Detection of Context-Dependent Lexical Information from Unstructured Data using Word Embeddings Based on Machine Learning: An Assessment

Amit Shukla and Rajendra Gupta
Rabindranath Tagore University, Bhopal, India

Keywords: Lexical Information, Unstructured Data, Word Embeddings.

Abstract: In the current generation of context-dependent information solutions, data discovery and identification engines rely on rule-based models in most cases, and they are confined to context-independent lexical data in unstructured data such as formless text or images. Data elements that are always considered lexical, regardless of the context in which they reside, are known as context-independent lexical data. The unstructured lexical data is the data that isn't arranged according to a pre-determined data schema and can't be saved in traditional relational database. Text and multimedia are two types of unstructured data that are regularly analyzed. The paper presents a Context - Centered Extraction of Concepts (CCEC) word embeddings method the gives benefit from a neural-network methods ability to encode textual information by converting meaningful text information into numeric values. The result shows about 90 per cent accuracy in targeting context-dependent lexical information by considering the context of the words in a sentence/text.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

The term "context dependent" refers to a type of word representation that enables machine learning algorithms to distinguish words that have similar meanings. It is a feature learning technique that uses probabilistic models, dimension reduction, or neural networks on the word co-occurrence vector matrix to map words into real-number vectors (Kopeykina et.al., 2021). Consider the phrase 'Tiger,' which is context independent, but is context dependent when we say, 'The Tiger is harmful' or 'The Tiger may be dangerous.'

A typical context-dependent system consists of several tools and components that are divided into two categories: host-based and network-based systems. A tool that analyses network traffic and communication platforms is network-based context dependent information (e.g., email and instant messaging).

The state of the art in Data Loss Prevention (DLP) systems; automatic lexical data detection algorithms can be divided into two types. The first category includes rule-based approaches, which are widely used in commercial DLP software. The second

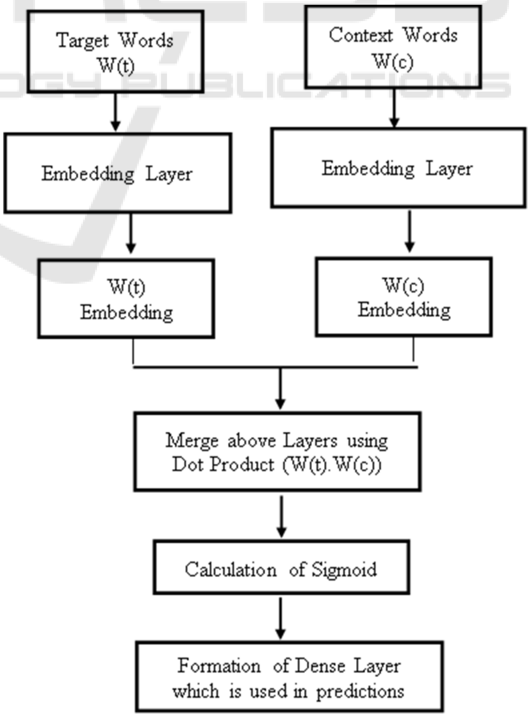

Figure 1: The Work flow of Embedding Layer.

Shukla, A. and Gupta, R.
Detection of Context-Dependent Lexical Information from Unstructured Data Using Word Embeddings Based on Machine Learning: An Assessment.
DOI: 10.5220/0012603900003739
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT 2023)*, pages 533-536
ISBN: 978-989-758-661-3
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

category includes methods that use Machine Learning (ML) and Natural Language Processing (NLP).

2 LITERATURE REVIEW

The authors (Myasnikov et.al. 2021) proposed a method for detecting lexical data in tweets. Vacation tweets, inebriated tweets, and sickness tweets were all separated from the rest of the data. The data was then manually classified as lexical or non-lexical tweets. The author (Chow et.al. 2019)created a system that recognize lexical content depending on user input. The authors in (Kamakshi et.al., 2021) emphasized the importance of categorizing private information before training machine learning algorithms. They also suggested a framework for automatically detecting lexical data based on the user's criteria. The Table 1 shows the approaches for detecting the Context-dependent Unstructured Lexical Information using a number of methods:

Table 1: Approaches for Detecting Context-dependent Unstructured Lexical Information.

Approach	Type	Description
Primary Approach	Deep Learning	Neural network is used to classify context-dependent and context-independent Sensitive data
TF–IDF	Traditional Machine Learning	Conventional machine learning algorithm trained using TF–IDF
Extracted Features	Traditional Machine Learning	Conventional machine learning algorithms trained using features extracted from data
Rule-Based	Heuristic	Data are classified as lexical data based on pre-defined rules

3 METHODS FOR DETECTING CONTEXT-DEPENDENT UNSTRUCTURED LEXICAL INFORMATION

The unstructured lexical data is the data that isn't arranged according to a pre-determined data schema and can't be saved in an out-of-date relational database. Text and multimedia are two types of unstructured data that are regularly employed. Many newspapers, as well as e-mails, images, videos, webpages and audio files are available on internet as unstructured source of data (Akoka et.al. 2019).

