2016 International Conference on Advances in ICT for Emerging Regions (ICTer): 155 - 161

An Automated Method to Extract Information in the Biomedical Literature about Interactions between Drugs

D. Mahendran and R.D. Nawarathna*

Department of Statistics and Computer Science, University of Peradeniya, Peradeniya 20400, Sri Lanka

*ruwand@pdn.ac.lk

***Abstract* ---- Text mining techniques are useful in extracting hidden information about the biomedical interactions such as Protein-Protein, Drug-Drug and Protein-Drug. Recently, there is an increased interest in automated methods due to the vast growth in the volume of published text regarding biomedical interactions. This work mainly focuses on extraction of Drug-Drug interactions (DDIs) in biomedical research articles from well-known databases such as DrugBank and MedLine. The proposed approach is developed based on feature engineering through natural language processing (NLP) techniques such as bag-of-words approach, tokenization, part-of-speech (POS) tagging, lemmatization and so on. This uncomplicated and easy to implement set of features are combined into a feature vector which is used to train a machine learning model. The effectiveness of the proposed approach was measured by conducting several experiments on the "DDI Extraction 2013" corpus. The system showed encouraging F-measure value of 76.9%.**

***Keywords* ---- Drug-Drug Interaction (DDI), Natural language processing, Bag-of-words**

I. INTRODUCTION

Biomedical text mining refers to the application of text-mining techniques to unstructured text in the biomedical domain to extract useful information [1]. Biomedical literature mostly includes published articles, journals, essays and other text about new drugs, genetics, biochemistry, proteins and so on. The volume of published biomedical text is expanding at an increasing rate. As more and more text becomes available there can be observed a great interest in extracting useful information hidden in these articles. Applications of biomedical text mining include genome and gene expression annotation, drug–target discovery, biomedical interaction identification, drug repositioning, and electronic health records maintenance. Out of these applications, processing of articles to identify important biomedical interaction is considered as an integral part in the biomedical text mining [2]. Biomedical interaction provides the details about a relationship between two biological entities such as two proteins, two genes, two drugs and many others. Knowledge about the biomedical interactions helps to understand the mechanisms of the living organisms better. Three such important interactions are Protein – Protein interactions (PPI), Protein – Drug interactions (PDI), and Drug – Drug interactions (DDI). This research mainly focuses on DDI information extraction.

A. Drug-Drug Interactions (DDI)

Some drugs interact with one another when taken into the body at the same time. DDIs occur when one drug influences the level or activity of another drug. The drugs involved here can be prescribed medications, over-the-counter medicines, even vitamins and herbal products. Knowledge of all potential interactions is very important for physicians who prescribe varying combinations of drugs for the patients [3]. Not all drug-drug interactions result in the same manner. For example, Aspirin and blood thinners like Warfarin Coumadin help to reduce the risk of heart attacks and strokes [4]. Both drugs are in use for a long time to prevent dangerous blood clots from forming during heart attacks. However, both drugs taken together can cause unpleasant side effects such as excessive bleeding.

Extraction of information about interactions between drugs is an important field of research for both biological and medical researchers. New drug-drug interactions are discovered every day. As a result of that a huge number of biomedical articles and research papers are published every year presenting these interactions. Scientists in the biomedical field are aided with some online databases, namely, MEDLINE, PubMed and DrugBank [5] for this purpose. There are online medical apps created by the medical experts, for example Medscape and Micromedex, that use a backend database of drug-drug interactions. These apps are used by doctors to validate interactions between drugs when prescribing them. Therefore, these databases need to be updated more often for doctors to accurately prescribe drugs to patients as well as to identify more drug-drug interactions.

Although many efforts have been made, much information about interactions still remains in the unstructured textual format. Extracting information manually from the literature is extremely time-consuming and the explosion of information in the biomedical domain in recent years has made this task almost impractical. This situation has naturally led to an interest in automated techniques. Automation of the drug-drug interactions from the literature, would improve the drug databases that store this information by a great deal. This also solves the problem of managing the continued growth of new literature on drugs being published.

