data

MDPI

Article

Context Sensitive Verb Similarity Dataset for Legal Information Extraction

Gathika Ratnayaka *, Nisansa de Silva, Amal Shehan Perera, Gayan Kavirathne, Thirasara Ariyarathna and Anjana Wijesinghe

Department of Computer Science and Engineering, University of Moratuwa, Moratuwa 10400, Sri Lanka; nisansadds@cse.mrt.ac.lk (N.d.S.); shehan@cse.mrt.ac.lk (A.S.P.); gayankavirathne.14@cse.mrt.ac.lk (G.K.); thirasara.14@cse.mrt.ac.lk (T.A.); anjana.21@cse.mrt.ac.lk (A.W.)
* Correspondence: gathika.14@cse.mrt.ac.lk

Abstract: Existing literature demonstrates that verbs are pivotal in legal information extraction tasks due to their semantic and argumentative properties. However, granting computers the ability to interpret the meaning of a verb and its semantic properties in relation to a given context can be considered as a challenging task, mainly due to the polysemic and domain specific behaviours of verbs. Therefore, developing mechanisms to identify behaviors of verbs and evaluate how artificial models detect the domain specific and polysemic behaviours of verbs can be considered as tasks with significant importance. In this regard, a comprehensive dataset that can be used as an evaluation resource, as well as a training data set, can be considered as a major requirement. In this paper, we introduce *LeCoVe*, which is a verb similarity dataset intended towards facilitating the process of identifying verbs with similar meanings in a legal domain specific context. Using the dataset, we evaluated both domain specific and domain generic embedding models, which were developed using state-of-the-art word representation and language modelling techniques. As a part of the experiments carried out using the announced dataset, Sense2Vec and BERT models were trained using a corpus of legal opinion texts in order to capture domain specific behaviours. In addition to *LeCoVe*, we demonstrate that a neural network model, which was developed by combining semantic, syntactic, and contextual features that can be obtained from the outputs of embedding models, can perform comparatively well, even in a low resource scenario.

Keywords: information semantics; word embeddings; deep learning; natural language processing

Citation: Ratnayaka, G.; de Silva, N.; Perera, A.S.; Kavirathne, G.; Ariyarathna, T.; Wijesinghe, A. Context Sensitive Verb Similarity Dataset for Legal Information Extraction. *Data* **2022**, *7*, 87. https://doi.org/10.3390/data7070087

Academic Editors: Emma Tonkin and Kristina Yordanova

Received: 3 April 2022
Accepted: 25 May 2022
Published: 28 June 2022

1. Introduction

Legal opinion texts consisting of descriptions on: incidents, involved parties, legal opinions, and judgements related to previous court cases, are considered as an integral part of case law. The information available in these documents can be used in various forms such as arguments, counter-arguments, justifications, and evidence. Therefore, developing automated mechanisms to understand and extract important information from legal opinion texts can be considered as an important step in bringing artificial intelligence into the legal domain. Semantics within a sentence as well as relationships between sentences play a crucial role when it comes to bringing context to automatic legal information extraction [1,2]. When it comes to sentences and their rhetorical structures, verbs are pivotal, as they have a significant impact on the meaning of a sentence due to their semantic and syntactic properties [3–5]. The argument structure of verbs can facilitate several legal information extraction tasks such as argument extraction [6], semantic role labelling, sentiment analysis [7], and discourse analysis. Moreover, verbs can also be considered as an instrumental part of determining the semantics of an event and how different parties are related to an event [5]. Identifying such semantics has a significant value, especially when it comes to extracting information from legal opinion texts, as much

emphasis is given to the events/incidents related to the particular court case, involved parties and actions performed by each party.

Example 1.

- *Sentence 1.1: Lee has demonstrated that he was prejudiced by his counsel's erroneous advice.*
- *Sentence 1.2: The Government contends that Lee cannot make that showing because he was going to be deported either way; going to trial would only result in a longer sentence before that inevitable consequence.*
- *Sentence 1.3: Applying the two-part test for ineffective assistance claims from Strickland v. Washington, 466 U. S. 668, the Sixth Circuit concluded that, while the Government conceded that Lee's counsel had performed deficiently, Lee could not show that he was prejudiced by his attorney's erroneous advice.*

Example 1 contains three sentences extracted from *Lee v. United States* [8], which demonstrate the impact of verbs and related semantics when it comes to legal information extraction. The verb *prejudice* in Sentence 1.1 plays a key role, because determining whether Lee was prejudiced by his counsel's advice can be considered as a key consideration in this case. Moreover, establishing whether Lee has *demonstrated* he was prejudiced or not can be considered as another legal test. Sentence 1.2 demonstrates how verbs can be useful when extracting arguments and other facts presented by a party at a case. Verbs such as *contend*, *argue*, *claim* can be considered as some of the keywords that would be useful when extracting such argumentative properties [9]. Sentence 1.3 demonstrates how verbs such as *applying* can be used to identify situations where existing laws or legal processes are being used to evaluate a scenario. A sophisticated mechanism to identify verbs conveying similar meanings would definitely benefit the latter two tasks. Furthermore, a recent study [10] demonstrates how verb similarity can be used to identify sentences that provide different opinions on a legal test related to the same subject within a legal opinion text. The same study claims that the lack of a sophisticated mechanism to identify verbs with similar meanings in legal opinion texts as the major limitation when it comes to detecting contradictory statements or arguments.

Evaluation resources and datasets are vital when it comes to evaluating the performance of model architectures, which are developed to identify semantic similarity between two textual units. *SimVerb-3500* [5] can be considered as a well-structured evaluation set of verb similarity, which has been developed to overcome some of the issues related to prior evaluation resources such as SimLex-999 [11]. The information available in SimVerb-3500 can be useful in developing mechanisms to identify verbs with similar meanings, as it has provided information on the verb pair type (eg: synonyms, hyponyms) in addition to the rating given to each verb pair. However, most of these existing evaluation resources, including *SimVerb-3500*, are focused on rating semantic similarity between two words, rather than explicitly rating whether two words in a word pair are having a similar meaning or not.

Additionally, in most of the current evaluation resources, the context has not been considered when the similarity between the verbs has been rated by human annotators. In other words, only the two verbs have been considered when determining their similarity. However, the same word may have multiple different meanings based on the context in which it is being used. For example, consider the sentences given in Example 2.

Example 2.

- *Sentence 2.1: Jon moved to United States.*
- *Sentence 2.2: Jon returned to South Korea.*
- *Sentence 2.3: Jon returned the balance to the customer.*

The verb *moved* in Sentence 2.1 and the verb *returned* in Sentence 2.2 suggest mobile behaviour. But, the verb *returned* in Sentence 2.3 suggests a behaviour of giving back. Therefore, while the two verbs in Sentence 2.1 ad 2.2 have a significant similarity when the

meanings are considered, we cannot observe that kind of similarity between the two verbs in Sentence 2.1 and 2.3. Thus, it is important to consider the associated context when rating two verbs based on their similarity.