In order to assess the comparative capacity and weaknesses of different clustering algorithms for unstructured data, a precise standard must be used to quantify the comparative capacity and weaknesses of each strategy using unstructured data characteristics such as Velocity, Volume, and Variety. Table 2 shows the list the Lexical Information Categories and its related words (Mouza et.al., 2019).

Table 2: A list showing the Lexical Information Categories and its related words.

Category	Words
Place and facility	beach, coast, hotel, conference, island, airport, flight
Positive	go, going, gonna, leave, leaving, pack, booked, before, will, until, wait, plan, ready, here Icome
Negative	need, wish, not, no, want, wanna, back, went, may, might, maybe, had, recent, was, were, could, should, hope, got, suppose, if

4 PROPOSED METHODOLOGY

The proposed method Context-Centered Extraction of Concepts (CCEC) comprises an embedding layer that may be utilized to train neural networks using text input. The integer encoding of the input data is required with each word represented by a separate number. This data preparation phase can be completed with the tokenizer. It must specify three arguments:

- The number of words in the text data's vocabulary is input dim. The vocabulary will be 11 words long if the data is integer encoded with values ranging from 0 to 10.
- Output dim: This is the size of the vector space where words will be embedded. It determines the size of the output vectors from this layer for each word.
- Input length: For each input layer, this is the maximum length of input sequences.

The positive Input Sample training data is in the type [(target, context), 1], with target denoting the target or center word, context denoting the surrounding context words, and label 1 denoting if the pair is meaningful. For Negative Input Samples, the training data will be in the same format like [(target, random), 0]. In place of the real surrounding words, randomly selected words are mixed in with the target words, with a label of 0 indicating that the combination is meaningless (Gómez-Hidalgo et.al. 2016).

The proposed model's operation is illustrated in Fig. 1 and described in the steps below:

- The target and context word pairs are fed to individual embedding layers, resulting in dense word embedding for each of these two words.
- Use a 'merge layer' to compute the dot product of these two embeddings to get the dot product value.
- The dot product's result is then sent to a dense sigmoid layer, which outputs 0 or 1. (Sigmoid layers commonly have a return value (on the y axis) in the range of 0 or 1.) To update the embedding layer, the output is compared to the real label.

The CCEC is analyzed in terms of four aspects. First, we evaluate our method's effectiveness. The underlying major outcomes of CCEC's in obtained in second phase. In the second phase, the detected central node is similar to the underlying node which is examined individually. Then, in terms of concepts with links, we explore CCEC qualitatively. Finally, we test the robustness of our method by altering the parameter, which decides how many of the top-ranking candidate concepts to include as concepts at the end of the second phase.

5 EXPERIMENT RESULTS

In this part of study, it is started with the outcomes of conventional machine learning approaches and then move on to the results of TF-IDF and Rule-based methods. While evaluating the various approaches, the data is divided into two partitions: a training set containing 80% of the data and a testing set containing the remaining 20% of the data.

We have chosen a tweet dataset from kaggle.com that comprises roughly one million tweets. The F-Measure, which is derived from information retrieval, assesses the accuracy of pairwise relationship judgments and is also known as pairwise F-Measure. The Precision (P) is derived by dividing the number of accurate decisions - texts from the same category being assigned to the same cluster by the number of assignments, or the number of text pairs with the same cluster membership. The proportion of couples assigned to the same cluster who share the same category membership is known as Recall (R). As a result, the following contingency table is generated.

Precision and recall are calculated as:

$$P = \frac{TP}{TP+FP} \qquad R = \frac{TP}{TP+FN}$$

and the F- Measure is calculated as :

$$F_1 = \frac{(\beta^2+1)PR}{\beta^2 P+R}$$

Where β is a variable function. Now, following is the explanation of the conventional and deep machine learning algorithms and result performance of rule based algorithms also.

6 CONVENTIONAL MACHINE LEARNING CLASSIFIER FOR CONTEXT DEPENDENT LEXICAL FEATURES

The five alternative supervised machine learning methods were recognized utilizing data sensitivity and TF–IDF characteristics. The performance results for the five conventional machine learning classifiers when using context-dependent lexical features are shown in Table 3.

Linear SVM had the highest accuracy, with an F-measure is of 65 percent. The context-dependent characteristics did not function well, as the data demonstrate. The most essential features in identifying tweets, according to the feature importance study, are location, time, and place.

