Journal of Theoretical and Applied Information Technology
30th April 2025. Vol.103. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCING CRIME DATA ANALYSIS THROUGH WORD-VECTOR CONVERSION: A RESNIK-GLOVE APPROACH

○ SAJNA MOL H S[1] , GLADSTON RAJ S[2]

[1]Research Scholar, Centre for Development of Imaging Technology (C-DIT), University of Kerala, Trivandrum, Kerala, India.

[2]Professor, Department of Computer Science, Govt. College Kariavattom, Trivandrum, Kerala, India.

E-mail: [1]sajnahsalam@gmail.com, [2]gladston@rediffmail.com

ABSTRACT

Crime data analysis is being performed now according to the requirement of Natural Language Processing (NLP) and machine learning algorithms that are able to extract insights from vast amounts of unstructured text data. There have been several efforts to apply classical word embedding models such as Glove, BERT, Skip-Gram, and Bag of Words, but they never identify the explicit meaning of crime tapes. This paper presents a novel hybrid method in which Resnik's similarity measure is combined with the GloVe algorithm to improve the semantic representation of text crime data. Our method employs the Resnik-GloVe model to generate word vectors for crime descriptions and addresses that both capture global word co-occurrence as well as semantic similarity. An Entropy swish based convolutional dense neural network (ES-CDNN) is learned to enhance classifier accuracy by incorporating them. The experiments performed with the 2016 San Francisco crime dataset validate that the Resnik-GloVe method actually generates an outstanding result by obtaining 98.33% accuracy whereas GloVe achieved 96.25%, BERT 94.26%, Skip-Gram 92.15%, and Bag of Words reached as low as 90.66%. The suggested methodology utilizes improved classification of crime data and helps law enforcement officers and policymakers analyze crime patterns more effectively. The research contributes to the area of crime analytics by addressing important shortcomings in existing word embedding methods and demonstrating the real benefits of integrating semantic similarity measures with conventional NLP models.

Keywords: *Crime Data Analysis, Word-To-Vector Conversion, Resnik Glove, Textual Information Representation, Word Embedding Algorithms.*

1. INTRODUCTION

Crime record analysis is a critical component of contemporary law enforcement, needing sophisticated methods to cull useful patterns from heterogeneous and frequently unstructured text data. Crime data, consisting of incident descriptions, locations, and times, is a rich source of information that can aid crime forecasting, intervention policies, and policy decisions. The data is challenging to analyze owing to semantic imprecision, language differences, and contextual dependencies of crime reports. This article emphasizes enhancing crime data analysis by utilizing a proficient word-to-vector transformation process with Resnik-GloVe technology.

A dataset of 2016 San Francisco crime statistics downloaded from GitHub, coupled with a crime surveillance video dataset, forms the basis of this investigation. Popular word embedding models like GloVe, BERT, Skip-Gram, and Bag-of-Words have been utilized extensively in NLP applications. These models fall short in usage when applied to crime analytics data. GloVe is good at capturing global co-occurrence statistics but poor in hierarchical semantic comprehension. Skip-Gram highlights local word relations but is poor at handling long-range dependencies. Bag-of-Words models are naive and do not understand contextual meaning, whereas BERT's high computational cost renders it infeasible for real-time crime analysis.

This study assumes that crime statistics in textual form accurately reflect patterns of crime in the external world. There are, however, a number of limitations including potential data set bias, differences across jurisdictions in the wording of crime reports, and processing limitations in large-scale deployments. Crime text classification performance depends greatly on text description consistency and completeness, which may vary by

jurisdiction and reporting system. Additionally, Resnik-GloVe enhances semantic understanding but is no better than hierarchical relationship quality in external knowledge bases.

1.1 Research Novelty

While GloVe, BERT, and Skip-Gram are normal word embedding approaches for general NLP applications, they do not function effectively when there is hierarchical semantic discontinuity in crime text. They are such as BERT, which functions flawlessly but requires a lot of computers and cannot be used in real-time analysis of crimes, and GloVe, which is based on co-occurrence of words and does not take into consideration the semantic relationships.

