and Artificial Intelligence (CCAI), IEEE, Xi'an, China, 24–26 May 2024; pp. 297–302. [CrossRef]
26. Martin, C.; Yang, H.; Hsu, W. KDDIE at SemEval-2022 Task 11: Using DeBERTa for Named Entity Recognition. In Proceedings of the SemEval 2022—16th International Workshop on Semantic Evaluation, Seatle, WA, USA, 14–15 July 2022; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 1531–1535. [CrossRef]
27. Kumar, N.; Kumar, S. Enhancing Abstractive Text Summarisation Using Seq2Seq Models: A Context-Aware Approach. In Proceedings of the 2024 International Conference on Automation and Computation (AUTOCOM), IEEE, Dehradun, India, 14–16 March 2024; pp. 490–496. [CrossRef]
28. Alessio, I.D.; Quaglieri, A.; Burrai, J.; Pizzo, A.; Aitella, U.; Lausi, G.; Tagliaferri, G.; Cordellieri, P.; Cricenti, C.; Mari, E.; et al. Behavioral sciences 'Leading through Crisis': A Systematic Review of Institutional Decision-Makers in Emergency Contexts. *Behav. Sci.* **2024**, *14*, 481. [CrossRef]
29. Necula, S.C.; Dumitriu, F.; Greavu-Șerban, V. A Systematic Literature Review on Using Natural Language Processing in Software Requirements Engineering. *Electronics* **2024**, *13*, 2055. [CrossRef]
30. Hladek, D.; Stas, J.; Pleva, M.; Ondas, S.; Kovacs, L. Survey of the Word Sense Disambiguation and Challenges for the Slovak Language. In Proceedings of the 17th IEEE International Symposium on Computational Intelligence and Informatics, IEEE, Budapest, Hungary, 17–19 November 2016; pp. 225–230. [CrossRef]
31. Thompson, R.C.; Joseph, S.; Adeliyi, T.T. A Systematic Literature Review and Meta-Analysis of Studies on Online Fake News Detection. *Information* **2022**, *13*, 527. [CrossRef]
32. Bowring, A.; Telschow, F.J.E.; Schwartzman, A.; Nichols, T.E. NeuroImage Confidence Sets for Cohen's deffect size images. *Neuroimage* **2021**, *226*, 117477. [CrossRef]
33. Elkahky, A.; Webster, K.; Andor, D.; Pitler, E. A Challenge Set and Methods for Noun-Verb Ambiguity. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 2562–2572. [CrossRef]
34. Aksoy, M.; Yanık, S.; Amasyali, M.F. A comparative analysis of text representation, classification and clustering methods over real project proposals. *Int. J. Intell. Comput. Cybern.* **2023**, *16*, 6. [CrossRef]
35. Elagbry, H.E.; Attia, S.; Abdel-Rahman, A.; Abdel-Ate, A.; Girgis, S. A Contextual Word Embedding for Arabic Sarcasm Detection with Random Forests. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Online, 19 April 2021; pp. 340–344. Available online: https://aclanthology.org/2021.wanlp-1.43 (accessed on 10 September 2024).
36. Sinha, A.; Shen, Z.; Song, Y.; Ma, H.; Eide, D.; Hsu, B.-J.P.; Wang, K. An Overview of Microsoft Academic Service (MAS) and Applications. In Proceedings of the WWW'15: 24th International World Wide Web Conference, Florence, Italy, 18–22 May 2015; ACM Digital Library: Florence, Italy, 2019; pp. 243–246. [CrossRef]
37. Chaimae, A.; Yacine, E.Y.; Rybinski, M.; Montes, J.F.A. BERT for Arabic Named Entity Recognition. In Proceedings of the 2020 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Marrakech, Morocco, 25–27 November 2020; pp. 1–6. [CrossRef]
38. Kuling, G.; Curpen, B.; Martel, A.L. BI-RADS BERT and Using Section Segmentation to Understand Radiology Reports. *Imaging* **2022**, *8*, 131. [CrossRef]
39. Gani, M.O.; Ayyasamy, R.K.; Fui, Y.T. Bloom's Taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique. *Educ. Inf. Technol.* **2023**, *28*, 15893–15914. [CrossRef]
40. Campagne, R.V.L.; van Ommen, D.; Rademaker, M.; Teurlings, T.; Frasincar, F. DCWEB-SOBA: Deep Contextual Word Embeddings-Based Semi-automatic Ontology Building for Aspect-Based Sentiment Classification. In Proceedings of the Semantic Web: 19th Internatinal Conference, Hersonissos, Greece, 29 May–2 June 2022; ACM Digital Library: Hersonissos, Greece, 2022; pp. 183–199. [CrossRef]
41. Gedela, R.T.; Baruah, U.; Soni, B. Deep Contextualised Text Representation and Learning for Sarcasm Detection. *Arab. J. Sci. Eng.* **2024**, *49*, 3719–3734. [CrossRef]