Furthermore, unsupervised representation learning approaches such as XLNet [12], BERT [13], and ElMO [14], which are based on autoregressive (AR) or autoencoding (AE) pretraining objectives, have surpassed traditional word embedding approaches such as Word2Vec [15] and Sense2Vec [16] in several tasks related to natural language processing. Existing verb similarity evaluation resources, which do not take the context into account, would not be able to reap the maximum benefit when it comes to evaluating methodologies that can be used to measure verb similarity using these novel representational learning approaches.

The domain of the considered text may also be significant when it comes to the meaning of a verb. For example, in general use, the verb *plea* suggests a behaviour of requesting. However, in the legal domain, the verb *plea* often suggests a behaviour of stating guilt or innocence. Due to this kind of domain-specific behaviour, domain-independent evaluation resources that do not take context into account would have drawbacks when they are being used in the legal domain.

In order to overcome the abovementioned challenges in the existing resources, we developed a context-based verb similarity dataset *LeCoVe* that has been developed using legal opinion texts related to United States criminal cases. This dataset, which consists of 959 verb pairs, is publicly available (Lecove is publicly available at https://osf.io/bce9f/, accessed on 24 May 2022). The two sentences that are used to extract the two verbs were also included in the dataset. The verb pair scoring has been performed by human annotators, considering the respective sentence when interpreting the meaning of the considered verb. As the next step, we have used *LeCoVe* to evaluate popular word embedding models to quantify their performance on identifying verbs conveying similar meanings in the legal context. During the evaluations, we have developed Sense2Vec models that have been trained using a corpus of legal opinion texts. Furthermore, we have also proposed a methodology to identify verbs that convey similar meanings using BERT and evaluated that approach using *LeCoVe*. The Sense2Vec embedding models and further trained BERT models have been made publicly available. Finally, we developed a neural network to identify verbs with similar meanings by taking outputs from the existing word representational models as the input features. Our experiments demonstrate that the neural network model, which was developed combining the outputs of several models, outperforms the performances of models when they were used individually.

2. Related Work

2.1. Word Vector Embeddings and Language Modelling Systems

Converting a textual word into a numerical representation is a basic requirement in Natural Language Processing. The word representation approaches, which take semantic, syntactic and contextual properties of words, can be useful in identifying similarity between words. Neural Word Embedding approaches such as word2vec [15] and Glove [17], which use distributional similarity based representations, have been successfully used in various NLP tasks. However, most of these approaches model only one representation for a word, even though a word can have many meanings. Sense2vec [16], which is also a distributional similarity based word representation, attempts to overcome this issue by providing multiple embedding for a word, considering aspects such as Part of Speech tagging. One of the major limitations in these distributional similarity based techniques is that they do not consider the sequential context when learning the representations. As a result, these techniques would not be able to capture how the meaning of a word varies according to the sentence and the position of the word in the sentence. Models that are pre-trained based on different language modelling techniques such as ELMO [14], BERT [13], and XLNet [12] consider the sequential context, thus overcoming the above mentioned issues. Such language modelling techniques have become a key feature behind the state-of-the-art models related to many

NLP tasks. Therefore, datasets such as *LeCoVe*, which focus on considering the context when interpreting the similarity of two words, can be considered as a valuable contribution, as they can be used to evaluate different language modelling techniques and also to develop models that identify words with similar meanings using such techniques.

2.2. Resources on Verb Similarity

Semantic similarity measures that can be obtained from different word representation techniques can be made use to identify verbs with similar meanings. Classifying verb pairs as verbs with a similar meaning or not by defining a threshold based on semantic similarity can be considered as one such approach [10]. Evaluation resources are important when it comes to deciding on a suitable threshold to perform such a classification. In evaluation resources such as SimLex-999 [11] and *SimVerb-3500* [5], the way in which a meaning of a word can vary according to the context, has not been considered. Lack of contextual information would not only create issues in interpreting the meaning of a verb, but also it will create limitations when it comes to evaluating language representational models that are pre-trained on auto-regressive or autoencoding language modelling. A study by [18] provides a dataset that considers the context based on the sentences when rating the semantic similarity of two words, but that dataset contains only 399 verb-verb pairs. Moreover, these datasets [5,11,18] are focused on giving a rating based on semantic similarity instead of classifying word pairs as similar or dissimilar. As a result, when evaluating different distributional models using such datasets where similarity values have been annotated as ratings, correlation measurements, such as the Pearson correlation, have been used [11]. However, the objective of this study is to utilize distributional word embedding models as well as the language modelling techniques to differentiate verb pairs that have similar meanings in the legal domain from the verb pairs which do not have similar meanings (in the legal domain). A study [10] on identifying the shift of perspectives in the legal documents has used a verb pair dataset (the context of a verb has not been considered when annotating this dataset) that has been annotated based on whether two verbs are similar or not to evaluate existing semantic similarity measures available in WordNet [19] such as the Wu Palmer similarity [20]. For each of the selected semantic similarity measures, the precision, recall, and f-scores related to the task of identifying verbs with similar meanings have been calculated by varying a similarity threshold. The similarity threshold has been used to classify verb pairs into two classes: 1. verb pairs with similar meanings; 2. verb pairs with dissimilar meanings. We have adapted a similar approach for our evaluations. However, the semantic similarity measures that are in WordNet have been developed for generic English and have not been developed for the legal domain. Similarly, the datasets SimLex-999 [11] and *SimVerb-3500* [5] were not prepared specifically for the legal domain and they do not consider the context as well. Though the dataset provided in [18] considers the context, it has not been developed for the legal domain and consists of only 399 verb–verb pairs. As a result, there may be drawbacks when using those resources to analyse the behaviour of verbs in relation to different contexts in legal documents.

2.3. Semantic Similarity for Legal Information Extraction

Semantic similarity measures and datasets are crucial for any domain when developing and evaluating models and other mechanisms that are intended towards extracting information from the text. The importance of developing semantic types, annotated datasets and making them publicly available is also highlighted in a recent study on computational legal reasoning [1]. Another study that is also focused on the legal domain [21] demonstrates Word2Vec, and lexicon-based semantic similarity methods can be combined together to develop a more accurate domain-specific semantic similarity. Such semantic similarity measures have been utilized in the legal domain for: ontology population [22,23], document retrieval [24], and deriving representative vectors [25]. A recent study [10] empirically claims that a verb similarity based methodology outperforms the other two approaches, which are based on analysing sentiment and inconsistencies between triples, when it comes

to identifying sentences that provide different opinions on the same topic. The same study emphasizes the importance of developing a sophisticated verb similarity measure for the legal domain.

