Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020)
IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9

Review on Word2Vec Word Embedding Neural Net

Soubraylu Sivakumar,
Computer Science Engineering,
Koneru Lakshmaiah Education Foundation,
Guntur, India.
sivas.postbox@gmail.com

Lakshmi Sarvani Videla,
Computer Science Engineering,
Koneru Lakshmaiah Education Foundation,
Guntur, India.
sarvani.mtech@kluniversity.in

Rajesh Kumar T.,
Computer Science Engineering,
Koneru Lakshmaiah Education Foundation,
Guntur, India.
t.rajesh61074@gmail.com

Nagaraj J.,
Computer Science Engineering,
Koneru Lakshmaiah Education Foundation,
Guntur, India.
nagrajan31@gmail.com

Shilpa Itnal,
Computer Science Engineering,
Koneru Lakshmaiah Education Foundation,
Guntur, India.
shilpaitnal@kluniversity.in

D. Haritha,
Computer Science Engineering,
Koneru Lakshmaiah Education Foundation,
Guntur, India.
haritha_donavalli@kluniversity.in

***Abstract*— The word2vec model consists of more useful applications in different NLP tasks. The semantic meaning given by word2vec for each word in vector representations has served useful task in machine learning text classification. They are employed in finding analogy, syntactic, and semantic analysis of words. Word2vec falls in two flavors CBOW and Skip-Gram. Given a context, they used to predict a word and vice versa are also true. In order to optimize the efficiency of word2vec, they have introduced two computational techniques namely hierarchical softmax and negative sampling. The proposed research work is more focused on introducing the models, computational technique, and various fields of word2vec applications. Word2vec is compared based on the metrics and their performance is evaluated by comparing with other existing models.**

Keywords— Skip Gram (SG), continuous bag of words (CBOW), hierarchical softmax (HS), negative sampling (NS), binary tree (BT)

I. INTRODUCTION

Word embedding is a process of converting a word into a number. Any machine learning (ML) or deep learning (DL) algorithm will only require a continuous vector of input to process their training process. They cannot handle strings with plain text for their natural language processing (NLP) model [27]. Each word in the vocabulary is mapped to a real number vector by using word embedding process. Traditionally, feature extraction technique like BOW and TF-IDF creates a vector size for the document, which is equal to the vocabulary size. They don't apply any dimensional reduction method to convert the sparse into a dense vector representation. The vector representation created by the word embedding has two advantages over the conventional methods (i) an efficient dense representation of the word in the vector space and (ii) contextual related standardized words have closer vector value.

The BOW approach creates a sparse matrix that consumes more memory for a huge corpus. It doesn't handle the contextual information. An n-gram based BOW technique, doesn't maintain the relationship between the terms beyond n-gram words. TF-IDF has a similar Pros and Cons of the BOW model, but they are computationally expensive than the BOW model when the corpus is a huge amount. This Word2Vec embedding model retains the semantics of words in a sentence or document and the contextual integrity of the sentence is not lost. The size of the embedding matrix is very small it maintains information related to the aspect of the term. It is faster than any other technique because of its simpler in design. This model scales well and gives good results for the smaller and larger dataset. This approach is used in the text classification task along with Convolution Neural Network (CNN) [23] and Recurrent Neural Network (RNN).

In conventional language processing, the document classification is done with TF-IDF score. They provided proportional importance of words in the document without handling semantic meaning. Word2Vec is a Neural Network (NN) model that encodes the semantic information of each term in the corpus given an unlabeled training data. It evaluates the cosine resemblance between the word vectors to understand the semantic similarity. Similar meaningful words have similar vectors, while dissimilar words have diversified vector. They have been used in various supervised language processing tasks such as Sentiment Classification, NER, POS-tagging, and document analysis. It falls in two flavors, Skip Gram (SG) and Continuous Bag of Words (CBOW). Back propagation and stochastic gradient descent method serves to learn the word vector. Both the models have a one concealed layer.

The word2vec architecture and its approaches are dealt with in section 2. In section3, the literature review and the applications of Word2vec in various fields are given. Finally, section 4 of the importance of Word2vec in various fields is concluded.

II. ARCHITECTURE

A. Skip Gram

The input to the SG is a single word W_i and the output is a context of words ($W_{0,1}$, … ,$W_{0,N}$). The 'W' is the weight

matrix. It lies among the input and concealed layer. It is a 'V' x 'N' matrix. 'V' stands for no. of words in vocabulary and 'N' Stands for no. of nodes in the concealed layer and it is the training parameter. H_i is the input to the concealed layer and it is the weighted sum of the input vector given in (1). The k^{th} row of 'W' will be output of the concealed layer. Fig. 1. Architecture diagram of Skip Gram model [22].

$$h = X^T W = V_{WI} \quad (1)$$

The input to the j^{th} node of concealed layer of the c^{th} output word is given in equation (2).

$$U_{c,j} = V'^T_{wj} . h \quad (2)$$

The output of the j^{th} hidden node of the c^{th} output word is computed finally using softmax function which is a multinomial distribution is given in (3).

$$p\left(W_{c,j} = W_{0,c} \mid W_I\right) = Y_{c,j} = \frac{\exp\left(U_{c,j}\right)}{\sum_{j'=1}^{V} \exp\left(U'_j\right)} \quad (3)$$

The loss function 'E' (4) will be

$$E = -\sum_{C=1}^{C} U_{j^* c} + C.log \sum_{j'=1}^{V} \exp\left(U_j\right) \quad (4)$$