With the inclusion of Resnik's semantic similarity score on top of GloVe embeddings, this research offers a novel hybrid Resnik-GloVe model to improve the representation of crime texts. Resnik-GloVe preserves global word co-occurrence patterns as well as hierarchical semantic significance, improving class accuracy in criminal reports compared to traditional models. This approach is appropriate for use by law enforcement agencies as it allows them to better identify crime trends.

1.2 Problem Statement

Even though word embeddings for NLP have made improvements, current models are not efficient for semantic difference-based analysis of crimes and context limitation analysis. Word co-occurrence-based models like GloVe and Skip-Gram are vulnerable to word co-occurrence without addressing hierarchical crime terminology relationships [1]. BERT offers contextual embeddings but computational costs make it infeasible for real-time crime analysis [2].

Recent studies prefer the utilization of hybrid methods combining semantic similarity measures with word embeddings with the aim of enhancing classification performance [3][4]. Their utilization in crime text analytics, nonetheless, is minimal.

In response to these limitations, this paper proposes a hybrid Resnik-GloVe approach that improves semantic representation by leveraging hierarchical similarity metrics. This proposal fills the computational complexity gap for better crime text classification accuracy and is more adaptable for law enforcement use.

1.3 Research Objectives

- To improve the semantic representation of crime descriptions through the incorporation of Resnik's similarity alongside GloVe embeddings.
- To compare the performance of the Resnik-GloVe model in enhancing crime classification accuracy with the performance of current models.
- To create a scalable system for real-time crime pattern detection based on the new model.

We hypothesize that combining Resnik's semantic similarity measure with GloVe embeddings improves the contextual representation of crime descriptions, which results in higher classification accuracy than conventional word embedding methods.

In addition, we anticipate that the Resnik-GloVe model will perform better in crime text classification because it can identify hierarchical relationships between crime-related words and thus is more appropriate for law enforcement use and crime trend analysis.

The primary intention of using Resnik-GloVe in crime data analysis is to bridge semantic gaps in crime descriptions, especially in cases of multilingual and terrorism data. Resnik's information-content similarity measure also enhances the ability of GloVe embeddings to capture hierarchical word relationships. This improves the modeling of crime descriptions and addresses so that it can detect hidden patterns and trends in criminal activity. With greater semantic understanding, our proposed methodology enhances the overall goal of increasing the efficiency and precision of crime data processing.

The 2016 San Francisco crime data is a multi-faceted and rich real-world dataset posing traditional analytical methods. Crime reports frequently have nuanced language, colloquial terms, and inferential content needing advanced NLP methods to analyze properly. Resnik-GloVe technology with its ability to effectively capture semantic relationships is a vital addition to this analytical framework. By combining Resnik's similarity measure with GloVe embeddings, we resolve contextual inconsistencies in crime reports, thus enhancing crime classification accuracy and predictive analytics.

With advancing technology and increasing complexity of crime datasets, efficient methods of crime pattern discovery and analysis play a more significant role. Resnik-GloVe method advocated in this work is a state-of-the-art crime analytics that provides a highly efficient semantic-boosted word vectorization framework optimized for police use. Our results show that the combination of Resnik-GloVe with Entropy Swish-based Convolutional Dense Neural Network (ES-CDNN) has a 98.33% accuracy, which far exceeds other models like GloVe (96.25%), BERT (94.26%), Skip-Gram (92.15%), and Bag-of-Words (90.66%). This proves the use of semantic similarity measures in the workflows of crime data analysis to be effective.

This work not only supports current law enforcement needs but also lays the groundwork for developing smart systems to deal with the increasing complexity of crime information. Through the resolution of semantic problems in crime descriptions, this work supports the development of NLP techniques in criminal justice, which in turn helps in crime prevention, urban safety planning, and policy-making.

This work, therefore, introduces and utilizes a state-of-the-art word-to-vector conversion method through the use of Resnik-GloVe with a test case using the 2016 San Francisco crime data. This study is important to increasing the interpretability of crime information, narrowing the semantic chasms of criminal stories, and facilitating the advancement of crime analysis techniques.