II. RELATED WORK

In the studies that have been conducted so far on extracting biomedical interactions, three common approaches can be underlined, specifically, dictionary-based, rule-based and machine learning based approaches. Early studies were performed mainly based on ontologies and dictionaries due to lack of labeled corpora. Dictionary-based methods identify drug names in biomedical texts by using lists of terms in drug dictionaries whereas the rule-

978-1-5090-6078-8/16/$31.00 ©2016 IEEE

based methods are based on a set of rules defined to extract information about the interactions and classify according to the created rules. Machine learning based methods build machine learning models based on labeled corpora to identify drug names.

ChemSpot [6] is a named entity recognition tool for identifying chemicals and drugs in the text. It also detects trivial names, abbreviations, molecular formulas and IUPAC entities. Basically, the method uses a dictionary-based approach. Sanchez-Cisneros et al. [7] presented a method for Drug Name Recognition (DNR) which is a combination of both ontology-based and dictionary-based methods.

To encourage researches on information extraction two challenges were organized: DDIExtraction 2011 and DDIExtraction 2013 [8] [9]. Most DDI Extraction studies have been conducted using the labeled corpora provided during these two challenges. The DDIExtraction 2011 challenge is mainly focused on the identification of all possible pairs of interacting drugs in sentences, without specifying anything further about the interactions [8]. A machine learning-based system was presented by He et al [10] for DNR using the DDIExtraction 2011 corpus. The DDIExtraction 2013 challenge included the task of DNR as well as the classification of the drugs [9]. Bokharaeian et al [11] suggested a method for annotating the DDI-DrugBank corpus with negation cues and scopes. They investigated the negated statements in the corpus and found that they consist of approximately 21% of its sentences. Bjorne et al. have developed a system based on the Turku Event Extraction System [3]. This system relies heavily on deep syntactic parsing to build a representation of the relations between drug mentions. They have tested both support vector machine (SVM) and regularized least-squares (RLS) classifiers and their system achieves a performance of 63% F-measure value on the DDI Extraction 2011 task.

Both dictionary-based and machine learning-based methods have been used during DDIExtraction 2013 challenge. Machine learning-based methods achieved the best performance comparatively. The main contribution of the proposed method of Kim et al [13] is the rich feature based approach using linear SVMs. Since non-linear composite kernel approaches tend to be complex, this linear kernel approach [13] is the practical alternative for large-scale problems and their system achieves an F-measure value of 67%, as compared to 65% and 60.9% in other existing systems.

Most machine learning-based methods have used singleton features which can only capture one linguistic characteristic of a word that is not sufficient to describe the information for DNR when multiple characteristics are considered. Therefore, Shengyu Liu et al. [14] explore feature conjunction and feature selection. They have selected 8 types of singleton features and combine them into conjunction features. But they reported that although the performance was better than previous systems, it was not good enough. Recently Arantza Casillas et al [15] presented a hybrid system which can perform both named entity recognition of drugs and diseases and adverse drug reaction event extraction medical texts in Spanish. Wei Zheng et al [16] came up with the system which uses a graph kernel based method to identify DDIs from biomedical literature using the DDIExtraction 2013 corpus. Using the graph representation of a parsed sentence relations among long–range and short–range words are obtained. They have reported F-measure values of 81.8% and 68.4% for detection and classification of DDIs respectively.

Although many approaches have been used to extract DDIs over the past few years still the accuracy of those automated extraction methods need to be improved.

Over the last decade natural language processing (NLP) together with machine learning techniques have been used to successfully extract information from biomedical texts. In this study a new model consists of a sophisticated set of features learned from NLP techniques and a popular classification algorithm [13] is developed. Regular Expressions, Tokenization, Stemming, Lemmtization, Bag-of-words and POS tagging [17][19] are used to create a feature vector containing a set representative features that precisely describe an interacting and non-interacting drug pair. The major advantage and the novelty of the proposed work is its well-defined set of features. Since the set of features are simple and easy to implement the proposed model can be used in practice without much problems.

The remainder of the paper is organized as follows. The proposed methodology is presented in Section III. The details and results of the experimental study are summarized in Section IV. Finally, some concluding remarks are stated in Section V.

III. METHODOLOGY

In this section, we describe the proposed NLP and machine learning [24] based DDI extraction method that is tasked with extracting information about interactions between two drugs. The main steps of the method can be summarized as shown in Fig. 1.