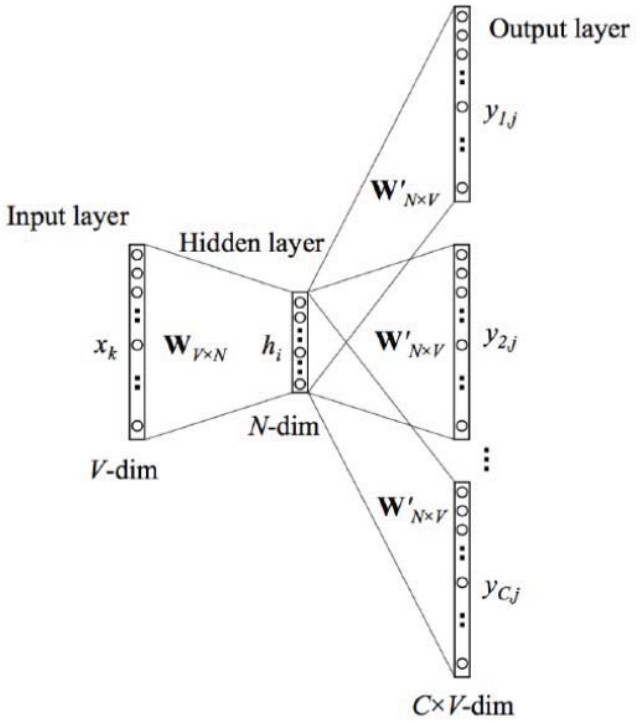

Fig. 1. Architecture of Skip Gram Model.

It is the probability of the output words given the input word. The index of the C^{th} output word is j*C. The variable $t_{c,j}$ is equal to 1, if the j^{th} node of the C^{th} output word is 1, otherwise it is 0. It is called as prediction error for the node c, j. The output matrix 'W' update equation (5) in gradient descent form is

$$W'^{(new)}_{ij} = W'^{(old)}_{ij} - \eta \sum_{c=1}^{C} \left(y_{c,j} - t_{c,j}\right).h_i \quad (5)$$

The input matrixes 'W' update equation (6) in gradient descent form is

$$W^{(new)}_{ij} = W^{(old)}_{ij} - \eta \sum_{j=1}^{V} \sum_{c=1}^{C} \left(y_{c,j} - t_{c,j}\right).w'_{i,j} \mathbf{x}_j \quad (6)$$

B. Continuous Bag Of Words

The input to the CBOW is a context of words ($W_{0,1}$,…,$W_{0;N}$) and the output is a single word Y. The 'C' is the window size. The weight matrix 'W' lies between the input and concealed layer. It is a 'V' x 'N' matrix. 'V' stands for no. of words in vocabulary and 'N' stands for no. of nodes in the concealed layer and it is the training parameter. Fig. 2 depicts the architecture diagram of CBOW model.

The 'h' is an N-dimensional vector in concealed layer. The 'N' x 'V' is a weight matrix that connects the concealed layer with output. The concealed layer 'h' (7) is computed by below equation.

$$h = \frac{1}{C} \mathbf{W} . \left(\sum_{i=1}^{c} x_i\right) \quad (7)$$

The matrix 'W' is weighted average of input vectors. In the output layer, the input to each hidden node is computed by equation (8).

$$u_j = V'^T_{wj} . h \quad (8)$$

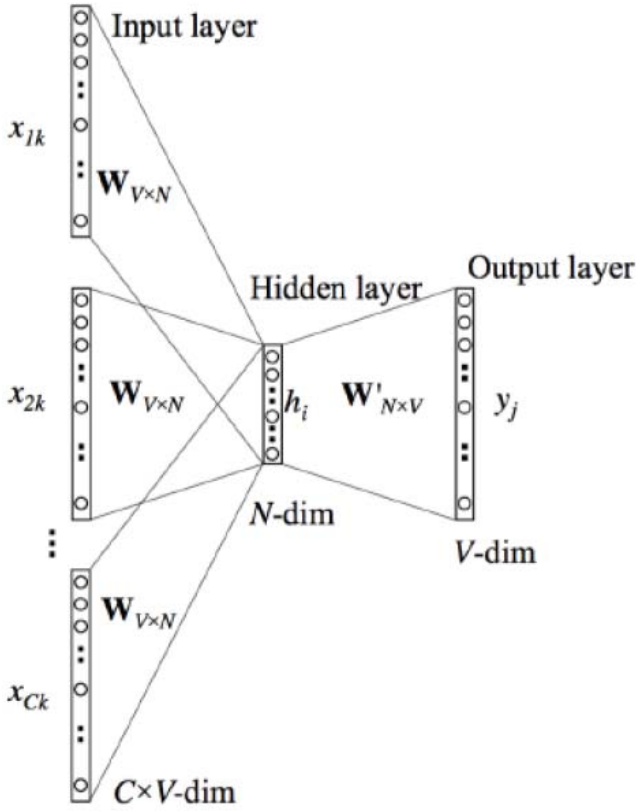

Fig. 2. Architecture of COWs model.