2. DATASET AND IMPLEMENTATION

2.1 Dataset and Justification

The San Francisco crime dataset of 2016, taken from GitHub, is the main data set for this research. It consists of actual crime occurrences with formatted features like crime descriptions, geographic coordinates, timestamps, and addresses. The dataset is appropriate for word-to-vector conversions in crime analysis because it contains a wide range of crime categories, differing locations, and prominent temporal patterns.

Crime statistics analysis is problematic in nature due to unstructured text descriptions of crime, unstable vocabulary, and geospatial dependencies. As a solution to these problems, this research makes use of an organized text-analysis pipeline that refines the semantic interpretability of crime stories. Through the blending of Resnik's similarity metric with GloVe embeddings, the suggested approach advances contextual word features while maintaining hierarchical crime relationships.

2.2 Implementation Environment

The execution of this research is done with Python in the PyCharm IDE, taking advantage of its powerful debugging capabilities and native integration with machine learning libraries. Python is used extensively in data science because it is scalable, easy to develop, and has strong natural language processing (NLP) support. Some of the key libraries employed are NLTK for text processing, Scikit-learn for clustering, Gensim for word embeddings, and TensorFlow for deep learning based classification. The PyCharm environment supports efficient debugging, code organization, and modular execution, which makes it suitable for NLP-based crime analysis.

2.3 Architecture of the workflow

Prior to word-to-vector conversion, initial operations like geospatial clustering and distributed storage are conducted to organize the dataset in an efficient manner. Haversine K-Means clustering algorithm is utilized to cluster crime incidents based on spatial proximity so that location-based crime patterns are maintained. The clustered data is then mapped and stored with the Hadoop Distributed File System (HDFS) for maximizing storage and retrieval efficiency in handling large-scale crime records. These preprocessing procedures serve as the foundation for further application of Resnik-GloVe embeddings to better represent crime terms semantically. The workflow architecture and detailed methodology are discussed in the next section.

2.4 Research Method Protocol

For the sake of reproducibility, this research adopts a systematic research approach with the following steps:

- Data Collection: The 2016 San Francisco crime dataset is retrieved from GitHub with crime descriptions, timestamps, and locations.
- Preprocessing: Text cleaning, tokenization, removal of stopwords, lemmatization, and POS tagging are applied to crime reports using NLTK to normalize textual data.
- Embedding Generation: GloVe embeddings are calculated and enriched with Resnik's similarity measure to

enhance semantic representation of crime words.
- Clustering Approach: Haversine K-Means algorithm is used for clustering crime incidents based on geospatial distance to assist the identification of crime trends.
- Storage & Processing: The Hadoop Distributed File System (HDFS) is used for big data storage scalability and easy access of large-scale crime data.
- Model Training & Classification: The ES-CDNN classifier, which is built using TensorFlow, is trained on Resnik-GloVe embeddings to accurately classify crime types.
- Performance Evaluation: The model is evaluated against baseline embeddings (GloVe, BERT, Skip-Gram, BoW) on the basis of key performance measures such as accuracy, F1-score, and semantic consistency to verify improvements.

This systematic process ensures that every step from data collection to analysis is clearly defined to enable reproducibility and usability to other crime databases.

■ 3. METHODOLOGY

Conventional and static word representation methods treat words as atomic units represented as indices in a dictionary and these methods do not represent the similarity between words [5]. A static word embedding function converts each word into a vector and in comparison to vocabulary size, these vectors are dense and have a lower dimensionality [6]. Word-to-vector conversion using Resnik GloVe technology has been carefully designed to adapt the best natural language processing (NLP) methods to the unique nuances of text-to-action conversion. The complexity of the definition of violence poses a unique challenge that must be considered for visualizing semantic relationships. Word2vec is a vector generation method based on word embedding. Its core idea is the distributed expression of words, which maps words to vectors with definable dimensions [7].