3. The Dataset

3.1. Design and Motivation

The verb pairs in the dataset are first classified considering whether the two verbs convey a similar meaning or not. After classifying a verb pair as similar or dissimilar, a rating from 0–10 will be given for the selected class (similar or dissimilar). When rating the verb pairs based on similarity, human annotators were instructed to obtain the meaning of each verb, considering the sentence that is related to that verb. After interpreting the meaning of the verbs using the context, only the meanings of the verbs will be considered when deciding the similarity. In other words, the related entities and other attributes in a sentence that are related to the considered verbs should not have an effect on the ratings given based on the similarity or dissimilarity of the verbs. If we consider Example 3 (which consists of sentences taken from *Lee v. United States* [8]), verb *moved* from Sentence 3.1 and verb *returned* from Sentence 3.2 suggest a behaviour of moving from one country to another. Therefore, these two verbs have a significant similarity in this context. As shown in the Example, the verb *move* in Sentence 3.1 is related to the United States while the verb *returned* in Sentence 3.2 is related to South Korea. However, that difference would not be considered when rating the two verbs based on their similarity, as the focus is given only to the meaning that is conveyed by each verb.

Example 3.

- *Sentence 3.1: Petitioner Jae Lee* ***moved*** *to the United States from South Korea with his parents when he was 13.*
- *Sentence 3.2: In the 35 years he has spent in this country, he has never* ***returned*** *to South Korea, nor has he become a U. S. citizen, living instead as a lawful permanent resident.*

3.2. Dataset Preparation

The criminal court cases were obtained from the SigmaLaw dataset. After randomly picking criminal court cases available in the corpus, the sentences in each document were split using Stanford CoreNLP [26] to create sentence pairs, from which the verb pairs were obtained. The sentence pairs were obtained from sentences that are adjacent, or only one sentence apart from each other in a legal opinion text. Nearby sentences were considered because it would be difficult to understand the context of the sentences that are far away from each other, even if it is from the same case. From the two sentences in a sentence pair, which are developed according to the above mentioned procedure, the sentence that appears first in a legal opinion text is known as the *target sentence*, and the other sentence is known as the *source sentence*. If we consider Example 3, Sentence 3.1 is the target sentence and Sentence 3.2 is the source sentence.

After constructing the sentence pair, the verbs from each of the sentences were extracted using Stanford CoreNLP PoS Tagger [27]. The verbs related to source sentence and target sentence were maintained separately in two lists, and the verbs that are lemmatized into words *be* or *have* were not included in the lists. Then, each of the verbs in the target sentence list was compared with each verb in the source sentence and the Wu-Palmer similarity score [20] between the two verbs were considered. A verb pair was added to the dataset only if the Wu-Palmer similarity score between the two verbs is greater than 0.75. This step was followed in order to maintain a proper balance between verb pairs with similar meanings and dissimilar meanings [10]. When a verb pair is selected to be included in the dataset, the details of the two verbs, as well as related sentences, will be added to the dataset. More specifically, these details include the *target sentence*, *source sentence*, *target verb*, *source verb*, *the lemmatized form of the target verb*, and *the lemmatized form of the source verb*.

3.3. Verb Pair Scoring

All the human annotators who participated in this process were provided with an understanding of the two classes for which the verb pairs will be classified (Similar/Dissimilar) using sets of examples related to each class. Then, other randomly selected examples were discussed with the inputs from human annotators, in order to make sure that all the human annotators have understood the class definitions and annotation guidelines properly. The human annotators were instructed to classify each verb pair either as a verb pair with a similar meaning or verb pair with dissimilar meaning by interpreting the meanings of the verbs, considering the context as described in Section 3.1. Human annotators were instructed to mark 1 for similar verb pairs and mark 0 for dissimilar verb pairs. After classifying a verb pair into one of these two classes, annotators were instructed to provide a rating from 1 to 10, based on how confident they are on their classifications. Table 1 illustrates some of the key statistics in *LeCoVe*, which have been identified using the ratings provided by human annotators.

Table 1. Frequency statistics.

Feature	Number of Verb Pairs
Two verbs with similar meaning (agreed by 3 human annotators)	170
Two verbs with similar meaning (agreed by atleast 2 human annotators)	285
Two verbs with similar meaning (agreed by atleast 1 human annotator)	463
Verb Pair with same lemmatized form, but different meaning (considering majority agreement)	6
Verb Pair with same lemmatized form and similar meaning	144
Number of unique verb pairs (lemmatized form)	714

4. Experiments and Evaluations

In this study, four human annotators were involved in the annotation process. As a result, the three human annotators who were involved in the rating process of one pair may be different individuals from those who were involved in that of another pair. Therefore, Fleiss' kappa [28] was chosen to measure inter-rater reliability. For the considered 959 pairs, we have observed the kappa value of 0.57 (belongs to the moderate agreement level as interpreted in a study on observer agreement on categorical data [29]). We observed that the kappa value is only shy of substantial agreement (0.61) by a narrow margin.

LeCoVe has been used to evaluate a number of popular Word Embedding Models and Language Modelling Techniques in regard to their capabilities on identifying verbs that have similar meanings. The objective of the experiments carried out in this study is to compare and contrast different word embedding model architectures and language modelling techniques, in order to get a quantitative understanding of their performances. Such evaluations will provide an idea on the areas where there is a potential for improvements in relation to determining verbs with similar meanings. Furthermore, the evaluations will provide insights on architectural design aspects, which will be helpful in developing novel and improved models that can determine verb similarity with substantial accuracy.

4.1. Resources for Evaluation

LeCoVe was used to evaluate several unsupervised representation learning approaches. The detailed description of the models that have been used are provided below.

4.1.1. Word2Vec Models

SigmaLaw dataset (SigmaLaw dataset can be found at https://osf.io/qvg8s/, accessed on 24 May 2022) [21] and contains three word2vec models that were pre-trained using a corpus of legal opinion text.

- Word2Vec (LR)—Trained using raw legal opinion text corpus.
- Word2Vec (LL)—Trained using lemmatized legal opinion text corpus.

- Word2Vec (LLR)—Trained using lemmatized legal opinion text corpus and then enhanced for lexical similarity.

In addition, we conducted our experiments with the publicly available word2vec model, which has been trained using the Google news corpus by Google (Word2Vec (G)).

4.1.2. Sense2Vec Models

One of the major drawbacks in Word2Vec is that it provides only one representation for a single word. As a result, in Word2Vec, both the noun form of a word, as well as the verb form, have the same vector representation. Therefore, Sense2Vec was also considered in our experiments, as it provides multiple vector representations for a single word, based on the word sense. As the first step of evaluating Sense2Vec models, the legal opinion text corpus available at SigmaLaw [21] was modified by lemmatizing each word and appending the POS tag behind each word (Spacy (https://spacy.io/, accessed on 24 May 2022) was used). Then, three Sense2Vec models were trained using the modified legal opinion text corpus (The Sense2Vec and BERT models developed in this study are available at https://osf.io/s8dj6/, accessed on 24 May 2022). The specific details related to the significant parameters that were used when training each of the Sense2Vec models are provided in Table 2.