The j^{th} column of the 'W' output matrix is V'_{wj}. The output U_j is passed through the soft-max function to obtain Y_j is given in equation (9).

$$Y_j = p\left(W_j \mid W_1,\ldots,W_c\right) = \frac{\exp\left(U_j\right)}{\sum_{j'=1}^{V}\exp\left(U'_j\right)} \quad (9)$$

The loss function 'E' (10) will be

$$E = -V_{wo}^{T}.h + log\sum_{j=1}^{V}\exp\left(V_{wj'}^{T}.h\right) \quad (10)$$

The conditional probability has to be maximized, given the input context to obtain the output word. The output matrix 'W' update equation (11) in gradient descent form is

$$V_{w_j}^{'(new)} = \mathbf{V}_{w_j}^{'(old)} - \eta.\left(y_j - \mathbf{t}_j\right).\mathbf{h}_i \quad (11)$$

Where 'η' is the learning rate and it is greater than 0. The input matrix 'W' update equation (12) in gradient descent form is

$$V_{W_{I,C}}^{'(new)} = V_{\_W_{I,C}}^{'(old)} - \eta\frac{1}{C}.EH \quad (12)$$

Where $W_{I,c}$ is the C^{th} word in the input context and EH is a vector of n-dimensional elements $\sum_{j=1}^{V}\left(y_j - t_j\right).w'_{ij}$ from i = 1,…,n.

Each update gradient descent equation requires the summation of the entire vocabulary 'V', which is expensive with respect to computation. NS and HS are the efficient computation techniques used in practice. The conceptual representation of Word2Vec is shown in the figure 3. In CBOWs, given the surrounding word the architecture will find the main word "jumps". The skip gram finds the surrounding words given the main word "jumps".

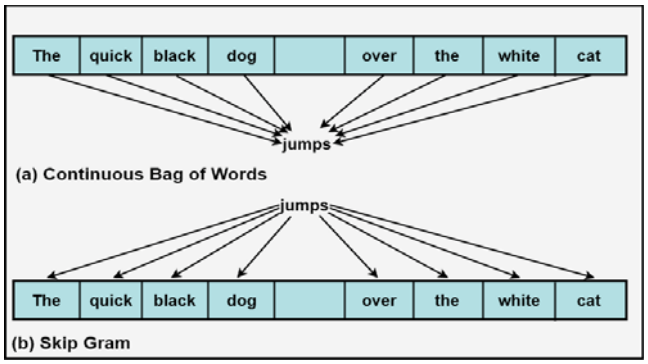

Fig. 3. Conceptual representation of Word2Vec architecture (a) CBOWs and (b) SG.

C. *Hierarchical Softmax*

In the NN model, the HS evaluates only the $log_2(W)$ nodes instead of measuring "W" output node to obtain the probability distribution. In HS the output layer is represented by a binary tree (BT) with the "W" words as leaves. The relative probability of the child nodes is explicitly represented by each node. A random walk through the BT assigns probability to words. Each term W in the BT can be reached out by a path from the BT root. Let n(W, i) be the i^{th} node along the path from the root to word W and let length of this path be L(W). For any node other than leaf, let ch(n) be an fixed child of "n" in the BT and let "x" be 1 if "x" is BT, otherwise -1. The $p(W_O|W_I)$ (13) for HS is defined as follows:

$$p(W \mid W_I) = \prod_{j=1}^{L(w)-1}\sigma(\| n(w,j+1) = ch(n(w,j)) \| v'^{T}_{n(w,j)} v_{wI}) \quad (13)$$

where x = 1/(1+exp(-x)). It is verified that $\sum_{w=1}^{W} p\left(w \mid w_I\right) = 1$. It entails that the computing cost of log $p(W_O|W_I)$ and log $p(W_O|W_I)$ is relative to $L(W_0)$, which is on average is no greater than log W. The HS has one interpretation V_w for each word 'W' and every inner node 'n' of the BT has one representation V'_η, while standard softmax has two interpretations V_w and V'_η to each word W. Thomas

Mikolov [1] et al. have given short codes for the frequent words in a binary Huffman tree. It provides a fast training.

D. Negative Sampling

Noise Contrastive Estimation (NCE) acts as an alternative to Negative Sampling. By means of logistic regression, NCE shows that a good model can differentiate data from noise an effective manner. SG mainly focuses on high quality vector representation; the log probability of the softmax is maximized by NCE. As long as SG retains their quality in vector representations, it is easy to simplify NCE. By objective function (14), NS is defined as

$$\log \sigma\left(v'_{wO}{}^{T} v_{wI}\right) + \sum_{i=1}^{k} E_{wi} \sim P_n(w) \left[\log \sigma\left(-v'_{wO}{}^{T} v_{wI}\right)\right] \quad (14)$$

Logistic regression are used to distinguish the destined word W_0 from the noise distribution $P_n(w)$. For each data sample [21], there are k negative samples. For large datasets the 'k' value will be 2-5, while for small datasets the 'k' takes the value 5-20. NCE considers both noise distribution and samples of the numerical probabilities, while NS uses only samples. The noise distribution $P_n(w)$ parameter is a free for both NCE and NS. The unigram distribution U(w) raised to the power three by fourth outperformed well is found by investigating a no. of choices for $P_n(w)$.

III. LITERATURE REVIEW AND ITS APPLICATION

Tomas Mikolov [1] et al. introduced a Skip-Gram model. During training, a significant speedup is obtained in the subsampling of frequent words. Less uncommon frequent words have improved accuracy. A negative sampling algorithm is put in place to provide an accurate representation for frequent words. An alternative to HS named negative softmax is introduced. Combining the word vector by simple vector addition gives meaningful results. A phrase is simply represented with a single token. For example, *vec(India) + vec(Capital)* is close to *vec(Delhi)* and *vec(India) + vec(river)* is close to *vec(ganga)*. Based on the cosine distance, they employed to find analogies. For example, *vec(tokyo)- vec(japan) + vec(india)* gives a *vec(delhi)*. They provide the syntactic analysis *(rain : rainy:: cloud : Cloudy)* and the semantic analysis *(country to capital city relationship)*. Combining the simple vector addition and token representation for the phrases helps to represent the longer pieces of text with minimal computational complexity.

