IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue, 7, March 2019 E-ISSN: 2347-2693

Extraction and Medical Coding of Adverse Events using Word Embedding

Rajdeep Sarkar[1], Devyani Sampale[2*], Harshita Rai[3]

[1,2,3]Analytics and Insights, Tata Consultancy Services

Corresponding Author: devyani.sampale@tcs.com, Tel.: +91-88307-24988

Available online at: www.ijcseonline.org

Abstract— Machine based identification and coding of Adverse Events mentioned in the natural language text received in Pharmacovigilance through the different sources be it emails, faxes, literature, complaint reports, forms, study literatures or phone call transcripts poses a compound problem as it deals with two sub problems, firstly Named entity recognition to detect events (symptoms, illnesses, adverse events) and secondly mapping events to a standard medical coding scheme such as MedDRA. Additionally, industrial applications are required to build their systems in compliance with regulatory requirements, thereby limiting their ability. In this paper we focus on mapping laymen terms or adverse event verbatim/ description to the standard medical terms. We use vector representation of standard medical term dictionary Medical Dictionary for Regulatory Activities (MedDRA) and that of the event verbatim along with their similarity score to establish these mappings.

Keywords— *Medical Coding*, Adverse Event Identification, Cosine Similarity, Word2Vec, Semi supervised machine learning.

I. Introduction

An adverse drug reaction (ADR) is a harmful reaction/side effects caused by taking a drug/medication. Recent studies have shown that ADRs caused by drugs, in their post marketing phase, are of serious health concern. Early detection of such ADRs, both in clinical trials and post marketing phase has become very important for the drug manufacturers as well as regularity organizations such as FDA and WHO. Organizations like the FDA or the European Medicines Agency (EMA) have put various efforts to accomplish voluntary and mandatory reporting of Adverse Drug Reactions through platforms like the FDA's Adverse Event Reporting System (FAERS), the Institute of Safe Medication Practices' Medication Error Reporting System (MERP) and MedWatch. It is mandatory for drug manufacturers to report adverse reactions and general public could report it voluntarily. These systems, sometimes, suffer from under-reporting [1], over-reporting, incorrect reporting, duplicated reporting, and sometimes late reporting of ADRs as well, which could cause serious regulatory penalties for the drug manufacturers. Due to such issues, research to speed up the automation of Pharmacovigilance process has gained attention in recent years. Pharmacovigilance (PV) [2], also known as drug safety, is the pharmacological process of collecting, monitoring, detecting, assessing the causal relationship and ensuring prevention of adverse effects with pharmaceutical products both during clinical trials and post marketing phase. It is a practice of monitoring the side effects of drugs, especially, after the drug has been issued for use in the market, in order to identify any adverse reactions caused after drug admission which were previously not reported or detected. This reporting is done through various sources like complaint forms, literatures, legal cases and social media which can be in any format like PDFs, XMLs, and scanned images. Most of the reporting is now done through electronic communication to standardize the information being shared and make this sharing faster. After receiving the reports, medical professional identifies the adverse events and assign medical coding for the identified terms. With these advancements in the transmission and standardization of the reporting methods the demand to automate Information extraction in the bio-medical domain, from these sources, has seen much focus in the recent past. Extraction of ADRs from the natural language text and identifying the standard term or code that the adverse event phrase refers to, based on to any industry standard dictionary like MedDRA is gaining attention of the research community working in “Pharmacovigilance” recently. In this paper we propose a method to identify the adverse drug reaction that have been mentioned in patient reports whether it is through clinical trial or spontaneous case reports. In order to identify these phrases we make use of grammatical structure of the sentences in the text. After extracting the verbatim from the natural language text, we propose a method to map these verbatim to standard MedDAR LLT and if we are able to map these verbatim phrases to the standard knowledge based then we identify it as an ADR.