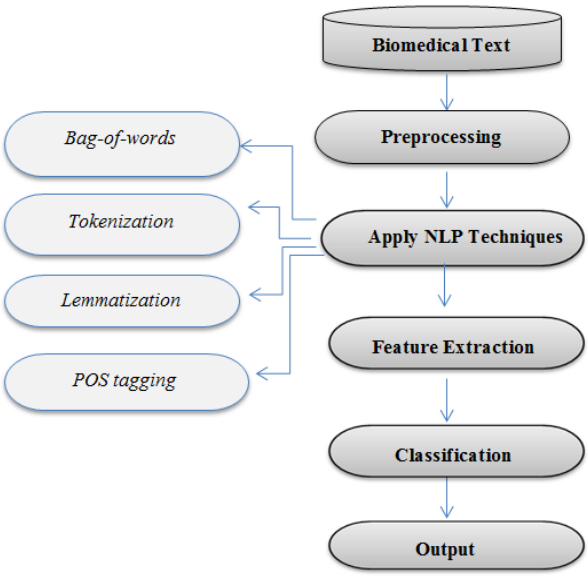

Fig. 1 Steps of the proposed method

When a biomedical text and a drug pair are given as the input into the model, all the sentences discussing about the given drug pair are extracted first. On each sentence the steps shown in Fig. 1 are performed in that order. Then, the model classifies each sentence as "interacting" or "non-interacting" through NLP-based feature extraction and a

learned model. Each step in the method is discussed in detail in the following subsections.

A. Preprocessing Steps

All sentences in the text corpus are preprocessed in order to normalize the corpus as well as to simplify the feature extraction. Regular expressions are used to search the drug pair and extract the corresponding sentences from the text as well as to extract features [18].

Some of the existing NLP techniques and tools are used for preprocessing. Preprocessing is performed as follows.

- Sentences with less than two drug names are removed as the proposed method classifies interactions between two drugs only.
- All letters are changed to lower case.
- Stop words and punctuations are removed.
- Both sentences and words are tokenized using a tokenizer [21]

Stop words are common words of the language that do not contribute to the semantics of the documents and do not contain any significance but has a high frequency. They are usually filtered out during search queries to prevent returning vast amount of unnecessary information [20]. For example, some of the stop words include: a, about, among, the, e.g., as, at and many others. Tokenization is used to break the given text into tokens [21]. A token is an instance of a sequence of characters grouped together as a useful semantic unit for processing. Both sentence and word tokenization are used as text processing is performed on sentence as well as word levels.

The proposed method is developed mainly around the bag-of words model [22] and part-of-speech (POS) tagging [19]. The bag-of-words model is a concept commonly used in text mining for text classification where the frequency of each word is used as a feature to train a classifier. In this work, the bag-of-words model includes all possible positive and negative keywords that describe the relationship between the two drugs of a given sentence. Positive keywords and negative keywords are the most common words used to describe two interacting and non-interacting drugs, respectively. Table I shows a few examples of positive and negative keywords in the bag-of-words model of the proposed method. The bag-of-words is decided by carefully analyzing the text corpus mainly based on domain knowledge, expert recommendations and word frequency analysis.

One of key steps of the proposed method is the use of "Lemmatization". Lemmatization is the process of grouping together the different forms of a word and returning of the base or dictionary form of a word so that they can be analyzed as a single word [19]. For example, the verb "to increase" has many other verb forms such as "increase", "increases", "increased", and "increasing". In this approach, lemmatization is applied to both sets of keywords (positive and negative) to group together the different forms of a particular word and replace it with the base form or the lemma of the word [4]. This reduces the number of words in the bag-of-words and provides more accurate results.

TABLE I
SOME OF THE POSITIVE AND NEGATIVE KEYWORDS INCLUDED IN THE BAG-OF-WORDS MODEL

Positive Keywords	Negative Keywords
interact	no
inhibit	not
not recommended simultaneously	without
contradict	neither/nor
influence	lack
affect / have effect	cannot
increase / enhance	absence
decrease / diminish	unchanged
carefully monitored	unlikely

B. Feature Extraction

During the feature extraction step, a representative set of features is computed for each sentence about a drug pair. Features are extracted using regular expressions. Prior to extracting the features, for each sentence a "DDI Scope" is defined. The DDI Scope depends on the position of the drug pairs and the number of words selected from either side of the two drugs. The range can differ according to the text corpus used. It was experimentally decided that DDI Scope of 5 words provides the best results. A sample DDI scope of a sentence is depicted in Fig. 5. In Fig. 5, the DDI scope of the drug pair containing "ketoconazole" and "terfenadine" is shown using a red colored oval.