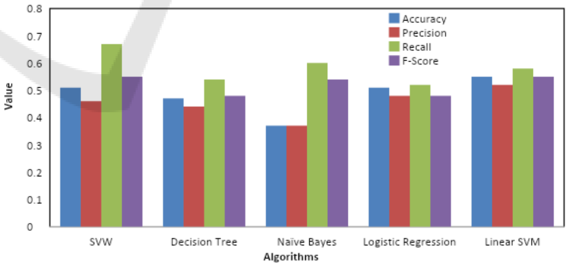

Figure 2: Performance of the Conventional Machine Learning classifiers using Context-Dependent Lexical Features on the Tweet Dataset after Word Embeddings.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 3: Comparing the performance of proposed model in terms of accuracy with related models before and after word embeddings.

Model Name	Type	Data Type	Accuracy (in %)	
			before Embeddings	after Embeddings
Conventional Machine Learning	Statistical Analysis	Tweets	54	65
Term Frequency	TF-IDF	Image	46	57
Rule based on ML	Rule based	Image	62	64
Feature Extraction based ML	FELI	Tweets	80	84
Feature Extraction	FELI	Image	70	74
Rule based Model	Rule based / Term Matching	Tweets	84	89

Fig.2 shows the performance of the Conventional Machine Learning classifiers using Context-Dependent Lexical Features on the Tweet Dataset after Word Embeddings. It is evident from the figure that Linear SVM outperforms for the tweet dataset using Conventional machine learning classification. The performance results for tweet dataset achieved an overall accuracy of 65 percent. The result shows that SVM shows around 55 percent accurate result which is more than Decision Tree and Logistic Regression algorithms. The Naïve Bayes algorithm almost performs similar to SVM. Therefore, achieving the more prominent result, the LSVM can be used for the lexical feature analysis. The text retrieved from the unstructured data produced prominent results. Because lexical information in structured (i.e., personally identifiable information) is lexical regardless of context, so this is the case. The sensitive data, such as birth dates, names, and gender, will always be labeled as such, and will be clearly distinguished from non-lexical data in the context. As a result, it is important to know that the regardless of the method, this type of data is simple to identify. Table 3 shows a comparison of model's performance to that of the earlier proposed methods. As demonstrated, the deep learning approach outperforms existing methods such as TF–IDF and models based on statistically derived features.

7 CONCLUSIONS

Different approaches were tested on two categories of lexical data: context-dependent lexical data and context-independent lexical data. Identifying context-independent lexical data is far easier than identifying context-dependent lexical data, regardless of the approaches utilised. Word extraction methods such as TF–IDF/Count Vectorizer are used to extract features from the text. Some keywords are more essential than others in determining a text category. These approaches, on the other hand, ignore the text's sequential structure. Deep learning algorithms, on the other hand, do not ignore the sequence structure while providing more weight to significant terms.

REFERENCES

Kopeykina L. and Savchenko A.V.(2021). Automatic Privacy Detection in Scanned Document Images Based on Deep Neural Networks, I. Russian Automation Conference (RusAutoCon), 11–16.

Myasnikov E., Savchenko A. (2021). “Detection of Lexical Textual Information in User Photo Albums on Mobile Devices”, Journal of Computing, 0384–0390.

Chow R., Golle P., Staddon J. (2019). Detecting privacy leaks using corpus-based association rules, Proceeding of the 14th ACM SIGKDD I.Conf. on Knowledge Discovery and Data Mining - KDD 08.

Kamakshi P., Babu A.V. (2019). “Automatic detection of lexical attribute in PPDM”, IEEE I. Conf. on Computational Intelligence and Computing Research, 1–5.

Akoka J., Comyn-Wattiau I., Mouza C.D., Fadili H., Lammari N., Metais E. and Cherfi S.S.-S. (2019). A semantic approach for semi-automatic detection of lexical data, Information Resource Management, J. 27 (4), 23–44.

Mouza C.D., Métais E., Lammari N., Akoka J., Aubonnet T., Comyn-Wattiau I., Fadili H. and Cherfi S.S.-S.d. (2019) Towards an automatic detection of lexical information in a database, 2nd I. Conf. on Advances in Databases, Knowledge, and Data Applications

Heni H. and Gargouri F. (2019). “Towards an automatic detection of lexical information in mongo database, Advanced Intelligent System Computer Intelligent System Design Application, 2019 138–146.

Park J.S., Kim G.W. and Lee D.H. (2018). Sensitive data identification in structured data through genner model based on text generation and NER, in: Proceedings of the 2020 International Conference on Computing, Networks and Internet of Things, in: CNIOT2020, Association for Computing Machinery, New York, NY, USA, 2020, pp. 36–40.

Trieu L.Q., Tran T.N., Tran M.K. and Tran M.T. (2018) Document sensitivity classification for data leakage prevention with twitter-based document embedding and query expansion, in: 2017 13th International Conference on Computational Intelligence and Security (CIS), 537–542.

Gómez-Hidalgo J.M., Martín-Abreu J.M., Nieves J., Santos I., Brezo F. and Bringas P.G. (2016). Data leak prevention through named entity recognition, in: 2010 IEEE Second International Conference on Social Computing, 1129–1134.