3.1 Data Preprocessing and Clustering

3.1.1 Preprocessing pipeline

Prior to the implementation of word embedding algorithms, crime data is passed through a number of steps in order to make the text cleaner and consistent for reading:

- Text Cleaning: Elimination of special characters, numeric values, and white spaces.
- Tokenization: Segmentation of crime descriptions into individual words for vectorizing.
- Stopword Removal: Elimination of unwanted terms such as "the," "and," "is," etc.
- Lemmatization: Converting changer words into root words for better word-vector consistency.
- POS Tagging: Determination of word parts of speech (e.g., nouns, verbs) in an effort to more precisely determine word context in crime reports.

3.1.2 Geospatial clustering using Haversine K-Means

As crime information tends to have a geographic nature, this research uses Haversine K-Means clustering to cluster crime events according to geolocation proximity.

The Haversine distance formula is used for calculating distances between crime locations:

$$d = 2r \cdot arcsin\left(\sqrt{sin^2\left(\frac{\Delta\varphi}{2}\right) + cos(\varphi_1)cos(\varphi_2)sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \quad (1)$$

Where,

- r is the Earth's rasius (≈ 6,371 km),
- φ_1,φ_2 are the latitudes of two crime locations,
- $\Delta\varphi$ and $\Delta\lambda$ are the difference in latitude and longitude, respectively.

Through the use of Haversine K-Means, crime events are grouped into spatially proximate clusters, creating an organized representation for further semantic processing.

3.2 Data Storage and Mapping using HDFS

HDFS, a large-scale crime data management tool, is incorporated into the process to manage big crime data effectively. HDFS facilitates:

- Structured as well as unstructured crime data's efficient distributed storage.
- Large-scale NLP processes' parallel

processing.

- Efficient data retrieval optimized for real-time crime analysis.

The dataset of crime is mapped onto HDFS, supporting high availability as well as fault tolerance. Such a storage infrastructure is essential to handle word-vector data transformation at scale.

3.3 Word- to- Vector Conversion using Resnik-GloVe

3.3.1 GloVe embedding process

The GloVe (Global Vectors for Word Representation) model builds word embeddings from word co-occurrence statistics:

$$X_{ij} = \frac{P(w_i,w_j)}{P(w_i)P(w_j)} \quad (2)$$

Where,

- X_{ij} is the ratio of co-occurrence probability of words w_i and w_j,
- $P(w_i, w_j)$ is the probability of the words co-occurring,
- $P(w_i)$ and $P(w_j)$ are probabilities of individual words.

But GloVe is not sufficient on its own to capture hierarchical semantic relations and therefore needs to incorporate Resnik's measure of similarity.

3.3.2 Resnik similarity for enhanced word representation

Resnik's similarity measure is used in the GloVe algorithm. This measure is based on integration, which is the measure of information shared by two elements. For two given concepts, Resnik defines their similarity as the information contained in their most specific (i.e., at least among most concepts). GloVe is trained in global vocabulary integration. GloVe is well known for its efficiency and scalability and it has become more and more popular in NLP and provides rich, context-aware word representation [8].

Resnik's similarity measure is given by:

$$Sim_{Resnik}(c_1, c_2) = -log\ P(c) \quad (3)$$

Where, P(c) is the probability of finding the most informative common ancestor of concepts c_1 and c_2.

Through the union of Resnik's similarity and GloVe embeddings, crime descriptions are translated into richer semantic representations, resulting in better crime pattern detection and classification accuracy.

The training process of the San Francisco crime database is essentially a word-to-vector conversion method. Using the rich and diverse content of this data, Resnik GloVe technology reaches a level of education that will learn to associate words with their meanings. This phase is characterized by the optimization of word vectors to ensure that the output is evaluated as a relationship in violence-related information. The efficiency of this process, knowing the needs for inclusion of big crime data is the main goal.

The entire process, from pre-processing to parameter tuning and training, is designed to provide detailed and quantitative vector-to-vector conversion of the fault dataset. By tailoring Resnik GloVe technology to the unique characteristics of crime descriptions, this approach lays the foundation for the development of in-depth analysis in criminal record analysis.