Table 2. Parameter configurations of Sense2Vec models.

Parameter	SG-2	CBOW-10	SG-10
Model	Skip-gram	CBOW	Skip-gram
Size (dimensionality)	128	128	128
Min. Count	5	10	10
Context window size	5	10	10
Training algoriithm	Negative sampling	Hierarchical softmax	Negative sampling
Number of iterations	2	10	10

Furthermore, we have also evaluated Reddit Vectors 1.1.0 (Sense2Vec(R)), which is a publicly available Sense2Vec model that has been trained using a corpus obtained from Reddit.

4.1.3. BERT

BERT [13] has become a widely used language modelling technique for many NLP tasks including text classification, question answering, and natural language understanding. The key difference of BERT when compared with word embedding methods such as Word2Vec/Sense2Vec is that the representation that is related to a word, as provided by BERT, can vary on the context in which the word is being used. For the experiments in this study, we used the publicly available pre-trained BERT model ('bert-base-uncased') and other implementation mechanisms provided by Transformers (https://github.com/huggingface/transformers, accessed on 24 May 2022) library.

However, the publicly available pretrained BERT model was trained using a very large Wikipedia corpus and a book corpus. As a result, the performances of the pretrained BERT model in the legal domain has to be evaluated. In this regard, we have created a corpus using the Criminal Court Cases available at the SigmaLaw dataset in a way it can be used to train a BERT model following the instructions provided in BERT implementation (https://github.com/google-research/bert, accessed on 24 May 2022). In order to overcome the drawbacks that will be created when splitting the sentences, we add only sentences with at least five tokens to the training corpus. After suitably creating the text data (consists of 90,851 sentences) for pretraining using BERT, we have further trained the 'bert-base-uncased' model using the created text data taking parameters of the original model as the initial checkpoint. BERT is designed to learn two tasks during the training phase; masked LM (masked language modelling (predicting the tokens which are masked)) and next sentence prediction. While training the BERT using the legal corpus, we have observed the

performances of the model after one training step and after 500 training steps as shown in the Table 3. It can be observed that the accuracy of the pretrained BERT model in relation to the Criminal Case Corpus is low at the initial checkpoint. However, the accuracy increased significantly when the original model was further trained using the criminal case corpus. From this point onwards, pretrained model obtained from the BERT implementation by Google will be denoted using BERT(G), while the BERT model that was obtained after further training, using the criminal case corpus, will be denoted by BERT(L).

Table 3. Training BERT model using criminal court case corpus.

No. of Training Steps	Masked LM Accuracy	Masked LM Loss	Next Sentence Accuracy	Next Sentence Loss
1	0.55	2.71	0.60	2.47
500	0.70	1.42	0.95	0.14

4.2. Evaluating Distributional Word Representation Models Based on Cosine Similarity

When evaluating models that are based on Word2Vec/Sense2Vec, a threshold (t) based on the similarity score has been considered. First, the cosine similarity between the two vectors corresponding to two verbs in a verb pair is calculated. If $\mathbf{U}$ is the vector representation corresponding to the *Source Verb* and $\mathbf{V}$ is the vector representation corresponding to the *Target Verb*, cosine similarity between two vector representations are calculated as $\mathbf{U}^\intercal\mathbf{V}$. Then, the cosine similarity values of each pair, which are in the range between -1 to 1, were linearly scaled to a range of 0 to 1 (Scaled Value (sv) = $\frac{U^TV+1}{2}$). The values obtained after scaling was considered as the similarity scores of verb pairs. If the similarity score between two verbs, as given by the considered model, is higher than or equal to the threshold ($sv \geqslant t$), the verb pair is considered to be classified by the system as having two verbs that have a similar meaning. Otherwise, it will be considered as a verb pair that has been classified as having dissimilar meaning. It should be noted that the lemmatized form of the verbs has been considered when evaluating the Word2Vec/Sense2Vec models trained on lemmatized legal opinion text corpus. As each verb pair has been annotated by three human annotators, the class (Similar/Dissimilar) agreed by at least two human judge has been considered as the classification provided by the human annotators for a considered verb pair. Each model has been evaluated varying the similarity score that is being used as the threshold to classify the verb pairs based on similarity. As our main intention is on identifying verbs with similar meanings, the evaluation was performed in relation to the *Similar* class.

The results obtained from the evaluations are provided in Table 4. The way in which precision and recall were calculated in these evaluations can be described as follows. Let S be the number of verb pairs classified by the system as having verbs with similar meaning and let H be the number verb pairs classified as having verbs with similar meaning according to human annotations. Then, $Precision = \frac{S \wedge H}{S}$ and $Recall = \frac{S \wedge H}{H}$.

Table 4. Recall (R) and F-Measure (F) received from different thresholds of considered models.

Model \ Threshold	0.60		0.65		0.70		0.75		0.80		0.85		0.90	
	R	**F**	**R**	**F**	**R**	**F**	**R**	**F**	**R**	**F**	**R**	**F**	**R**	**F**
Word2Vec(G)	0.85	0.62	0.75	0.70	0.64	**0.71**	0.52	0.66	0.45	0.60	0.33	0.49	0.29	0.45
Word2Vec(LR)	0.80	0.65	0.74	0.70	0.68	**0.72**	0.60	0.70	0.53	0.66	0.41	0.57	0.33	0.49
Word2Vec(LL)	0.80	0.51	0.72	0.57	0.62	0.62	0.55	0.66	0.52	**0.67**	0.51	0.67	0.51	0.66
Word2Vec(LLR)	0.79	0.65	0.74	0.70	0.66	0.72	0.62	**0.73**	0.60	0.72	0.55	0.70	0.54	0.69
Sense2Vec(R)	1.00	0.46	1.00	0.47	0.98	0.50	0.86	0.54	0.71	0.62	0.58	0.64	0.52	**0.67**
Sense2Vec(SG-2)	0.92	0.55	0.83	0.58	0.78	0.64	0.70	0.67	0.65	0.71	0.62	**0.72**	0.56	0.70
Sense2Vec(CBOW-10)	0.94	0.53	0.89	0.63	0.81	0.69	0.68	0.69	0.64	0.72	0.61	0.72	0.59	**0.72**
Sense2Vec(SG-10)	0.82	0.62	0.75	0.66	0.71	0.69	0.65	**0.74**	0.61	0.73	0.56	0.70	0.55	0.69