Features used in building the proposed model is listed in Table II. Each feature represents one of the key factors in deciding the chance of getting an interacting or non-interacting drug pair. In Table II, Feature F1 to F8 are derived based on the bag-of-words. Feature F9 is derived based on POS tagging. Features are derived using the whole sentence as well as DDI scope only. Extracted positive and negative keywords (See Table 1) in the bag-of-words are lemmatized and used to create Features F1 to F8.

When Feature F1 (Total number of positive keywords in the whole sentence) is high and F2 (Whether one or more positive keyword exists in between the two drug names) is true then there is a high chance of observing an interaction between the two drugs. However, when Feature F3

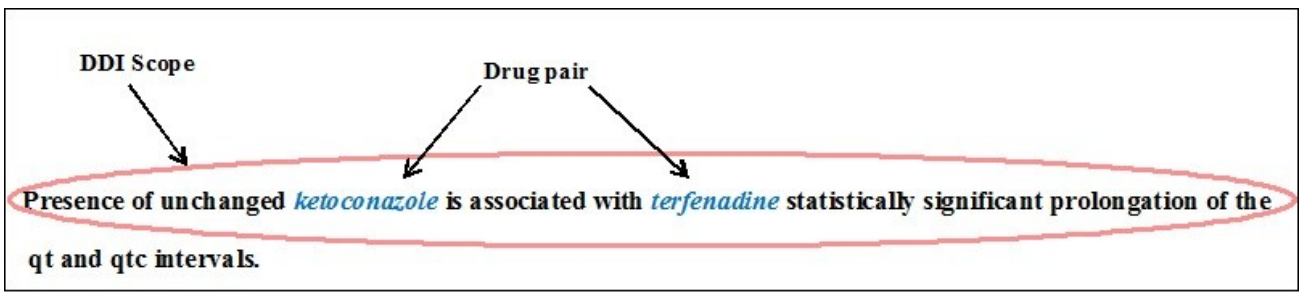

Fig. 2 DDI Scope (red oval) in a sentence about a drug pair

(Whether one or more positive keyword exists within scope but not in between the two drug names) is true, the chance of getting an interacting drug pair becomes lower.

TABLE II
THE SET OF REPRESENTATIVE FEATURES USED TO BUILD THE MODEL OF THE PROPOSED METHOD. FID = FEATURE ID

FID	Description
F1	Number of positive keywords in the whole sentence
F2	Whether one or more positive keyword exists in between the two drug names
F3	Whether one or more positive keyword exists within scope but not in between the two drug names
F4	Number of negative keywords in the whole sentence
F5	Whether a negative keyword exists in between the two drug names
F6	Whether a negative keyword exists within scope but not in between the two drug names
F7	Number of special words in the whole sentence
F8	Number of words in between the two drugs
F9	Number of verbs in between the two drugs with the following part of speech tags: VB, VBD, VBG, VBN, VBP, VBZ

The set of features includes features to represent non-interacting pairs as well. To consider negation in the text, three features (F4, F5 and F6) are defined that use the negative keywords in the bag-of-words. When Feature F4 (Number of negative keywords in the whole sentence) is high and F5 (Whether a negative keyword exists in between the two drug names) is true, there is a high chance of getting a non-interacting drug pair. However, the chance of getting an interacting drug pair gets increased when F6 (Whether a negative keyword exists within scope but not in between the two drug names) is true.

Moreover, the bag-of-words includes a set of special words that increase the probability of receiving an interacting drug pair when included in a sentence. These words basically help to enhance the meaning of positive keywords in the bag-of-words. The special words contain words such as "concomitant administration", "simultaneously", "concurrently", and etc. Such words are identified, extracted and lemmatized using the WordNet Lemmatizer. Feature F7 (Number of special words in the whole sentence) defined based on this special set of words. High value for F7 means that the drug pair may interact. Moreover, it was experimentally found that when Feature F8 (Total number of words in between the 2 drugs) is high; the chances of detecting two non-interacting drugs increase considerably. That means to specify a non-interaction, more words are required in the sentences.