4. RESULTS AND IMPLICATIONS

Resnik GloVe application on San Francisco crime data gave positive outcomes by improved representation of crime-related information. The power of a well-functioning word-to-vector mapping process extends past the mere utilization of the model. Improved representation of crime-related information allows improved analysis of crime information. The police would highly assist themselves if they knew more about crime trends. The ability to recognize social relations in accounts can help organizations improve decision-making, with consequences for police strategy and resource allocation.

Word- to- vector conversion is the major step in the crime prediction system for classifying the crime since the extracted description of address is represented in the form of the word. The classifier does not accept the word formation. Hence, the obtained extracted data is converted from word to vector using the Resnik-GloVe algorithm. In general, GloVe is an unsupervised learning algorithm for obtaining vector representation of words.

In the crime pattern prediction system, this is the major step for classifying the crime since the extracted description and address are represented in

the form of the word. The classifier does not accept the word formation. Therefore, the obtained extracted data (r_m) is converted from word to vector using the R-GloVe algorithm. Generally, an unsupervised learning algorithm to attain vector representations of words is termed GloVe. By factoring the word-word co-occurrence matrix, word representations are learned by the Resnik-GloVe. Minimizing the reconstruction error is the goal. The factorization $(\tilde{\Gamma}_m)$ for (r_m) is derived as,

$$\sum_{u=1}^{m}\sum_{v=1}^{m} f(r_{uv})(x_u y_v + b_u + b_v - \log(r_{uv}))^2 = \tilde{\Gamma}_m \quad (4)$$

Where, (b_u) and designate the scalar bias terms related to the words (u) and (v), respectively, (x_u) and (y_v) are trained vectors, a (b_v) nd $f(r_{uv})$ is the following weighted function that eliminates commonly occurring words (like stop words), which is shown as,

$$f(r_{uv}) = \begin{cases} \left(\frac{r}{r_{\max}}\right)^{0.75} & if\ r < r_{\max}, \\ 1 & otherwise, \end{cases}$$

Here, $(r_{\max})$ implies the maximum number of words.

The Resnik-GloVe model learns word representations by factoring the word-word co-occurrence matrix. Its ideal is to minimize the reconstruction error. When embedded in a convolutional dense neural network based on Entropy-Swish, the Resnik-GloVe algorithm can be used with other word embedding algorithms such as GloVe, BERT, Skip-Gram, and Bag-of -Words (BoW). It is compared with the real-time identification of the network (ES-CDNN) algorithm as shown in Table 1. Local and global data in the database. Its flexible structure and Entropy Swish functionality make it adaptable to a variety of environments and applications, providing a powerful force for deep learning-based analytics projects.

After adding Resnik-GloVe, the accuracy of the ES-CDNN algorithm reaches 98.33%. More importantly, the results demonstrate the utility of Resnik GloVe technology in improving crime detection capabilities. The model is capable of recognizing the complexity of criminal language and will directly enable law enforcement to better understand and provide an important tool to improve decision-making.

4.1 GloVe (Global Vectors for Word Representation):-

GloVe is a word embedding algorithm that learns the vector representation of a word based on global word frequency statistics. It captures relationships between words by identifying patterns that appear in large texts. The GloVe model is built on the intuition that the ratios of co-occurrence probabilities among words potentially encode some kind of a relation among words [9]. The GloVe performance is better than other word embeddings because it applies to non-zero elements and a subset of a corpus, not a whole corpus or a separate window of the significant corpus [10].

The basic premise of GloVe is that experts believe the co-occurrence of words within a particular context holds some contextual meaning. The GloVe model aims to decompose the co-occurrence matrix into two lower dimensional matrices that capture the word relationships.

Let the co-occurrence matrix be given by X. Each element X(i, j) tells us how often word i appears in the context of word j. The aim of GloVe is to find out two matrices W and C such that W*C comes closest to the representation of X that marks the co-occurrence matrix.

The GloVe objective function can be expressed as:

$$J = \sum_{i,j=1}^{V} f(X(i,j))(w_i c_j + b_i + b_j - log(X(i,j)))^2 \quad (5)$$

Where,

- X(i,j) represents the co-occurrence count of word i with word j.
- w_i and c_j denote the word and context word vectors, respectively.
- b_i and b_j are bias terms associated with the word and context word vectors.
- f(X(i,j)) is a weighting function applied to the co-occurrence count, typically in the form $f(x) = \left(\frac{x}{x_{max}}\right)^{\alpha}$, where α and x_{max} are constants that adjust the influence of rare

co-occurrence pairs.