4.3. Deriving Contextual Word Embeddings from BERT

Contextual word embeddings can be considered as an important language feature, which can be extracted using BERT to interpret the meaning and semantics of a word considering the context in which the word is being used. The key difference in deriving contextual word representations when compared to obtaining Word2Vec/Sense2Vec word representations is that in BERT, it is needed to input the associated sentence (context) of a verb in order to obtain the contextual representation for the particular verb. In Word2Vec/Sense2Vec, the vector representation for a word is static; however, in BERT, it varies according to the context. The 'bert-based uncased' model is made up of 12 hidden layers and within a single layer, each token in a sentence will be represented by 768 hidden units. Previous studies demonstrate that obtaining the contextual word embedding by averaging the vector representations for a particular word provided by the last four hidden layers has demonstrated promising results. Therefore, we followed the same methodology to obtain the contextual vectors. However, it was observed that contextual embeddings for some words cannot be obtained due to the way in which BERT performs tokenization. For example, the verb *substantiate* is tokenised into four sub-tokens *sub*, *##stan*,*##tia*,*##te*. In order to overcome this issue, we developed a methodology to identify sub-tokens of a particular word and then obtained the mean of the vector representations of sub-tokens as the contextual representation of the considered word (improved version). However, if a subtoken is the lemma of the considered verb, we directly take the contextual embedding of the subtoken as the embedding of the verb. After obtaining the contextual embeddings for two verbs in verb pairs, both BERT models (BERT(G) and BERT (L)) were evaluated considering the cosine similarity values following a similar approach, as described in Section 4.2. In this evaluation, we took cosine similarity values as they are, without performing any linear scaling. The obtained results are shown in Table 5.

Table 5. Recall (R) and F-Measure (F) received from different thresholds of BERT models.

Model \ Threshold	0.50		0.525		0.55		0.60		0.65		0.70		0.75	
	R	**F**	**R**	**F**	**R**	**F**	**R**	**F**	**R**	**F**	**R**	**F**	**R**	**F**
BERT(G)	0.80	0.65	0.77	0.69	0.72	0.68	0.65	**0.71**	0.57	0.68	0.45	0.60	0.34	0.50
BERT(G) Improved	0.85	0.66	0.81	0.70	0.75	0.69	0.67	**0.72**	0.59	0.69	0.46	0.61	0.38	0.54
BERT(L)	0.72	**0.71**	0.69	**0.71**	0.65	0.70	0.58	0.69	0.50	0.64	0.50	0.64	0.38	0.54
BERT(L) Improved	0.75	0.72	0.72	**0.73**	0.66	0.71	0.60	0.70	0.51	0.65	0.39	0.54	0.27	0.43

4.4. Result Analysis

When the evaluation results related to Word2Vec models and Sense2Vec models are considered (as shown in Table 4), one of the key common characteristics that can be observed is Word2Vec or Sense2Vec models that have been trained using a relatively small, but domain-specific corpus, which tends to outperform similar models that have been trained using a relatively large corpus, which is either domain generic or from another domain. If Figure 1 is considered, it can be observed that the two models Word2Vec (LLR) and Word2Vec(R), which were trained using the corpus of legal opinion text outperform Word2Vec (G), which was trained using the Google News Corpus. Similarly, if Figure 2 is considered, all three Sense2Vec models (SG-2, CBOW-10, SG-10), which have been trained using the legal opinion text corpus outperform the Sense2Vec model, which was trained from a corpus obtained from Reddit. As mentioned in the previous sections, the annotation process of *LeCoVe* has been carried out considering the meaning of a word relative to the legal domain. In such a context, these results suggest that Word Embedding models trained using a domain-specific corpus are able to encode the domain-specific meanings of words more comprehensively than domain-generic models. Therefore, it can be considered that domain-specific models and domain-specific datasets such as *LeCoVe* are of significant value, when it comes to legal information extraction.

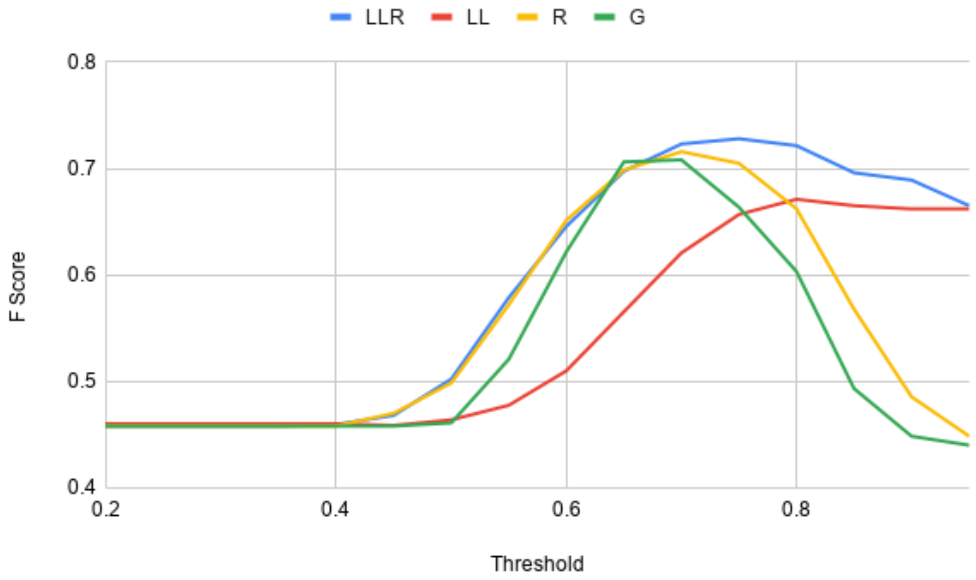

Figure 1. F measures obtained for different thresholds of similarity scores of Word2Vec models (Table 4).

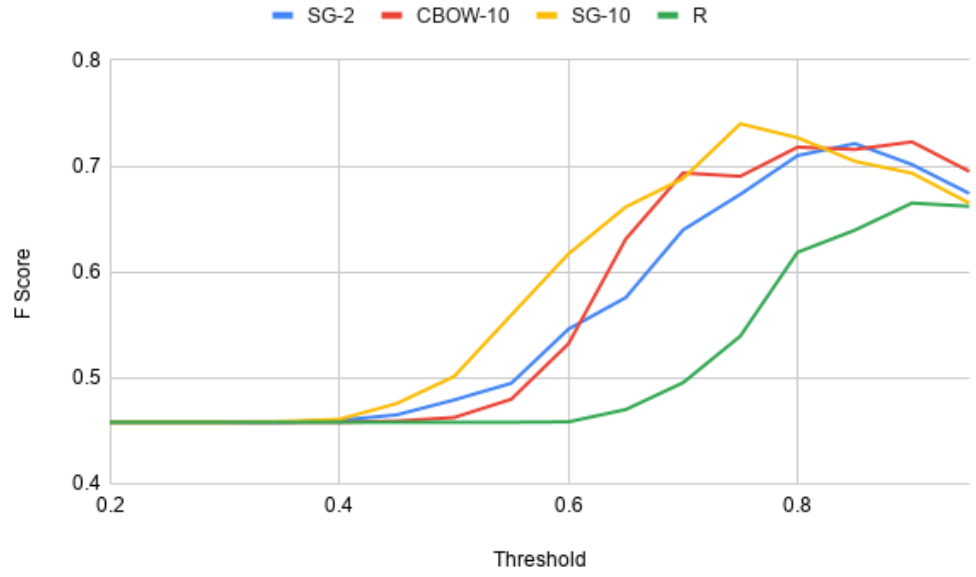

Figure 2. F measures obtained for different thresholds of similarity scores of Sense2Vec models (Table 4).