II. Related Work

The research in the early days focused on building rules, lexical chains, term mappings and ontologies for Named

Entity Recognition (NER) [3] and Relation Extraction (RE) [4] tasks. In [5] the authors have discussed approaches that have used ontologies and dictionaries to identify entities such as adverse events and drug in online medical forums data, the limitation in methods that it make use of dictionaries, these dictionaries are not exhaustive and thus, are not scalable. In [6] the authors have used manually annotated to train their model which predicts the adverse event, the drawback of using such annotations is that we need to put in addition effort to annotated new data if the model needs to be retrained with the change in input data over a period of time. The trend, more recently, has been to engage probabilistic methods. This has been possible due to the increase in computing power, availability of resources online and constant upgradation of machine learning techniques. [7] presents an approach that uses word embedding models for entity recognition in bio medical data. Word embedding have been pervasively utilized in biomedical Natural Language Processing (NLP) applications because of the vector portrayals of words catching valuable semantic properties and etymological connections between words. Distinctive literary assets (e.g., Wikipedia and biomedical writing corpus) have been used in biomedical NLP to prepare word They have used Word embedding have been pervasively utilized in biomedical Natural Language Processing (NLP) applications because of the vector portrayals of words catching valuable semantic properties and etymological connections between words. Distinctive literary assets (e.g., Wikipedia and biomedical writing corpus) have been used in biomedical NLP to prepare word embedding and these word embedding have been normally utilized as highlight contribution to downstream machine learning models. Be that as it may, there has been little work on assessing the word embedding prepared from various literary assets. Strategies In this investigation, they exactly assessed word embedding prepared from four unique corpora, in particular clinical notes, biomedical distributions, Wikipedia, and news. For the previous two assets, they prepared word embedding utilizing unstructured electronic record (EHR) information accessible at Mayo Clinic and articles (MedLit) from PubMed Central, individually. For the last two assets, we utilized openly accessible pre-prepared word embedding, GloVe and Google News. The assessment was done subjectively and quantitatively. For the subjective assessment, they discretionarily chosen medicinal terms from three therapeutic classes (i.e., turmoil, manifestation, and tranquilize), and physically assessed the five most comparable words processed by word embedding for every one of them. We likewise broke down the word embedding through a 2-dimensional representation plot of 377 medicinal terms. For the quantitative assessment, we led both characteristic and outward assessment. For the natural assessment, we assessed the restorative semantics of word embedding utilizing four distributed datasets for estimating semantic comparability between medicinal terms, i.e., Pedersen's dataset, Hliaoutakis' dataset, MayoSRS, and UMNSRS. For the outward assessment, we connected word embedding to various downstream biomedical NLP applications, including clinical data extraction, biomedical data recovery (IR), and connection extraction (RE), with information from shared assignments discretionarily chosen medicinal terms from three therapeutic classes (i.e., turmoil, manifestation, and tranquilize), and physically assessed the five most comparable words processed by word embedding for every one of them. We likewise broke down the word embedding through a 2-dimensional representation plot of 377 medicinal terms. For the quantitative assessment, we led both characteristic and outward assessment. For the natural assessment, we assessed the restorative semantics of word embedding utilizing four distributed datasets for estimating semantic comparability between medicinal terms, i.e., Pedersen's dataset, Hliaoutakis' dataset, MayoSRS, and UMNSRS. For the outward assessment, we connected word embedding to various downstream biomedical NLP applications, including clinical data extraction, biomedical Information recovery (IR), and relation extraction (RE), with information from shared assignments. They have also used similarity methods to determine the similarity between the extracted entities. [8] presents a comparison between the different similarity method including semantic similarity, cosine similarity among others.