In analyzing crime data, GloVe embeddings can capture semantic information by identifying crime and location, thus improving the efficiency and classification of the ES-CDNN algorithm. The accuracy of the ES-CDNN algorithm is 96.25% when GloVe is used for transformation.

4.2 BERT (Bidirectional Encoder Representations from Transformers):-

BERT embeddings are contextual and bidirectional, allowing them to capture complex syntactic and semantic relationships in sentences. These applications are therefore likely to achieve accuracy in criminal analysis, especially tasks that require understanding of context and speech. BERT is mainly used in a fine-tuning mode in most NLP tasks, and it is used as a feature-based mode and as an encoder for text representation [11]. The main purpose of BERT is to train bidirectional representation from an unlabeled dataset [12].

The BERT model is a bidirectional model. Unlike its predecessors, which were unidirectional and so read the text in a particular direction, the main model of BERT goes through the entire text in both directions simultaneously, which presents the property of "bidirectionality"[13].

BERT uses a transformer architecture that incorporates self-attention mechanisms to grasp the relationships between words in a sentence. It takes a bidirectional approach, meaning it looks at both the left and right context of each word simultaneously. Initially, the model is trained on a large dataset and then fine-tuned for specific tasks, such as classifying crime data.

The fundamental mechanism of the model is self-attention, and the attention score between two words is computed in the following way:

$$Attention(i,j) = \frac{exp(Q_i{}^{|}K_j)}{\sum_{k=1}^{n} exp(Q_i{}^{T}K_k)} \quad (6)$$

Where,

- Q_i is the query vector of word i.
- K_j is the key vector of word j.
- The softmax function ensures the attention scores sum to 1 across all words in the sequence.

While breaking the data analysis task in the ES-CDNN algorithm, BERT embeddings can achieve accuracy by capturing the relationship between the description of the crime and the location, thus improving the accuracy of crime prediction and classification. The accuracy of the ES-CDNN algorithm is 94.26% when BERT is used for transformation.

4.3 Skip-Gram:-

Skip-Gram embeddings are trained to predict the semantic content of a given target and can capture native language coordination. The Skip-Gram technique to word embedding has been extensively researched and developed in the NLP field, with various tweaks and enhancements presented over the years [14].

The Skip-Gram model operates by taking a target word and predicting the context words that surround it within a specified window. The goal is to maximize the probability of the context words based on the target word. In formal terms, the model is trained to optimize the following objective function:

$$J = \sum_{t=1}^{T} \sum_{-C \le j \le C, j \ne 0} log\ P(w_{t+j}|w_t) \quad (7)$$

Where,

- w_t is the target word at time t.
- w_{t+j} is the context word within the window of size C around w_t.
- $P(w_{t+j}|w_t)$ represents the probability of the context word w_{t+j} given the target word w_t, calculated as:

$$P(w_{t+j}|w_t) = \frac{exp(v'_{w_{t+j}}{}^{|}v_{w_t})}{\sum_{w=1}^{V} exp(v'_{w}{}^{T}v_{w_t})} \quad (8)$$

Where,

- v_{w_t} is the vector representation of the target word w_t.
- $v'_{w_{t+j}}$ is the vector representation of the context word w_{t+j}.

When the position in the sentence is closer to the position of the central concept, the concept vector corresponding to the central concept in the concept vector space, that is, the close relationship between the concepts in the sentence can be better reflected according to the relationship between the concept vectors [15].

When integrated into the ES-CDNN algorithm, Skip-Gram embeddings can use San Francisco crime report data to perform average-to-accurate crime detection tasks. When Skip-Gram is used for transformation, the accuracy of the ES-CDNN algorithm is 92.15%.