Proving the observations provided in the study related to the SigmaLaw dataset [21], Word2Vec(LLR) model outperforms all other Word2Vec models used in the evaluations, as shown in Figure 2. When it comes to Sense2Vec models, it can be observed that the overall performance of the Sense2Vec(SG-10) model is better than the other models used in the evaluations (Figure 2). This behaviour indicates that a Sense2Vec model, which is trained using the skip-gram algorithm, tends to perform better than a Sense2Vec model trained using the Continuous Bag Of Words (CBOW) algorithm with the same dimensionality and the same number of training iterations.

As illustrated in Figure 3, it can be considered that the performances of the Sense2Vec(SG-10) model is on a par with the Word2Vec(LLR) model. It should be noted that Word2Vec(LLR) has been enhanced for lexical similarity using lexical resources [21], while such enhancement has not been performed on the the Sense2Vec(SG-10) model developed in this study.

This demonstrates the potential of Sense2Vec models to overperform Word2Vec models when it comes to tasks such as identifying the similarity between words, which belongs to a particular word category (e.g., nouns, verbs, adjectives). It should be further noted that the vocabulary size of Sense2Vec models developed in this study are higher than Word2Vec models available in the SigmaLaw dataset, which have been trained using the lemmatized legal corpus. The reason is, in Sense2Vec, each sense of the word (eg: noun form and verb form) will create different vocabulary instances, while in Word2Vec, one word will have only one vocabulary instance. Though that is the case, the corpus used when training both ense2Vec models and Word2Vec models are the same, thus the number of tokens used in training is the same. This can also have a disadvantage on Sense2Vec models. Thus, if the size of the training corpus is increased, it can be assumed that the Sense2Vec(SG-10) model has the potential to perform better than the Word2Vec(LLR) model.

Another important behaviour that can be observed when analyzing the results obtained from experiments, is that the model that provides the highest F-Score varies depending on the considered threshold. As illustrated in Figure 3, Sense2Vec (CBOW-10) model shows the highest F-Scores for thresholds above 0.8, though that model is outperformed by Sense2Vec(SG-10) and Word2Vec(LLR) models for thresholds between 0.4 and 0.8. This behaviour indicates that there is a potential to improve the accuracy of detecting verbs with similar meaning by developing an ensemble model considering different Word2Vec and Sense2Vec models discussed here. It also provides signs that it is possible to incorporate the outcomes obtained from different models as features when developing novel machine learning or deep learning models to identify verbs with similar meaning.

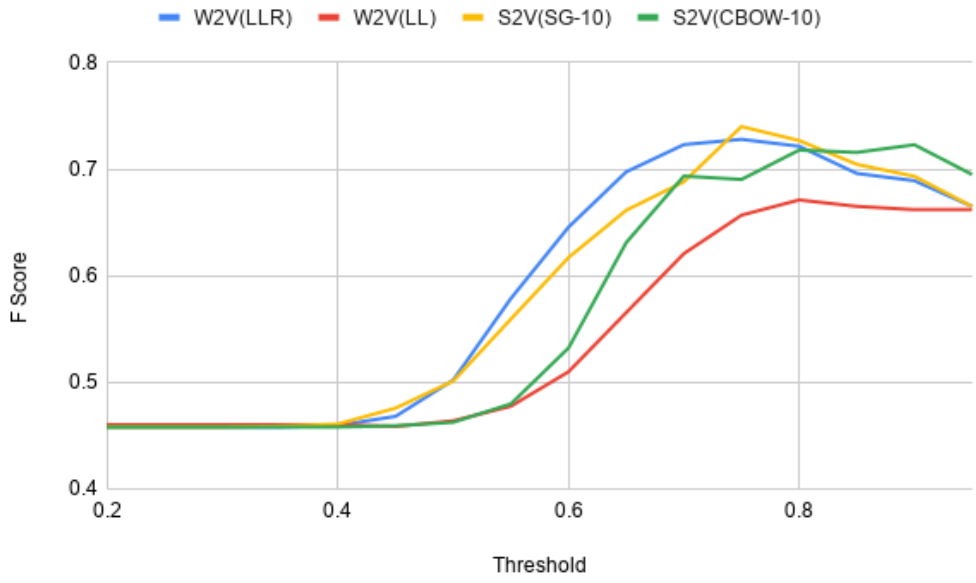

Figure 3. Comparison between Word2Vec and Sense2Vec models (Table 4).

When it comes to BERT-based approaches, it can be observed that the highest F-Scores of the four approaches lie in the range of 0.71–0.73, which is comparable with that of the Word2Vec and Sense2Vec based approaches. The results also demonstrate that the sub-token based improvements that were introduced in this study has increased the performance of both BERT(G) and BERT(L) methods. Moreover, it can be observed that the improved version BERT(L) model that was developed by post training BERT(G) model using legal domain specific corpus has outperformed all other BERT-based models when it comes to identifying the verbs with similar meanings in a legal context.

Overall, when interpreting the results given in Tables 4 and 5, it can be observed that the f-measures are below 0.75 in all the approaches that were considered. This emphasizes the importance of developing new models, which could identify verbs with a similar meaning in a comprehensive manner.

4.5. Context Sensitive Similar Verbs Classifier

The observations discussed in the previous section suggest that it would be of greater interest to analyze how we can combine these models together to develop a single model that can identify verbs with similar meanings. We have taken the cosine similarity values obtained for two verbs in a verb pair from the selected models as the input features. Word2Vec(LLR) and Sense2Vec(SG-10) were selected, as those two models have demonstrated the best results in previous experiments. However, in both models, only the lemma of the two verbs are considered. Therefore, we considered Word2Vec(R) as a measure to obtain legal domain specific word representations for the raw forms of verbs. Word2Vec(G) and Sense2Vec(R) models were also considered, as they were trained using a relatively large corpus (google news and reddit); thus, they have the potential to be more effective when capturing domain independent semantic relationships. Word2Vec provides a semantic representation for the words, while Sense2Vec adds syntactic properties to the representation, as the PoS tags are considered. However, none of these representations considered the word order in a sentence. Therefore, we have considered the outputs from the BERT(L) improved model to incorporate contextual properties.