Ying Sha [2] et al. have provided a MUSIC (Modeling User Style for Identifying aCcounts across social networks) framework to identify the user across social networks (SN) to evaluate the context style similarity. The framework is a step in two steps. During the first step, the framework defines the user account on a different social networks by the similarity in the content style. Second, the problem arises from multiple accounts on SN to a classification problem in a single account. They have used Word2Vec to obtain word vector, mean-pooling to obtain a document vector and finally, they performed classification. The dataset used is Twitter, Facebook, and Google+ from SN for classification. The collected user account is viewed as a positive instance. The negative instance is randomly chosen from the user account from each social media and they are not from the same individuals. The above framework achieves an F_1-Score of 89.3 for finding the user having multiple accounts or not.

Yan Zhang [3] et al. investigates the impact of 3 factors on S. A.: Sentiment polarity distribution, language models, and model settings. The key factor is addressed by the different data sampling techniques to find sentiment category distribution. The additional factor is done by combining different language models named hybrid approach. The final factor is addressed by separate model settings. The dataset employed in the method is posted from online diabetes Chinese forum *(https://bbs.tnbz.com/)*. The metrics used in the approach are accuracy and P-value. An individual approach, NBSVM [28] with unigram and bigram provides an accuracy of 83.73%. The hybrid approach, PV + NBSVM (unigram and bigram) provides an accuracy of 86.20%.

Giannis Nikolentzos [4] et al. have a proposed Multivariate Gaussian Document representation. They measure the similarity of the document based on the distribution similarity. They have used Reuters, Amazon, TREC, Snippets, BBC sport, Polarity, Subjectivity, and Twitter for their analysis on the state of the art approach. Some of the baselines methods used in predictions are Bag of Words, NBSVM, Centroid, Word Mover Distance, and CNN [19]. The average representations of the words are the mean of each distribution. The variation in dimensions of the mean w. r. t. other is measured using the covariance matrix. Cosine similarity finds the similarity between the mean vectors. Their empirical evaluation outperforms in various state of art methods. The Word Movers Distance computes the distance between the documents. Word2Vec is used to find the distance and K-Nearest Neighbor is employed to classify the documents. The accuracy and F_1-score of the proposed Gaussian are high for the dataset Reuters, Amazon, TREC, BBCSport, and Twitter. The Naïve Bayes-Support Vector Machine has high accuracy for Polarity and Subjectivity.

Liqiang Niu [5] et al. have proposed a unified framework model. Most of the distributed models focus on neighborhood context properties and it learns individually task-specific representation. The proposed model focuses on multi-attributes and learns jointly. They consider three core attributes: document, topic and lemma. In this framework, they have implemented Skip-Gram and CBOWs. ***TW:*** Learning Topic Representations, given the current topic Z_t and word W_t, Skip-Gram predicts (W_{t-2}, W_{t-1}, W_{t+1}, W_{t+2}) surroundings words and CBOWs uses (W_{t-2}, W_{t-1}, W_{t+1}, W_{t+2}) surrounding words, the word W_t and topic Z_t is predicted. ***DW:*** Learning Document Representations, the Skip-Gram gives the W_t current word and D_t document to predict surrounding words (W_{t-2}, W_{t-1}, W_{t+1}, W_{t+2}) and CBOW gives

the (W_{t-2}, W_{t-1}, W_{t+1}, W_{t+2}) nearby words, predicts the W_t and D_t. ***LW:*** In morphology, a set of words canonical form is the lemma. In English, for example, "come", "comes", "arrive" and "coming" are forms of the same lexeme, with come as the lemma. ***TLW:*** Consider both topic and lemma attributes try to improve the word representation. The CBOW be given with the lemma L_t and topic Z_t, predicts the W_t based on the surrounding words (W_{t-2}, W_{t-1}, W_{t+1}, W_{t+2}). Given the word W_t, topic Z_t and lemma L_t, the surrounding context word is predicted by Skip-Gram. The dataset used in the experiment are DS-LOOK for topic evaluation, 20 NewsGroup for the evaluation of document, and Mikolov e tal [15] proposed to google dataset and Mikolov et al [16] proposed MSR dataset for the evaluation of improved word representation. For evaluation of topic representation are done using LDA and TW. The evaluation of document representation is done using LDA, PV-DM and PV-DBOW. The improved word representation is evaluated with Word2Vec, TW, LW, TLW and GloVe. In all the three different types of evaluation, the proposed methods TW, LW, DW and TLW outperform the baseline and start of art methods.

Zhe Zhao [6] et al. provided weighting schemes for raw BOW in which important terms are given more weight while compared to unimportant. These term weight guides the neural model (Word2Vec and Glove) to point on important words. The PV-DBOW, PV-DM and PV-GLOVE are used for weighting the text. The Paragraph Vector is implemented with two variants Hierarchical Softmax (HS) and Negative Sampling (NS). The dataset employed for weighting and classification is IMDb dataset. Naive Bayes, Pointwise Mutual Information, Average Likelihood, Odds Ratio, and Weighted Nave Bayes are used with the baseline as the weighting schemes. Among the possible variants of PV, the PV-DM with NS performs the best. The weighted NB produces a robust by producing the accuracies above the average case.

Pengda Qin [7] et al. introduced a retraining scheme by a sentence classification to adjust unsupervised word embedding in certain supervised tasks. The same word being present in different tasks represents different sentiment polarities. The word '*duck*' has different meaning in two different sentences.