Part-of-speech tagging (POS tagging) [19], is the process of categorizing a word in a text into a particular part of speech, based on both its definition and its context [16]. Words that are belong to the same part of speech generally display similar behavior in terms of syntax. Similarly Feature F9 (Number of verbs in between the 2 drugs) has a clear relationship in deciding whether the two drugs are interacting or non-interacting. F9 represents the total number of verbs that belong to one of the following POS tags: VB (verb, base form), VBD (verb, past tense), VBG (verb, gerund or present participle), VBN (verb, past participle), VBP (verb, non-3rd person singular present), and VBZ (verb, 3rd person singular present) [19].

C. Classification

In the final step of the proposed model, a feature vector is created for each sentence by combining the nine features mentioned in Section III.B. The feature vectors are used to train a classifier such as decision tree, support vector machine, multilayer perceptron and so on. Experimentally it was found that the C4.5 decision tree classifier [23] is the most suitable classifier for the proposed method and LibSVM which is a support vector machine classifier [23] is a good alternative. Then the trained model is used to classify the sentences in the corpus as interacting or non-interacting. Practically, the final label of a drug pair is determined by analyzing all the sentences in the corpus that describes the relationship between the two drugs. That means, if the majority of the sentences are classified as interacting, then the two drugs are declared as interacting and vice versa.

IV. EXPERIMENTAL SETUP AND RESULTS

A. Data Set

The proposed method was evaluated using the DDI Extraction 2013 corpus which consists of texts from both DrugBank and MedLine databases [9]. It is a benchmark dataset commonly used for research in DDI information extraction. 54 documents, originally in XML format, were selected and sentences having both names of the drug pair were used for the experiment as given in Table III.

TABLE III
DETAILS OF THE DATA SET USED FOR EXPERIMENT

	DrugBank Database	MedLine Database	Total
Documents	31	23	54
Sentences	148	52	200
Drug pairs	42	29	71

From the DrugBank database 148 sentences about 42 drug pairs in 31 documents were selected. 52 sentences describing DDIs of 29 drug pairs were considered from the MedLine database. The implementation was carried out using Python. Natural Language Toolkit (NLTK) was used for NLP tasks.

B. Performance of the Model and the Effect of the Choice of the Classifier

To see the performance of the model and the effect of the classifier on the performance, the proposed model was evaluated with the following classifiers [23]: Naïve Bayes (NB), k-nearest neighbor (KNN), support vector machines (LibSVM), multilayer perceptron (MLP), and decision tree (C4.5). Five classifiers were selected to represent five of the main classification algorithms. NB, KNN, LibSVM, MLP and C4.5 classifiers represent Bayesian, instance-based learning, support vector machine, neural network and decision tree type classification algorithms, respectively. The validation is performed using the 10-fold cross validation. We present our results using commonly used performance metrics: recall, specificity, precision, accuracy and F-measure. In all measures positive instance is a sentence classified as interacting drug pair and negative instance is a sentence classified as non-interacting drug pair. The percentage of correctly classified positive instances from the actual number of positive instances is

defined as recall. Precision is the percentage of correctly classified positive instances from the predicted positive instances and specificity is the percentage of correctly classified negative instances from the actual number of negative instances. The accuracy is defined as the number of correctly classified instances from all instances and F-measure is computed as (2 x recall x precision) / (recall + precision). Table IV shows the performance of each classifier on the data set given in Table III.