The architecture of the Neural Network, which consists of fully connected layers, is shown in Figure 4. Two hidden layers and also the input layer have a bias unit, though it is not demonstrated in the diagram. Tanh is used as the activation function for hidden layers in order to maintain the activation outputs in [−1,1] range. As we are performing a binary classification task, a single neuron with the Sigmoid activation function is used as the output layer instead of a Softmax Layer with two neurons. The reason is that we need a probability in relation to the Similar class (Class with label 1), rather than a probability for each class. The Sigmoid activation function will provide the probability of two verbs conveying a similar meaning, thus it can also be used as a similarity measure between the verbs. Binary Cross Entropy (BCE) loss is used as the loss function of the model. $BCELoss = -\frac{1}{N}\sum_{n=1}^{N}(y_i log(p(s_i) + (1 - y_i) log(1 - p(s_i)))$, where y_i is the class label (1 if similar, 0 otherwise) and $p(s_i)$ is the predicted probability of the point, being *similar*(class 1) for each point i in the dataset, which consists of N data points. During the training, gradient descent is used as the optimizer with a learning rate of 10^{-4}.

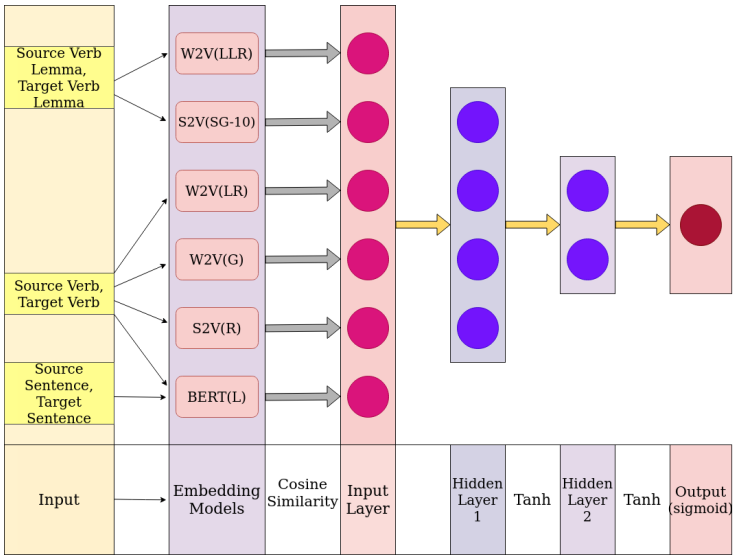

Figure 4. Architecture of the neural network model with fully connected hidden layers.

For the experiments, the data were randomly distributed as it is shown in Table 6. The Training Set, Test Set, and Validation Set are mutually exclusive from each other. One of the key challenges was the non even distribution of data in the dataset. Drawbacks are very much possible when training and evaluating a model with a significant portion of verb pairs with the same lemmatized form (with similar meanings). Therefore, out of 144

such pairs, we considered only 100 randomly selected pairs for the experiments. When the training was performed using a Training Set with data points that belong to 700 verb pairs, from which 545 verb pairs were from *dissimilar class*, the maximum F-Score that could be obtained after the evaluations using the Test Set was 0.79. Here, training was performed for 1500 epochs (considering the convergence of training and validation losses). As a measure of handling the class imbalance problem, 245 verb pairs were randomly chosen from the 545 verb pairs (that belong to the dissimilar class) to be used as data points in the Training Set. It was observed that training loss and validation loss started to converge after 1000 epochs (therefore the number of epochs for the training was set to 1000). After reducing the number of training points taken from the dissimilar class to 245, we observed slight improvements of the model, and the maximum F-Score obtained using the same Test Set was marginally increased to 0.80 (see Table 7).

Table 6. Distribution of data.

Data	Similar Meanings		Dissimilar Meanings	Total
	Different Lemmatized Form	Same Lemmatized Form		
Training	90	65	245	400
Validation	15	10	40	65
Test	36	25	89	150

Table 7. Comparison of Results.

Model	Threshold	Precision	Recall	F Score
Word2Vec Legal Raw (R)	0.35	0.78	0.70	0.74
Word2Vec Legal Lemmatized Enhanced(LLR)	0.15	0.64	0.92	0.75
Sense2Vec Legal Lemmatized (SG-10)	0.55	0.93	0.61	0.73
BERT(L) improved	0.50	0.79	0.69	0.74
Artificial Neural Network (Proposed in the study)	0.45	0.87	0.74	**0.80**

The results obtained from the evaluation of domain specific models (using the Test Set) are shown below. We considered the models that performed best in the experiments carried out in Section 4.2. Similar to the Section 4.2, the classification was conducted using a threshold value. Threshold values for Word2Vec/Sense2Vec/BERT are cosine similarity values without performing scaling. For the Artificial Neural Network (ANN) proposed in this study, the threshold was obtained from considering the output probability from the Sigmoid function. The maximum F-Score obtained (while maintaining a recall of at least 0.60) for each model is shown in Table 7 with the corresponding threshold values. It was also observed that maximum F-Score in relation to the test set (from all the considered models regardless of the recall or precision) is 0.80 and was obtained from the ANN proposed in this work.

5. Conclusions

This study discusses the importance of considering and analyzing the properties of verbs in legal documents and how the process of identifying verbs with similar meanings can be useful in different information extraction tasks such as sentiment analysis [30] and contradictory opinion detection [10]. Developing a verb similarity dataset in *LeCoVe*, which provides information related to the similarity of verbs based on the context it is being used and making it publicly available to the research community, can be considered as a key research contribution of this study. To the best of our knowledge, this is the first context-based verb similarity dataset developed for the legal domain. We have also used *LeCoVe* to evaluate the performances of different word representational models, considering the task of identifying verbs with similar meanings. The word representational models include existing models as well as Sense2Vec and BERT models, which have been

developed or improved within this study. One of the major drawbacks in most of the evaluation resources that have annotated the similarity between verbs, is that they do not consider the context associated with the verbs. In this study, we have addressed that issue by considering the sentence that a verb resides, when interpreting the meaning of that verb. As the sentence where a considered verb resides is also a part of the dataset, *LeCoVe* enables the use of approaches to identify verbs with similar meaning using unsupervised language representational techniques that consider the sequential context in a text. We have practically demonstrated this by using *LeCoVe* to evaluate pre-trained BERT models. Finally, considering the key observations of the evaluations conducted as part of validating the introduced data set, we propose a neural network-based approach. Using this neural network-based approach, we have demonstrated that the information that can be extracted from distributional word representational techniques and language modelling techniques can be combined together to develop more sophisticated and improved mechanisms to identify verbs with similar meanings. The findings of the study have the potential to facilitate several legal information extraction tasks, including contradiction detection, discourse analysis, and sentiment analysis.