4.4 Bag-of-Words:-

Bag of Words (BoW) is a simple word embedding process that represents data as a collection of word frequencies. It does not capture the sentence or content but provides an example of semantic information. Bag-of-Words provides one way to deal with text representation and apply it to a standard type of text arrangement [16]. Traditional information retrieval approaches operate on a bag of words representing documents by their word distribution [17].

In a bag-of-words (BoW) model, a document is represented as a vector of a specific length, where each element indicates how often a word appears in that document. The vector is defined as follows:

$$D = [f_1, f_2, \ldots, f_V] \quad (9)$$

Where,

- f_i is the frequency of word i in the document D.
- V is the total number of unique words in the corpus.

Bag-of-words embeddings represent data as a collection of word frequencies and can provide an initial level of exposure to data breaches in the ES-CDNN algorithm. The accuracy of the ES-CDNN algorithm is 90.66% when BoW is used for transformation.

Comparative word embedding method analysis revealed the following as the main findings:

The comparison table and graph of these word-embedding algorithms are shown in Table 1 and Figure 1 respectively..

- Plus: Resnik-GloVe provides the highest accuracy (98.33%) because of improved semantic integration, where there are hierarchically related words that relate to crime. While simple GloVe examines word co-occurrence, Resnik-GloVe improves contextual information through the consideration of semantic similarity and hence improving the classification performance of crime reports.
- Minus: Additional cost of computation compared to isolated GloVe as it is doing the computation of extra semantic similarity. GloVe can be the efficient time solution for computation, Resnik-GloVe involves additional computation and hence is time-consuming on the big dataset for crime.
- Interesting: Although BERT offers us rich contextual information, it requires humongous computational resources and is not feasible for real-time crime analysis. Resnik-GloVe balances on the tightrope between explicit semantic precision and computational expense, and hence is a more realistic alternative to crime pattern identification for law enforcement agencies which need accuracy as much as operational efficiency.

- *Table 1:Comparison of word embedding algorithms.*

Word Embedding Algorithms	Accuracy of ES-CDNN (%)
Resnik-GloVe	98.33
GloVe	96.25%
BERT	94.26%
Skip-Gram	92.15%
Bag-of-Words(BoW)	90.66%

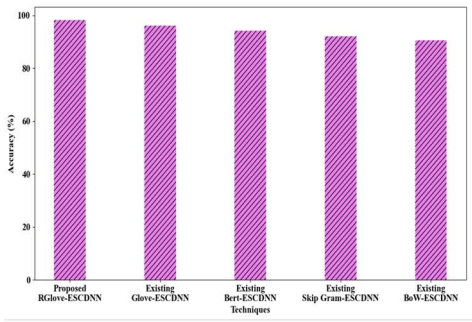

Figure 1: Comparison Graph of Word Embedding Algorithms

5. CONCLUSION

This article suggests a new Resnik-GloVe embedding approach that significantly improves semantic crime text analysis. The new model integrates hierarchical semantic relations to achieve superior accuracy and contextual representation compared to conventional embeddings.

The novelty in this work is:

- Enhanced crime text analysis accuracy through integration of Resnik-GloVe, improved identification of crime-related term associations.
- Scalable model for classifying crime data, thus suitable for large crime reports.
- Real-world application for law enforcement, improving crime trend detection and predictive modeling.

The findings show that this approach can be used to help law enforcement with better understanding of crime patterns and optimizing crime prevention. Future work will include implementing this framework on multilingual crime data and optimizing it for real-time application in mass-scale crime prediction systems.

REFERENCES:

[1] Qinhua Huang, Weimin Ouyang, "Word Embedding by Unlinking Head and Tail Entities in Crime Classification Model", *Proceedings of the 16th International Conference on Intelligent Computing (ICIC 2020)*, 2-5 October 2020, Bari, Italy. p. 555-564.

[2] Alina Arseniev-Koehler, Susan D. Cochran, Vickie M. Mays, Kai-Wei Chang, Jacob G. Foster, "Integrating Topic Modeling and Word Embedding to Characterize Violent Deaths", *Proceedings of the National Academy of Sciences*,Vol. 119, No.10, 2022, pp. 1-6.