Sent 1: The duck swims well in the pond: +ve polarities

Sent 2: Sourav Gangully was out for a duck, while playing a match against New Zealand: -ve polarities

The proposed work concentrated on fine-grained objectives to provide a task specific word embedding. For sentence classification using CNN [26], the back-propagation procedure is supported in word embedding to fine-tune on the labeled dataset. This approach is based on by PV-DM and implemented based on Word2Vec. The two methods are Task Specific-SG and Task Specific-CBOW. The PV-DM injects the word and the label a task specific data into word embedding. In all circumstances from the same sentences, a definite unique label vector is shared. The characteristics of corresponding classes are remembered using the label vector as a memory unit. Because of mutual prediction between words, Out Of Training Set (OOTS) words and Out Of Vocabulary (OOV) words are embedded into the task specific words with label information. Four datasets employed in our model are MR, SST-1, SST-2, and TREC. A pre-trained embedding is employed in the model are GoogleNews-Vectors-negative300.bin from Mikolovs Word2Vec and glove.840B.300d.Zip from Penningtons Glove method. The proposed work not only handles OOTS and OOV words, but the method is completely overcome overfitting problem because of the independence of task training procedure.

Felix Hill [8] et al. have introduced two unsupervised representation learning objectives. Domain portability, Training time, and performance is optimized to trade-off in their design. Sequential Denoising Autoencoders (SDAE) and FastSent a sentence level log-bilinear bag-of-words model. FastSent is a variant SkipThought objective. SkipThought is drawn to predict target sentences S_{j-1} and S_{j+1} given source sentence S_j. The feed-forward net applied to fixed size data is Denoising Auto Encoder. By introducing a noise function $N(S/p_0, p_x)$, it can be used for variable-length sentences by free parameters p_0, p_x ε [0, 1]. First, for a wordWin sentence S, noise function N deletes with probability p_0. Second, for the non-overlapping W_iW_{i+1} bigram in sentence S with probability p_x, then N swaps W_iW_{i+1}. Then train the LSTM based encoder-decoder architecture as NMT, given a corrupted version $N(S/p_0, p_x)$ the denoising objective predicts the original sentence S. This mode is called SDAE. SkipThought uses the adjacent sentence context to obtain the precious sentence semantics. FastSent exploit the invariable signal, but at a lower computational expense. A simple sentence in a context is given as BOW representation, the output adjacent sentences are also predicted in BOW form. For each word in the model vocabulary, the FastSent learns a target v_w and source u_w embedding. Toronto Books corpus was trained for supervised evaluation in six sentence classification tasks. The SICK and STS2014 datasets were trained for unsupervised evaluation. These two models performed well on specific tasks MSRP and SICK sentence relatedness.

Denis Gordeev [9] suggested an automatic detection of verbal aggression in Internet Communities. The U. S. government does not define any criteria for predicting the aggressive message. There is no registration in the Internet forum for prediction of the message. The task of detecting human emotions involves both sentimental analysis and aggressive analysis. Word2Vec finds the semantic relations and distance between words without any preprocessing or any annotation. The message applied in the training is from *4chnn.org* and 2ch.ck with 654,047 and 1,148,692 messages respectively. In a message, the important features for the English are the average semantic difference and the difference between the minimum and maximum distance within any pair in a wordset. The detected aggression ratio of English and Russian language is 88% and 59% respectively. The classifier pays more attention only to individual rude words while the entire message is not aggressive. Word2Vec model finds similarities between words, while random forest judges the message is aggressive or not. Some other information like parts of speech and other grammar characteristics like imperative verbs may be helpful in detecting the aggression.

Akihiro Toyoshima [10] et al. compared the Concept-

Base (CB) model and word distributed model for word association. The computer and humans engage in dialogue with each other in natural language. Association frequency table is utilized to find the word association. This assists in building the concept-base using the chain set method. Knowledge is known as a concept base. For example; I am going for a walk in the afternoon, the associate words are umbrella, water, and handkerchief. Chain-set based concept model extracts attributes from the headwords. Furthermore, it extracts the second level attribute based on the first order attributes as a headword. This extraction process continues to N order attributes. The association frequency table is a compilation of a headword and associative words as a set. They have used precision, recall, and F-measure. They utilized compound data EDR Electronic Dictionary and Wikipedia data as training data. Electronic Dictionary headwords are registered with a user dictionary of MeCas to analyze the Concept Dictionary and Japanese word dictionary. Wikipedia headword logs into with a user dictionary of MeCab to examine Wikipedia. Skip-gram provides a higher evaluation result than CBOW in semantic-syntactic word relationship text. The evaluation is made on two CB, five word2vec models, and the baseline CB. The baseline concept base is the highest F-measure of all models. A baseline method manually removed inappropriate attributes from CB and manually appended the right attributes of the concept. The Second-CB has a greater recall of all models. The Second-CB included most attributes is correct associative words. The result shows constructing CB based on chain-set; extracts new associative words. CB is under a high degree of semantic similar words because they extracted superordinate and subordinate words.

Frenando Enriquez [11] et al. showed a compared analysis of BOWs, Word2Vec models, and their integrated models. They also adapted the classification of cross-domain classifications. Word2Vec is not a part of DL, but some of the parts of the training scheme are related. The auto encoder has two noteworthy feature [20] that makes it popular. First, there is no need to annotate the images or text i.e., it opens the door to unsupervised learning. Second, they do everything they learn. i.e., when they process the input generates various representations. Word2Vec is the same as auto-encoder. Skipgram is better in handling smaller corpus, while CBOW is having an increase in speed and higher quality of representation. Logistic regression is used as a base classifier. Eleven different domains have been identified as a dataset which is referred by McAuley et al. [17, 18]. The combination method averages the weight to label it as positive and negative. They also adapted the classifier to cross-domain to identify the adaptability of the classifier in a new domain. The classifier provided a better accuracy result 9 out of 11 domains except in camera and electronics. They assessed the content of two domains. The two domains have a large no. of infrequent words which help the BOW classifier to correctly predict the output class. All the infrequent terms are domain specific, so the BOW model gives a better result than Word2Vec model. Their terms are not required in the generic corpus that one was used to build the Word2Vec model. In cross domain evaluation, the combination models obtain percentage of 10 out of 11 domains.