III. METHODOLOGY

In this paper we are dealing with a problem which can be sub divided into two problems. The first one is to extract event phase from a given adverse event verbatim and the second one is to associate these identified event phrase to industry standard dictionary like MedDRA [9]. The proposed approach begins with the identification of events from the verbatim, while processing natural language text there are scenario that a single event verbatim may contain multiple events in it. So, we begin by pre-processing the sentences and then extracting meaningful phrases from a verbatim. In order to extract the phrases, we use combination of sentence clauses and noun-chunks from that verbatim. To extract clauses and noun-chunks, we use dependency parsing of a sentence to traverse a tree like structure and generate noun chunks and clauses based on grammatical and lexical rules of the language. Depending on how these parts of speech (POS) tags are connected, sentence clauses of the form subject + verb + object are derived. The rules use POS tags [10] of the words in that sentence to identify the subject and object of the clauses which have the POS tag noun or proper noun and identity action as the words that are tagged as verb. Sentences clauses of a sentence could be called as crisp representation of that sentence which holds its exact context are can be used to extract event phrases.[11] The second problem aims at associating these extracted colloquial phrases to standard MedDRA term. Past approaches talk about using lexicon based techniques to create ontologies which map these colloquial terms with standard medical terms. As these ontologies were not exhaustive enough, there was a need for a technique that would capture and hold the context of these colloquial phrases which would be similar to the context of its corresponding medical term. This is where we used Word2vec which is a language modelling technique that generates word embedding to reconstruct linguistic contexts of words. Word embedding are nothing but mapping of words from the vocab to vectors of real numbers which are formed on the basis of the context in which the word is used. Word2vec is group of related shallow models with two-layer neural networks that are trained on a large corpus to produce a vector space, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space. We trained word2vec model on large corpus which described the medical concepts in colloquial terms or layman language. The corpus to train the model was created using open source case reports from PubMed and scrapped from medocaldictionary.com. This trained model is then used to create a knowledge base from MEdDRA. The model generates vectors for medical terms present in MedDRA, the vectored MedDRA which has the low level term and their respective vector for each medical condition present in the dictionary, this form the Knowledge base. Now we aim to generate resultant word vector for the event phrase which was extracted in the first step, the normalized vector for the phrase is computed based on the word vectors for each word in the event phrase. This resultant vector for event phrase is used to query the knowledge base which returns the low level term which is most similar event phrase. We have used cosine similarity to determine the similarity between the two vectors. This medical term is considered to be the coding derived for that particular event phrase.

The presented approach has used MedDRA as the standard dictionary. MedDRA is a multileveled framework consisting five levels. At the highest point of the chain of command are the 26 SOCs (take note of these relate to 'body frameworks' in COSTART, some still utilize this term wrongly in connection to MedDRA). The vast majority of the factual yields utilized by an administrative essayist for security revealing will be founded on favoured terms (viewed as a solitary therapeutic idea), gathered into SOCs as a rule. Beneath the favoured terms come LLTs, which regularly give equivalent words to favoured terms. The accessibility of a few LLTs for a favoured term helps with coding in light of the fact that there is probably going to be a nearby match with the verbatim terms recorded. Figure 1.1 shows the flow diagram for the proposed system.

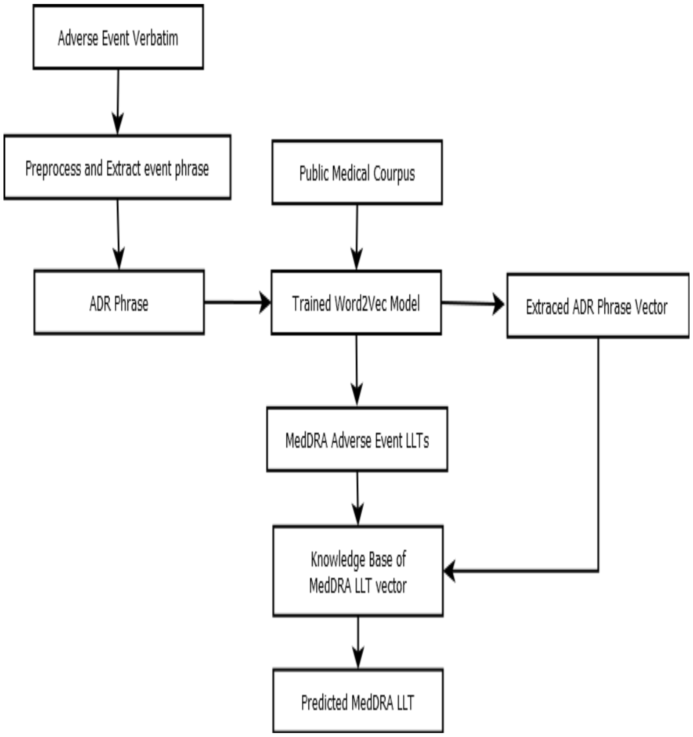

Figure 1.1

IV. Dataset Used

The dataset for the experimentation in this paper was created by combining patient report data from Mayo clinic and medical literature from PubMed. These patient report include case reports and clinical study literature document. We were able to scrape websites to extract medical narratives.