Author Contributions: Conceptualization, G.R., N.d.S. and A.S.P.; methodology, G.R., N.d.S., A.S.P.; software, G.R.; validation, G.R., N.d.S., A.S.P., G.K., T.A. and A.W.; formal analysis, G.R., N.d.S. and A.S.P.; investigation, G.R., N.d.S. and A.S.P.; resources, G.R., N.d.S., A.S.P., G.K., T.A. and A.W.; data curation, G.R., G.K., T.A. and A.W.; writing—original draft preparation, G.R., N.d.S. and A.S.P.; writing—review and editing, G.K., T.A. and A.W.; supervision, A.S.P. and N.d.S.; funding acquisition, A.S.P. and N.d.S.. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by University of Moratuwa Senate Research Committee Fund for publications.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset which was used for the experiments within this study is publicly available at https://osf.io/bce9f, accessed on 24 May 2022.

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

References

1. Walker, V.R.; Han, J.H.; Ni, X.; Yoseda, K. Semantic types for computational legal reasoning: Propositional connectives and sentence roles in the veterans' claims dataset. In Proceedings of the 16th International Conference on Artificial Intelligence and Law, London, UK, 12–16 June 2017; pp. 217–226.
2. Ratnayaka, G.; Rupasinghe, T.; de Silva, N.; Warushavithana, M.; Gamage, V.; Perera, A.S. Identifying relationships among sentences in court case transcripts using discourse relations. In Proceedings of the 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 26–29 September 2018; pp. 13–20.
3. Jackendoff, R.S. *Semantic Interpretation in Generative Grammar*; The MIT Press: Cambridge, MA, USA, 1972.
4. Levin, B. *English Verb Classes and Alternations: A Preliminary Investigation*; University of Chicago Press: Chicago, IL, USA, 1993.
5. Gerz, D.; Vulić, I.; Hill, F.; Reichart, R.; Korhonen, A. Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv* **2016**, arXiv:1608.00869.
6. Ashley, K.D.; Walker, V.R. Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law, Rome Italy, 10–14 June 2013; pp. 176–180.
7. Gamage, V.; Warushavithana, M.; de Silva, N.; Perera, A.S.; Ratnayaka, G.; Rupasinghe, T. Fast Approach to Build an Automatic Sentiment Annotator for Legal Domain using Transfer Learning. *arXiv* **2018**, arXiv:1810.01912.
8. Lee v. United States, 582 U.S. 2017. Available online: https://supreme.justia.com/cases/federal/us/582/16-327/ (accessed on 2 April 2022).
9. Ratnayaka, G.; Rupasinghe, T.; de Silva, N.; Warushavithana, M.; Gamage, V.S.; Perera, M.; Perera, A.S. Classifying Sentences in Court Case Transcripts using Discourse and Argumentative Properties. *Int. J. Adv. ICT Emerg. Reg.* **2019**, *12*, 1–10. [CrossRef]
10. Ratnayaka, G.; Rupasinghe, T.; de Silva, N.; Gamage, V.S.; Warushavithana, M.; Perera, A.S. Shift-of-Perspective Identification within Legal Cases. *arXiv* **2019**, arXiv:1906.02430.
11. Hill, F.; Reichart, R.; Korhonen, A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Comput. Linguist.* **2015**, *41*, 665–695. [CrossRef]

12. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv* **2019**, arXiv:1906.08237.
13. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
14. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L.; Allen Institute for Artificial Intelligence. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
15. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
16. Trask, A.; Michalak, P.; Liu, J. sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. *arXiv* **2015**, arXiv:1511.06388.
17. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
18. Huang, E.H.; Socher, R.; Manning, C.D.; Ng, A.Y. Improving word representations via global context and multiple word prototypes. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Korea, 8–14 July 2012; pp. 873–882.
19. Pedersen, T.; Patwardhan, S.; Michelizzi, J. *WordNet:: Similarity: Measuring the Relatedness of Concepts*; NAACL-HLT; American Association for Artificial Intelligence: Menlo Park, CA, USA, 2004; pp. 38–41.
20. Wu, Z.; Palmer, M. *Verbs Semantics and Lexical Selection*; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 1994; pp. 133–138.
21. Sugathadasa, K.; Ayesha, B.; de Silva, N.; Perera, A.S.; Jayawardana, V.; Lakmal, D.; Perera, M. Synergistic union of word2vec and lexicon for domain specific semantic similarity. In Proceedings of the 2017 IEEE International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, 15–16 December 2017; pp. 1–6.
22. Jayawardana, V.; Lakmal, D.; de Silva, N.; Perera, A.S.; Sugathadasa, K.; Ayesha, B.; Perera, M. Semi-supervised instance population of an ontology using word vector embedding. In Proceedings of the 2017 Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 6–9 September 2017; pp. 1–7.
23. Jayawardana, V.; Lakmal, D.; de Silva, N.; Perera, A.S.; Sugathadasa, K.; Ayesha, B.; Perera, M. Word Vector Embeddings and Domain Specific Semantic based Semi-Supervised Ontology Instance Population. *Int. J. Adv. ICT Emerg. Reg.* **2017**, *10*, 1. [CrossRef]
24. Sugathadasa, K.; Ayesha, B.; Silva, N.D.; Perera, A.S.; Jayawardana, V.; Lakmal, D.; Perera, M. Legal Document Retrieval Using Document Vector Embeddings and Deep Learning. In Proceedings of the Science and Information Conference, London, UK, 10–12 July 2018; Springer: Cham, Switzerland, 2018; pp. 160–175.
25. Jayawardana, V.; Lakmal, D.; de Silva, N.; Perera, A.S.; Sugathadasa, K.; Ayesha, B. Deriving a representative vector for ontology classes with instance word vector embeddings. In Proceedings of the 2017 Seventh International Conference on Innovative Computing Technology (INTECH), Luton, UK, 16–18 August 2017; pp. 79–84.
26. Manning, C.D.; Surdeanu, M.; Bauer, J.; Finkel, J.R.; Bethard, S.; McClosky, D. The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MA, USA, 23–24 June 2014; pp. 55–60.
27. Toutanova, K.; Klein, D.; Manning, C.D.; Singer, Y. *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*; NAACL-HLT; Association for Computational Linguistics: Stroudsburg, PA, USA, 2003; pp. 173–180.
28. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **1971**, *76*, 378. [CrossRef]
29. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
30. Ratnayaka, G.; de Silva, N.; Perera, A.S.; Pathirana, R. Effective Approach to Develop a Sentiment Annotator For Legal Domain in a Low Resource Setting. *arXiv* **2020**, arXiv:2011.00318.