TABLE IV
ACCURACY, RECALL, SPECIFICITY AND F-MEASURE VALUES OF THE PROPOSED METHOD FOR VARIOUS CLASSIFIERS

Classifier	Accuracy (%)	Recall (%)	Specificity (%)	F-measure (%)
NB	70.4	81.5	57.1	69.8
KNN	71.4	73.1	69.2	71.4
LibSVM	75.9	79.6	71.4	75.8
MLP	76.7	88.0	63.7	76.4
C4.5	76.9	79.6	73.6	**76.9**

From Table IV, it can be concluded that C4.5 decision tree classifier provides highest F-measure value of 76.9% while ensuring equally good values for accuracy, recall and specificity. Comparatively, Naïve Bayes and K-nearest neighbor classifiers have performed with less effectiveness as their F-measure and specificity values are quite low. MLP and LibSVM classifiers have shown similar F-measure values as C4.5 but with less specificity values. It should be noted that testing times of the MLP and LibSVM are somewhat higher compared to C4.5 classifier. When both recall and specificity values are considered LibSVM has performed better compared to MLP classifier. Hence, C4.5 decision tree algorithm can be recommended as the most suitable classifier for the proposed method with LibSVM which is a support vector machine classifier is a very good second choice for the DDI Extraction 2013 corpus. As DDI Extraction 2013 corpus is known as a representative corpus of drug-drug interactions, these conclusions can be valid in general as well.

C. *Effect of Each Feature on the Performance of the Model*

To determine the importance of each feature to the performance of the proposed method, a set of models was created as given in Table V. As listed in Table V, each model was created by excluding some of the features. Model M includes all nine features; Model M-F9 excludes Feature F9, Model-F1-F4 excludes Feature F1 and F4 and so on. C4.5 decision tree classifier was used as the classifier for this experiment as it showed the best performance with all nine features. Table V also shows the F-measure value computed for each model. Moreover, a graph to illustrate F-measure value of each model is provided in Fig. 3.

According to Table V, for Model M-F9 there can be seen a significant decrease in the F-measure value (71.8%). F9 is the number of verbs between the two drugs (See Table II). Even though it is not given in Table V there was a major drop in the specificity as well (65.9%) for Model M-F9. This shows the importance of the POS tagging based feature and its contribution to the overall performance. Also for Model M-F2-F5, we can see a considerable fall in the F-measure value (71.8%) compared to the other models. In Model M-F2-F5, two features, whether one or more positive keywords exist in between the two drug names (F2) and whether one or more negative keywords exist in between the two drug names (F5) are removed. This highlights the importance of Feature F2 and F5 to the proposed model. Similarly, features, number of special words in the whole sentence (F7), whether one or more positive keyword exists within scope but not in between the two drug names (F3) and whether a negative keyword exists within scope but not in between the two drug names (F6) seem to be important features as Model M-F7 and Model-F3-F6 provides comparatively lower F-measure values of 73.2% and 74.1%, separately. F-measure values for other models, M-F1-F4 and M-F8 do not decrease by a substantial margin. Therefore, it can be said that F1, F4 and F8 do not have a huge impact on the final result compared to other features. Overall it can be concluded that the Model M with all nine features provides the best performance and features F2, F3, F5, F6, F7 and F9 play a key role in determining the interaction between two drugs. Fig.3 provides a clear comparison of performance of each model.

TABLE V
PERFORMANCES OF THE MODELS USED TO CHECK THE CONTRIBUTION OF EACH FEATURE. IN HERE, M MEANS THE MODEL WITH ALL FEATURES, M-F9 MEANS THE MODEL EXCLUDING FEATURE F9 AND SO ON.

Model	Description	F-measure (%)
M-F9	Excludes the feature where POS tagging was used to count the number of verbs in between the two drug names (F9)	71.8
M-F1-F4	Excludes the features that counts the number of positive and negative keywords in the whole sentence (F1 and F4)	75.7
M-F7	Excludes the feature that counts the number of words present from the selected set of special keywords (F7)	73.2
M-F8	Excludes the feature that counts the total number of words in between the 2 drugs (F8)	75.4
M-F2-F5	Excludes two features (F2 and F5) that checks whether one or more positive or negative keywords exist in between the two drug names (F2 and F5)	71.8
M-F3-F6	Excludes two features that decide whether one or more positive or negative keywords exist within scope but not in between the two drug names (F3 and F6)	74.1
M	Includes all the features derived using the bag-of-words approach and POS tagging	**76.9**

V. CONCLUSION

The primary objective of this work is to design a system that can extract information about biomedical interactions from biomedical literature. The method focuses on biomedical text that discusses about interactions between drugs. For that purpose, a model is developed to read a sentence from a biomedical text that contains names of two