V. Results and Discussion

For the purpose of experimentation in this paper we have used of corpus of 17 million reports to train our Word2Vec model. The resultant vocabulary size of the model after training was around 300000 words which includes stop words. The proposed approach had a combined accuracy of 78% for both the steps. The standalone accuracy of the adverse event identification module is 76% and that of the model is 83%. We see a decrease in the accuracy of the model when use phrase to identity the LLT because of the accuracy of the identification module is effected if the natural language sentence not grammatically well formed or if the noise cannot be removed while we are pre-processing the text. The table 1.2 shows the similarity matrix for some sample phrases and their predicted LLT term.

Table: 1.2

Extracted Adverse Event Phrases	MedDRA LLT	Cosine Similarity
redness in eyes	eye redness	0.99
bleeding in lungs	lung haemorrhage	0.83
feeling tired	feeling hungry	0.94
infection in ears	ear infection nos	1.0
pyrexia	Pyrexia	1.0

The figure 2.1 shows the vectors for the ADR verbatim that the trained Word2Vec model generated.

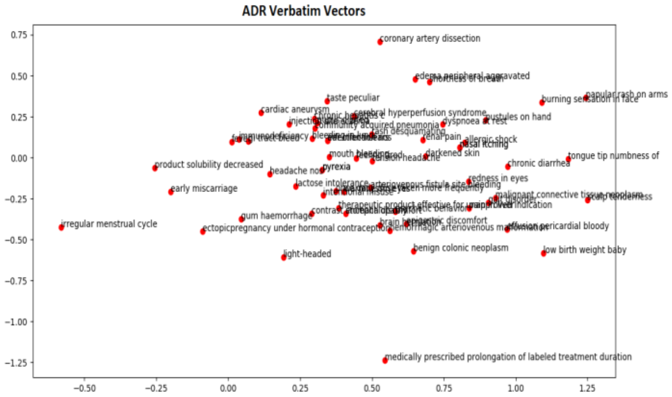

Figure: **2.1**

Figure 2.2 shows the vectors for the MedDRA LLTs that form the Knowledge base**.**

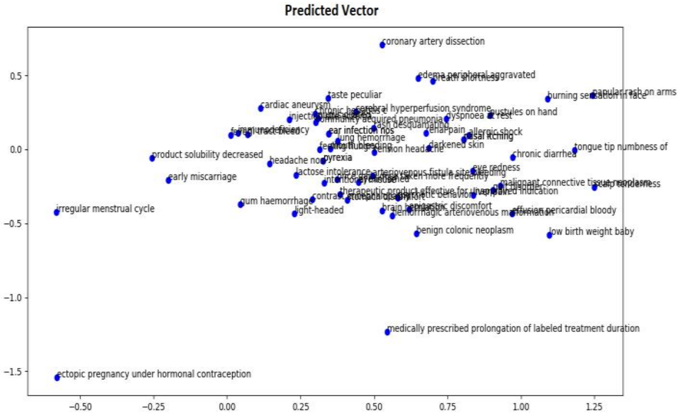

Figure**: 2.2**

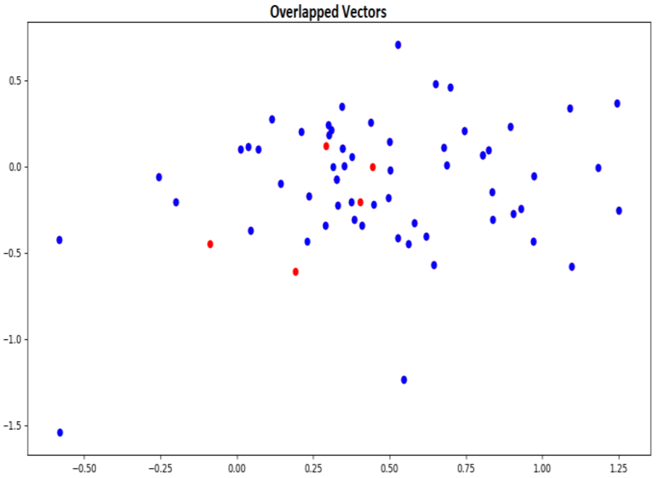

Figure**: 2.3**

The figure 2.3 shows the overlap between the two. The red point are the verbatim for which the proposed system was not able to predict the MedDRA LLTs and the blue points are the verbatim for which the we were able to predict the MedDRA LLTs.