42. Zhang, F.; Gao, W.; Fang, Y. News title classification based on sentence-LDA model and word embedding. In Proceedings of the 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 8–10 November 2019; pp. 237–240. [CrossRef]
43. Mehedi, K.; Fahim, H.; Moontaha, M.; Rahman, M.; Rhythm, E.R. Comparative Analysis of Traditional and Contextual Embedding for Bangla Sarcasm Detection in Natural Language Processing. In Proceedings of the 2023 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), Malang, Indonesia, 23–25 November 2023; pp. 293–299. [CrossRef]
44. Zhao, C. Multi-Feature Fusion Machine Translation Quality Evaluation Based on LSTM Neural Network. In Proceedings of the 2022 6th Asian Conference on Artificial Intelligence Technology (ACAIT), Changzhou, China, 4–6 November 2022. [CrossRef]
45. Roman, M.; Shahid, A.; Uddin, M.I.; Hua, Q.; Maqsood, S. Exploiting Contextual Word Embedding of Authorship and Title of Articles for Discovering Citation Intent Classification. *Complexity* **2021**, *2021*, 1–13. [CrossRef]
46. Elkaref, N.; Abu-Elkheir, M. GUCT at Arabic Hate Speech 2022: Towards a Better Isotropy for Hatespeech Detection. Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, Online, 20–25 June 2022; European Language Resources Association: Marseille, France, 2022. Available online: https://aclanthology.org/2022.osact-1.27/ (accessed on 30 July 2024).

47. Liu, F.; Lu, H.; Neubig, G. Handling homographs in neural machine translation. *arXiv* **2017**, arXiv:1708.06510.
48. Hailu, B.M.; Assabie, A.; Yenewondim, B.S. Semantic Role Labeling for Amharic Text Using Multiple Embeddings and Deep Neural Network. *IEEE Access* **2021**, *11*, 33274–33295. [CrossRef]
49. Harnmetta, P.; Samanchuen, T. Sentiment Analysis of Thai Stock Reviews Using Transformer Models. In Proceedings of the 2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE), Bangkok, Thailand, 22–25 June 2022. [CrossRef]
50. Walker, N.; Peng, Y.-T.; Cakmak, M. Neural Semantic Parsing with Anonymization for Command Understanding in General-Purpose Service Robots. In Proceedings of the RoboCup 2019: Robot World Cup XXIII, Sydney, NSW, Australia, 2–8 July 2019; ACM: Berlin/Heidelberg, Germany, 2019; pp. 337–350. [CrossRef]
51. Hang, G.; Liu, J. Big-Data Based English-Chinese Corpus Collection and Mining and Machine Translation Framework. In Proceedings of the 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 1–13 November 2021; pp. 418–421. [CrossRef]
52. Agarwal, N.; Sikka, G.; Awasthi, L.K. Web Service Clustering Technique based on Contextual Word Embedding for Service Representation. In Proceedings of the International Conference on Technological Advancements and Innovations (ICTAI), Tashkent, Uzbekistan, 10–12 November 2021. [CrossRef]
53. Kumar, A.; Albuquerque, V.H.C. Sentiment Analysis Using XLM-R Transformer and Zero-shot Transfer Learning on Resource-poor Indian Language. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* **2021**, *20*, 1–13. [CrossRef]
54. Karnysheva, P.S.A. TUE at SemEval-2020 Task 1: Detecting Semantic Change by Clustering Contextual Word Embeddings. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona: International Committee for Computational Linguistics, Online, 12–13 December 2020; pp. 232–238. [CrossRef]
55. Tran, O.T.; Phung, A.C.; Ngo, B.X. Using Convolution Neural Network with BERT for Stance Detection in Vietnamese. In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), Marseill, France, 20–25 June 2022; European Language Resources Association (ELRA): Marseill, France, 2022; pp. 7220–7225. Available online: https://aclanthology.org/2022.lrec-1.783.pdf (accessed on 10 September 2024).
56. Alibadi, Z.; Du, M.; Vidal, J.M. Using pre-trained embeddings to detect the intent of an email. In Proceedings of the 7th ACIS International Conference on Applied Computing and Information Technolog, Honolulu, HI, USA, 29–31 May 2019; Association for Computing Machinery: Honolulu, HI, USA, 2019. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.