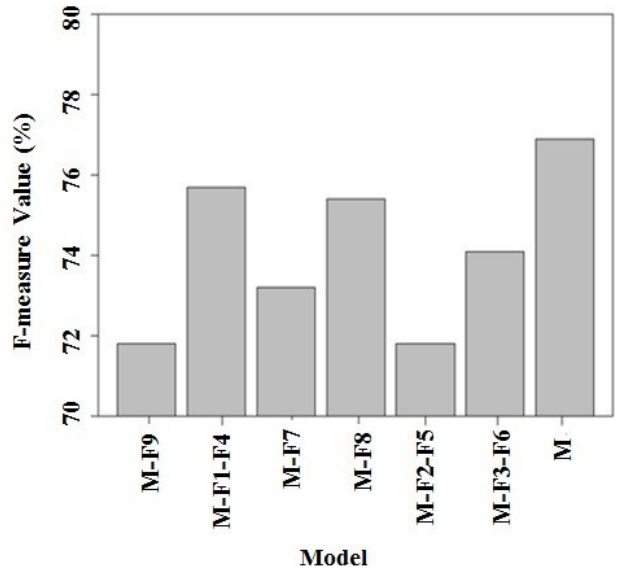

Fig. 3 F-measure values of each model used to test the effect of each feature on the performance of the proposed model. Here, Model M includes all nine features; Model M-F9 excludes Feature F9, Model-F1-F4 excludes Feature F1 and F4 and so on.

drugs and determine whether the two drugs interact or do not interact with each other (Drug-drug interaction (DDI)).

The proposed model developed mainly based on two popular natural language processing (NLP) techniques; bag-of-words approach and part-of-speech (POS) tagging. The model uses nine features derived using bag-of-words and POS tagging that are combined as one feature vector. A classifier is trained with all feature vectors where it learns to classify a given sentence about two drugs as "interacting" and "non-interacting". Since the method uses a set of uncomplicated features, implementation process is quite straightforward and consumes a fewer resources. The proposed model classifies the sentences in the "DDI Extraction 2013" corpus with encouraging F-measure value of 76.9% when nine features are used with a C4.5 decision tree classifier. However, support vector machine type LibSVM classifier showed equally good F-measure value. Further by conducting experiments, the contribution of each feature of the model was investigated separately. It was found that the model with all nine features provides the best performance.

Although the proposed model was developed based on a DDI corpus, it can be easily extended to detect other biomedical interactions such as protein-protein interactions (PPI) and protein-drug interactions (PDI). For both PPI and PDI detection the bag-of-words model should be modified slightly to occupy the contents of the PPI and PDI text. Also feature set can be extended further with more NLP techniques such as bigrams and parse trees. One of the challenges in developing the model was the less number of negative sentences about non-interacting drug pairs (21.8%) in the DDI Extraction 2013 corpus. This imbalanced nature can be combated by incorporating techniques such as resampling, cost matrix and etc.

REFERENCES

[1] R. Rodriguez-Esteban, "Biomedical text mining and its applications," *PLOS Computational Biology,* vol. 5, pp. 12, Dec. 2009

[2] C. D. Manning, P.Raghavan and H. Schütze, *Introduction to Information Retrieval*, UK: Cambridge University Press, 2008.

[3] J.Bjorne, A.Airola, T.Pahikkala and T.Salakoski, "Drug-Drug interaction extraction from biomedical texts with SVM and RLS classifiers," in *Proc. of the 1st Challenge Task on Drug-Drug Interaction Extraction, DDI Extraction 2011,* 2011, pp. 35-42.

[4] National Council on Patient Information and Education (NCPIE), "Be med wise: Frequently asked questions (FAQs) about Drug-Drug interactions involving OTC medications," Feb. 15 2013. [Online]. Available: http://www.bemedwise.org/quiz_facts/facts.htm. [Accessed: Nov. 15 2015].

[5] M. Rastegar-Mojarad, "Extraction and Classification of Drug-Drug Interaction from Biomedical Text Using a Two-Stage Classifier," *Theses and Dissertations,* pp. 304, 2013

[6] T. Rocktäschel, M. Weidlich, and U. Leser, "ChemSpot: A hybrid system for chemical named entity recognition," *Bioinformatics* vol. 28 (12), pp. 1633-1640, Apr. 2012.