Saurav Ghosh [12] et al. have given a vocabulary driven approach of Word2Vec for characterizing diseases from unstructured Health map news corpus. The vocabulary driven way is called Dis2Vec. It provides a word embedding to generate automated disease [24] taxonomy. These news articles just try to focus on all disease characteristics such as exposures, transmission methods, symptoms, and transmission agents. The domain knowledge is integrated with Word2Vec in the form of a predetermined list containing diseases related terms such as names, symptoms, and transmission methods. To find the transmission agent for the plague, the cosine similarities are calculated between the embedding's of plague word and all possible terms of transmission agents. Top words are extracted after sorting. The corpus article contains textual data, disease tag, location information in terms of latitude/longitude, and reported data. The articles are preprocessed by the RLP tool for sentence splitting, tokenization, and lemmatization. The terms being under a frequency less than 5 times are ignored. A total of 39 diseases are selected that represent a diversity of diseases [25] ranging from emerging (H7N9) to endemic (dengue) to rare (plague).

Thin Nguygen [13] et al. have estimated the degree of Adverse Drug Reactions (ADR) for psychiatric medications from social media. The quantified ADR from the SIDER database and the estimated ADR from social media corpora is measured using Pearson correlation coefficient. Word2Vec exceeds the coverage of variants of ADR terms of 0.08 and 0.50 (Baseline) to 0.29 and 0.59 (Word2Vec). The social media ADR rate is related to the known ADR rate is the primary aim. The second aim is tantamount to employ vector arithmetic to identify additional terms to improve the ADR rate. The datasets used are drawn from LiveJournal (33 million), Reddit (200 million posts, and 1.6 billion comments), and twitter (1.89 billion comments). Filtering is applied on the dataset using ten Psychiatric drugs (either brand or generic name) to obtain 602799 instances. Some noise is also captured which was irrelevant. The word closest to salivation was lacrimation in the Word2Vec corpus which is has no bearing. The use of extended lexicon helps to find the documents which contain alternatives mentioned. Eg: ringing ear alternatives tinnitus. The drowsiness is an ADR of Xanax, an adverse event X is detected for drug D by resolving the formula: Xanax + drowsiness $\approx$ D + X. In some cases, similar terms were worn for the indication of the drugs. In the proposed method, language characterizes are not regarded as negation.

Sl. No.	Proposed Approach	Dataset Used	Year	Achievements	Limitations
1.	Skip Gram Model	News articles from Google dataset and phrase analogy dataset	2013	Syntactic and Semantic tasks are simplified.	Not able to handle OOV words.
2.	Modeling User Style for Identifying aCcounts across social networks	Twitter, Facebook and Google+ from Social Network	2016	Identify the user across social networks to evaluate the context style similarity	It cannot be optimized for particular tasks.
3.	A hybrid approach of Paragraph Vector + Naïve Bayes and Support Vector Machine (unigram and bigram)	Online diabetes message collected from Chinese forum	2016	Sentiment polarity distribution, language models and model settings are resolved.	Sub-word level sharing is not represented in a proper way.
4.	Multivariate Gaussian Document	Reuters, Amazon, TREC, Snippets, BBC sport, Polarity, Subjectivity and Twitter	2017	Word2vec is used to find the distance between the words.	Not able to handle OOV words.
5.	Unified framework model based on Attributes	20 NewsGroup, Google and MSR dataset	2015	This model focuses on multi-attributes like and learns jointly. They consider three main attributes: document, topic and lemma.	They do not address the issue of conceptual words.
6.	Weighted Naïve Bayes	IMDB	2017	Important and relevant terms are given high weights.	It cannot be optimized for particular tasks.
7.	A retraining scheme by a sentence classification to adjust unsupervised word embedding in certain supervised tasks	MR, SST-1, SST-2 and TREC	2017	The Out Of Training Set (OOTS) and Out Of Vocabulary (OOV) words are handle in a better way.	Global co-occurrence statistics is not considered.
8.	Sequential Denoising Auto encoders (SDAE) and FastSent	SICK and STS2014 dataset	2016	Domain portability, Training time and performance are optimized to trade-off in their design.	Not able to handle OOV words.
9.	Word2Vec model find similarity between words and Random Forest is used for classification.	Aggression Messages from 4chnn.org and 2ch.ck	2015	The classifier pays more attention only to individual rude words while the whole message is not aggressive.	They do not address the issue of conceptual words.
10.	Chain-set based concept model extracts attribute from the headwords and further second level attributes are extracted from first order attributes.	EDR Electronic Dictionary and Wikipedia data	2016	Constructing Concept Base based on chain-set; extracts new associative words and Concept Base has high degree of semantic similar words.	Sub-word level sharing are not represented in a proper way.
11.	Word2vec with Logistic Regression	User-generated reviews from 12 different domains	2016	Cross domain evaluation obtained a better result in different domain datasets.	They do not address the issue of conceptual words.

12.	Vocabulary driven approach called Dis2Vec	HealthMap News Article	2016	A total of 39 diseases representing a diversity of diseases ranging from emerging (H7N9) to endemic (dengue) to rare (plague) are addressed.	Global co-occurrence statistics is not considered.