[3] Jay Choi, David Kilmer, Michael Mueller-Smith, Sema A Taheri, "Hierarchical approaches to Text-based Offense Classification", *Science Advances*,Vol.9, No.9, 2023, pp. 1-15.

[4] Yongmin Yoo, Tak-Sung Heo, Yeongjoon Park, Kyungsun Kim, "A Novel Hybrid Methodology of Measuring Sentence Similarity", *Symmetry*,Vol. 13, No. 8, 2021, pp. 1-10.

[5] Deepak Suresh Asudani, Naresh Kumar Nagwani, Pradeep Singh, "Impact of word embedding models on text analytics in deep learning environment: a review", *Artificial Intelligence Review*, Vol.56, No.9, 2023, pp. 10345-10425.

[6] S. Joshua Johnson, M. Ramakrishna Murty, I. Navakanth, "A detailed review on word embedding techniques with emphasis on word2vec", *Multimedia Tools and Applications*, Vol.83, No.13, 2024, pp. 37979-38007.

[7] Xi Yang, Kaiwen Yang, Tianxu Cui, Min Chen, Liyan He, "A Study of Text Vectorization Method Combining Topic Model and Transfer Learning", *Processes*, Vol.10, No.2, 2022, pp.1-16.

[8] Shilpi Kulshretha, Lokesh Lodha, "Performance Evaluation of Word Embedding Algorithms", *International Journal of Innovative Science and Research Technology*, Vol.8, No.12, 2023, pp.1555-1561.

[9] Martin Canaan Mafunda, Maria Schuld, Kevin Durrheim, Sindisiwe Mazibuko, "A Word Embedding Trained on South African News Data", *The African Journal of Information and Communication (AJIC)*, Vol.1, No.30, 2022, pp.1-24.

[10] Shapol M Mohammed, Karwan Jacksi, Subhi R. M. Zeebaree, "A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol.22, No.1, 2021, pp.552-562.

[11] Qicai Wang, Peiyu Liu, Zhenfang Zhu, Hongxia Yin, Qiuyue Zhang, Lindong Zhang, "A Text Abstraction Summary Model based on BERT Word Embedding and Reinforcement Learning", *Applied Sciences*, Vol.9, No.21, 2019, pp.1-19.

[12] Rukhma Qasim, Waqas Haider Bangyal, Mohammed A. Alqarni, Abdulwahab Ali Almazroi, "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification", *Hindawi Journal of Healthcare Engineering[online]*, Vol.1, 2022, pp.1-17.[Accessed 31 August 2024]. Available from: https://doi.org/10.1155/2022/3498123 .

[13] Mayara Khadhraoui, Hatem Bellaaj, Mehdi Ben Ammar, Habib Hamam and Mohamed Jmaiel, "Survey of BERT-Base Models for Scientific Text Classification: COVID-19 Case Study", *Applied Sciences*, Vol.12, No.6, 2022, pp.1-19.

[14] Dr. Abhigyan Dwivedi, Sanjay Kumar Anand, "Word Embedding using Skip-Gram Approach", *Interdisciplinary Journal of Contemporary Research*, Vol.10, No.3, 2023, pp. 1-5.

[15] Yachun Tang, "Research on Word Vector Training Method Based on Improved Skip-Gram Algorithm", *Hindawi Advances in Multimedia[online]*, Vol. 1, 2022, pp.1-8. [Accessed 31 August 2024]. Available from: https://doi.org/10.1155/2022/4414207 .

[16] Nisha V M, Dr. Ashok Kumar R, "Implementation on Text Classification using Bag-of-Words Model", *Proceedings of the Second International Conference on Emerging Trends in Science and Technologies for Engineering Systems (ICETSE-2019)*, 17-18 May 2019, Karnataka. SSRN-Elsevier Digital Library, 2019, pp.1-8.

[17] David Rau, Mostafa Dehghani, Jaap Kamps, "Revisiting Bag of Words Document Representations for Efficient Ranking with Transformers", *ACM Transactions on Information Systems*, Vol.42, No.5, 2024, pp.1-27.