The other dictionary and Bag of word based approaches are accurate when the ADR verbatim contain proper medical terms and not when they have layman description of the medical event, also the both the approaches are not able to take into the account the context of the ADR i.e. whether the ADR is mentioned in a negation or not. Our approach is able to handle both the scenarios, we are able to identify negations in the sentence context while we are extracting the event verbatim from the sentences.

VI. Conclusion and Future Scope

The proposed method shows accuracies that are better than any of the rule or dictionary based approaches as it does not depend on an exhaustive dictionary instead it can take into account the context of the event phrase. The resulting vector is based on the vector of the constituting terms. This approach does have some drawbacks due to the fact that Wordc2Vec does not take into consideration the different meaning of a word that could exist and thus may lead to ambiguity when the word has different part of speech tags but same vector representation. For the future scope we could use word embedding model such as Sense2Vec [12], which take into account the context or the differing meaning in which a word can be used and have differing vector representation for each sense. This could handle the inherent ambiguity in the language of the text.

ACKNOWLEDGMENT

The authors would like to thank Mrs. Aparna Joshi, Strategy & Incubation Head, Pune, Data and Analytics (DnA) Practice, Tata Consultancy Services for her early support, suggestions and guidance throughout the research. The authors are also grateful for support from Mrs. Sonal Jadhav, Technical Lead, Strategy & Incubation Head, Pune, Data and Analytics (DnA) Practice, Tata Consultancy Services for reviewing and providing meaningful suggestions which added value to this paper

REFERENCES

[1] Hazell, Lorna, and Saad AW Shakir. "Under-reporting of adverse drug reactions." *Drug safety* 29.5 (2006): 385-396.

[2] World Health Organization. "The importance of pharmacovigilance." (2002).

[3] Derczynski, Leon, et al. "Analysis of named entity recognition and linking for tweets." *Information Processing & Management*51.2 (2015): 32-49.

[4] Lin, Yankai, et al. "Neural relation extraction with selective attention over instances." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2016.

[5] Gupta, Sonal, et al. "Induced lexico-syntactic patterns improve information extraction from online medical forums." *Journal of the American Medical Informatics Association* 21.5 (2014): 902-909.

[6] Liu, Xiao, and Hsinchun Chen. "AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums." *International conference on smart health*. Springer, Berlin, Heidelberg, 2013.

[7] Wang, Yanshan, et al. "A comparison of word embeddings for the biomedical natural language processing." *Journal of biomedical informatics* 87 (2018): 12-20.

[8] Pedersen, Ted, et al. "Measures of semantic similarity and relatedness in the biomedical domain." Journal of biomedical informatics 40.3 (2007): 288-299.

[9] Morley, Greg. "Adverse event reporting: A brief overview of MedDRA." Medical Writing 23.2 (2014): 113-116.

[10] T Baldwin, MC de Marneffe, B Han, YB Kim Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition-2015

[11] Leaman, Robert, et al. "Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks." Proceedings of the 2010 workshop on biomedical natural language processing. Association for Computational Linguistics, 2010.

[12] Trask, Andrew, Phil Michalak, and John Liu. "sense2vec-A fast and accurate method for word sense disambiguation in neural word embeddings." arXiv preprint arXiv:1511.06388 (2015).

Authors Profile

Mr. Rajdeep Sarkar pursed Bachelor of Engineering from University of Kokata. He is currently working as a Data Scientist with nearly 15+ years of industry experience in NLP, Machine Learning and Deep Learning.

Ms Devyani Sampale pursed Bachelor of Engineering in 2013 from University of Aurangabad. She is currently working as a Data Scientist with nearly 5+ years of industry experience in NLP, Machine
Learning.

Ms Harshita Rai pursed Bachelor of Computer Applications in 2014 and Masters of Computer Applications in 2017 from University of Allahabad. She has industry experience in NLP and Machine Learning for 1.2 years.