Table 1. Limitations of word2vec with existing model

Sl. No.	Proposed Approach	Dataset Used	Year	Achievements	Limitations
13.	The degree of Adverse Drug Reactions (ADR) for psychiatric medications from social media is estimated by word2vec.	Adverse Drug Reactions (ADR) from SIDER database	2017	It helps to find an adverse event X for drug D by resolving the formula: Xanax + drowsiness ≈ D + X.	Sub-word level sharing are not represented in a proper way.
14.	Majority voting method is applied to combine surface and deep features.	SemEval 2013, SemEval 2014, Vader, STS-Gold, PL04, Sentiment140 and IMDB	2017	M_{SG} is an ensemble of features that combine both surface and deep learning features.	Global co-occurrence statistics is not considered.

Table 1. Limitations of word2vec on existing model (continued…)

Oscar Araque [14] et al. provided better feature representation capabilities and performance better than conventional feature-based techniques. A new classifier is exercised by combining word embedding model and linear ML algorithm. The traditional approach is time-consuming because of manual feature engineering. Complex features are extracted automatically by the DL approach with a large amount of data to perform well. Two ensemble techniques are proposed which aggregate the baseline classifier with the other surface classifier. Two models are suggested to combine surface and deep features from various sources. Taxonomy is presented for classifying the models. Taxonomy included surface features(S), generic automatic word vectors (G), and affect word vectors (A). In other words, the combination can be no ensemble method at all, through an ensemble of classifiers or the advantage feature ensemble. These frameworks allow us to provide in choosing the most appropriate and efficient method for specific applications. The rigid rule model combines predictions from various classifiers using a simple voting rule i.e., by majority voting. This rule counts the predictions of major classifiers and assigns the input to the class with most component predictions. In the Meta classifier model, the outputs of the major classifiers are regarded as features for this model. They can adapt to diverse situations.

Random forest is supposed to provide a high performance metrics for sentiment analysis. M_{SG} is an ensemble of features that combine both surface and DL features. M_{GA} is completely extracted using DL techniques. M_{SGA} is the proposed model combine surface, generic, and affects word vector. The measure used in this work is the macro-averaged F_1 score. The dataset used are SemEval 2013, SemEval 2014, Vader, STS-Gold, PL04, Sentiment140, and IMDB datasets. Taxonomy acts as a framework to characterize the existing approach to the ensemble of traditional and deep techniques. The combination of information from diverse sources helps in improving the classifier results. A CEM_{SGA} and $M_{SG + bigrams}$ model have the best performing alternatives. They try to extend the proposed models to other languages and even paradigms like emotional analysis. The numerous limitations that exist in the various models are given in table 1.

IV. DISCUSSION

Global co-occurrence of two terms across the different reviews is not dealt with by the existing model. The word that is not part of the vocabulary of dataset is given less importance during vector generation. Sub words level sharing is not represented in a proper way. A word having different meaning in various contexts is not handled by the word2vec. Optimizing this model to a different domain and task is difficult. Few words in the vector space are not distributed uniformly, and they create an insufficient utilization of vector space.

V. CONCLUSION

Thus, the proposed research work has studied the concept of word2vec NN and its two models (CBOW and SG). A computational efficient methods used in word2vec is also discussed (HS and NS). It is used as a framework named as MUSIC to identify the illegal access across the SN. Word analogy can be discovered with the semantic meaning between words. It is uses to identify the verbal aggression in the internet community. Word2vec is employed in medical fields for estimating the Adverse Drug Reactions for psychiatric medications through social media. It is utilized as a Dis2Vec i.e., disease to vector to solve around 39 diseases using names, symptoms and transmission methods. In NLP, it is used to solve problems related to language translation between the humans and the computer with the help of chain-set based concept model.

Construction of co-occurrence matrix for the vocabulary needs more storage and time. In the sentiment analysis, the opposite word pair like “good” and “bad” is closely located in vector space. This model finds it difficult to handle this pair. A new embedding matrix has to be constructed when it is applied for a new language or different domain.

V. REFERENCES

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean, ”Distributed Representations of Words and Phrases and their Compositionality”, Neural Information Processing Systems conference, pp 1-9, 2013.

[2] Ying Sha, Qi Liang, and Kaijian Zheng, ”Matching User Accounts across Social Networks based on Users Message”, The International Conference on Computational Science (ICCS) 2016, Volume 80, pp. 24232427, 2016.

[3] Yan Zhang, Yong Zhang, Jennifer Xu, Chunxiao Xing, and Hsinchun Chen, “Sentiment Analysis on Chinese Health Forums A Preliminary Study of Different Language Models”, International Conference on Smart Health, pp 68-81, 2016.

[4] Giannis Nikolentzos, Polykarpos Meladianos, Francois Rousseau, Michalis Vazirgiannis and Yannis Stavrakas, ”Multivariate Gaussian Document Representation fromWord Embeddings for Text Categorization, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2, pp. 450455, 2017.

[5] Liqiang Niu, Xin-Yu Dai, Shujian Huang and Jiajun Chen, ”A Unified Framework for Jointly Learning Distributed Representations of Word and Attributes”, Proceedings of Machine Learning Research, Vol. 45, pp. 143-156, 2015.

[6] Zhao, Tao Liu, Shen Li, Bofang Li, and Xiaoyong Du, ”Guiding the Training of Distributed Text Representation with Supervised Weighting Scheme for Sentiment Analysis”, Data Science and Engineering, Volume 2, Issue 2, pp. 178186, 2017.

[7] Pengda Qin,Weiran Xu, and Jun Guo, ”A Targeted Retraining Scheme of Unsupervised Word Embeddings for Specific Supervised Tasks”, Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 3-14, 2017.