[7] D. Sanchez-Cisneros, P. Mart´ınez, and I. Segura-Bedmar,, "Combining dictionaries and ontologies for drug name recognition in biomedical texts," in *Proc. of the 7th international workshop on Data and text mining in biomedical informatics, DTMBIO 2013*, New York, USA, 2013, pp. 27-30.

[8] I. Segura-Bedmar, P.Martinez, and D.Sanchez-Cisneros., "The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts," in *Proc. of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)* Huelva, Spain, 2011, pp. 1-9.

[9] I. Segura-Bedmar, P.Martinez, and M.Herrero-Zazo, "SemEval-2013 Task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013)," in *Proc. of* the *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, 2013, pp 341-350.

[10] L. He, Z. Yang, H. Lin, and Y. Li, "Drug name recognition in biomedical texts: a machine-learning-based method," *Drug Discovery Today,* vol. 19, no. 5, pp. 610–617, May 2014.

[11] B.Bokharaeian, A. Diaz, M.Nevesy, and V.Francisco, "Exploring negation annotations in the DrugDDI corpus," in *Proc. of Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM 2014)*, Harpa, Iceland, 2014.

[12] M.H.Zazo , I.Segura-Bedmar, P.Martínez, and T.Declerck, "The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions," *Journal of Biomedical Informatics,*vol. 46, pp. 914-920, Oct. 2015.

[13] S.Kim, H.Liu, L.Yeganova, W. J. Wilbur, "Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach," *Journal of Biomedical Informatics,* vol. 55, pp. 23-30, Jun. 2015.

[14] S.Liu, B.Tang, Q.Chen, X.Wang, and X.Fan, "Feature engineering for drug name recognition in biomedical texts: Feature conjunction and feature selection," *Computational and Mathematical Methods in Medicine,* vol. 2015*,* Article ID 913489, 9 pages, 2015.

[15] A.Casillas, A.Pérez, M.Oronoz, K.Gojenolab, and S.Santisob, "Learning to extract adverse drug reaction events from electronic health records in Spanish," *Expert Systems with Applications,* vol. 61, pp. 235–245, Nov. 2016.

[16] W.Zhenga, H.Lina, Z.Zhaoa, B.Xua, Y.Zhanga, Z.Yanga, and J.Wanga, "A graph kernel based on context vectors for extracting drug–drug interactions," *Journal of Biomedical Informatics,* vol. 61, pp. 34-43, Jun. 2016.

[17] B.Liu, L.Qian, H.Wang, and G.Zhou, "Dependency-driven feature-based learning for extracting protein-protein interactions from biomedical text," *in Proc. of the 23rd International Confernce on Computational Linguistics (Coling 2010), Beijing, China*, 2010, pp. 757-765.

[18] S. Markel and V. Rajapakse, "Pattern representation:" in *Silico Technologies in Drug Target Identification and Validation*, NewYork, Taylor and Francis group, 2006, pp. 13-40.

[19] S.Bird, E.Klein and E.Loper, "Categorizing and Tagging Words," in *Natural Language Processing with Python*, United states of America, O'Reilly Media, 2009, pp. 179.

[20] V.Trajkovik and A.Mishev, "Influence of stop-words removal on sequence patterns identification," in *ICT Innovations 2013*, Switzerland, Springer, 2014, pp. 67-76.

[21] C.D. Manning, P.Raghavan and H.Schütze,, "Tokenization," in *Introduction to Information Retrieval*, UK, Cambridge University Press., 2008, pp. 23-26.

[22] "Bag-of-Words Model," Jan 2000. [Online]. Available: https://en.wikipedia.org/wiki/Bag-of-words_model. [Accessed Nov. 10 2015].

[23] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classication*, John Wiley and Sons, New York, 2001, pp. 517-581.

[24] A.Thamrongrattanarit, M.Shafir, M.Crivaro, B.Borukhov, and M.Meteer, "What can NLP tell us about BioNLP?," in *Proc. of the 2012 Workshop on Biomedical Natural Language Processing, BioNLP* 2012, *Stroudsburg, PA, USA*, pp. 122-129, 2012.