[8] Felix Hill, Kyunghyun Cho and Anna Korhonen, ”Learning Distributed Representations of Sentences from Unlabelled Data”, Proceedings of NAACL-HLT, pp. 13671377, 2016.

[9] Denis Gordeev, ”Automatic detection of verbal aggression for Russian and American imageboards”, International Conference on Communication in Multicultural Society, pp. 71-75, 2015.

[10] Akihiro Toyoshimaa and Noriyuki Okumurab, ”A Comparison of Concept-base Model and Word Distributed Model as Word Association System”, International Conference on Knowledge Based and Intelligent Information and Engineering Systems, pp. 385 394, 2016.

[11] Fernando Enriquez, Jose A. Troyano, Tomas Lopez-Solaz, ”An approach to the use of word embeddings in an opinion classification task”, Expert Systems With Applications, September 3, 2016.

[12] Saurav Ghosh, Prithwish Chakraborty, Emily Cohn, John S. Brownstein and Naren Ramakrishnan, ”Characterizing Diseases from Unstructured Text: A Vocabulary Driven Word2vec Approach”, 25th ACM International on Conference on Information and Knowledge Management, October 2016.

[13] Thin Nguyen, Mark E. Larsen, Bridianne ODea, Dinh Phung, SvethaVenkatesh and Helen Christensen, ”Estimation of the prevalence of adverse drug reactions from social media”, International Journal of Medical Informatics, pp. 130137, 2017.

[14] Oscar Araque, Ignacio Corcuera-Platas, J. Fernando Snchez-Rada and Carlos A. Iglesias, ”Enhancing deep learning sentiment analysis with ensemble techniques in social applications”, Expert Systems With Applications, pp. 236246, 2017.

[15] Tomas Mikolov, Kai Chen, Gerg Corrado and Jeffrey Dean, “Efficient estimation of word representations in vector space”, 2013, CoRR, abs/1301.3781.

[16] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig, “Linguistic regularities in continuous space word representations”, In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 746-751, Atlanta, Georgia, June. Association for Computational Linguistics, 2013.

[17] McAuley, J., Pandey, R., Leskovec, J., “Inferring networks of substitutable and complementary products”, In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794, 2015.

[18] McAuley, J., Targett, C., Shi, Q., van den Hengel, A., “Image-based recommendations on styles and substitutes”, In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 43–52, 2015.

[19] Sivakumar S., Rajalakshmi R., Prakash K.B., Kanna B.R. and Karthikeyan C., “Virtual Vision Architecture for VIP in Ubiquitous Computing. In: Paiva S. (eds) Technological Trends in Improved Mobility of the Visually Impaired”, EAI/Springer Innovations in Communication and Computing. Springer, Cham, pp 145-179, 2020. DOI: 10.1007/978-3-030-16450-8_7

[20] S. Sivakumar and R. Rajalakshmi, “Comparative Evaluation of various feature weighting methods on movie reviews”, Springer 2017 International Conference on Computational Intelligence in Data Mining (ICCIDM), pp. 721-730, Nov 2017. DOI: 10.1007/978-981-10-8055-5_64

[21] K. B. Apoorva Sindoori, L. Karthikeyan, S. Sivakumar, G. Abirami and Ramesh Babu Durai, “Multiservice Product Comparison System with Improved Reliability in Big Data Broadcasting”, IEEE 2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM), March 2017. DOI: 10.1109/ICONSTEM.2017.8261256

[22] Rajesh Kumar T., G. R. Suresh, and S.K. Raja. "Conversion of Non-Audible Murmur to Normal Speech Based on Full-Rank Gaussian Mixture Model." Journal of Computational and Theoretical Nanoscience 15.1 (2018), Pp:185-190.

[23] Rajesh Kumar T, Suresh GR, Kanaga Subaraja S, Karthikeyan C, ”Taylor-AMS features and deep Convolutional neural network for converting non audible murmur to normal speech”. Computational Intelligence.2020;1–24. ttps://doi.org/10.1111/coin.12281.

[24] R. Kasthuri, B. Nivetha, S. Shabana, M. Veluchamy and S. Sivakumar, “Smart Device for Visually Impaired People” in Jeppiaar Engineering College, Chennai – IEEE 2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM) on 23th - 24th March 2017. DOI: 10.1109/ICONSTEM.2017.8261257

[25] J. Jennifer, Monica Nathasha Marrison, J. Seetha, S. Sivakumar, P. Sathish Saravanan, “Dynamic Medical Machine for Rural Areas” in SSN College of Engineering, Chennai – IEEE 2017 International Conference On Power And Embedded Drive Control (ICPEDC) on 16th - 18th March 2017. DOI: 10.1109/ICPEDC.2017.8081135

[26] V. Soniya, R Swetha Sri, K Swetha Titty, R. Ramakrishnan, and S. Sivakumar, “Attendance Automation Using Face Recognition Biometric Authentication” in SSN College of Engineering, Chennai – IEEE 2017 International Conference On Power And Embedded Drive Control (ICPEDC) on 16th - 18th March 2017. DOI: 10.1109/ICPEDC.2017.8081072

[27] Joby, P. P., "Expedient Information Retrieval System for Web Pages Using the Natural Language Modeling", Journal of Artificial Intelligence, 2, no. 02, pp. 100-110, 2020.

[28] Raj, Jennifer S., and J. Vijitha Ananthi, "Recurrent neural networks and nonlinear prediction in support vector machines", Journal of Soft Computing Paradigm (JSCP), Vol. 1, No. 01, pp. 33-40, 2019.