**Revista de Chimie**
*https://revistadechimie.ro*
*https://doi.org/10.37358/Rev.Chim.1949*

# Biomedical Named Entity Recognition Using the SVM Methodologies and bio Tagging Schemes

**THIYAGU MEENACHISUNDARAM*, MANJULA DHANABALACHANDRAN**
Department of Computer Science and Engineering, CEG Campus, Anna University, Chennai, India

***Abstract:*** *Biomedical Named Entity Recognition (BNER) is identification of entities such as drugs, genes, and chemicals from biomedical text, which help in information extraction from the domain literature. It would allow extracting information such as drug profiles, similar or related drugs and associations between drugs and their targets. This venue presents opportunities for improvement even though many machine learning methods have been applied. The efficiency can be improved in case of biological related chemical entities as there are varied structure and properties. This new approach combines two state-of-the-art algorithms and aims to improve the performance by applying it to varied sets of features including linguistic, orthographic, Morphological, domain features and local context features. It uses the sequence tagging capability of CRF to identify the boundary of the entity and classification efficiency of SVM to detect subtypes in BNER. The method is tested on two different datasets 1) GENIA and 2) CHEMDNER corpus with different types of entities. The result shows that proposed hybrid method enhances the BNER compared to the conventional machine learning algorithms. Moreover the detailed study of SVM and the methodologies has been discussed clearly. The linear and non linear text classification can be mapped clearly in the section 3. The final section describes the results and the evaluation of the proposed method.*

***Keywords:*** *Named Entity Recognition, Conditional Random Field, Support Vector Machines, Hybrid Machine Learning approach*

## 1.Introduction

Named Entity Recognition (NER) refers to identifying and classifying terms belonging to a domain from unstructured text and mapping them to predefined categories. Generally there are three methods for entity recognition [1] i) Rule-based methods ii) Dictionary-based methods and iii) Machine learning based methods. Rule based systems will be effective if patterns can be defined based on Orthographic or Morphological features for all types of entities. But identifying and addition of patterns and disambiguating them is a tiresome task. The Rule-based method is neither robust not portable, that is it has to be kept updated for precise extraction and features varies for every domain. Dictionary based approach is useful if the vocabulary is complete and updated and also requires certain pre-processing such as normalization for matching the text with the vocabulary. Both the approaches does not extracts unseen entities, that is if there is variation in patterns or the term is not present in the vocabulary they are not extracted. Machine learning based approach solves the problem by learning the distinctive features to identify an entity. Primarily supervised machine learning algorithms are used for extracting named entities and it requires large annotated data to learn the features of entities [2]. Analysing characteristics of microbiological characteristics was studied by [3] and comparative analysis was carried out between chemical and microbiological character for analysing antibacterial activities [4].

Unsupervised algorithms can be used to explore and analyse the text for various segments of entities based on their common properties, but needs discriminating features for classifying [5]. Topic model, Distributional information or semantics similarity is used to cluster similar entities and classify them. Semi-supervised approaches are useful if large amount of un-annotated text is available but needs proper selection of seeds with selective features for efficient learning of the task. A common approach for

**email: t.m.thiyagu@gmail.com*

extracting entities across domain is not available as the features and class for every domain [6]. To identify multiple types of entities, the features set has to be equally diverse and to identify them different machine learning algorithms are necessary.

In this work NER from biomedical text is considered, as Biomedical Named Entity Recognition (BM-NER) task has huge impact on tasks such as information extraction and knowledge discovery in the domain. The availability of large unstructured text provides opportunity for knowledge discovery from it using NLP and Machine learning techniques. The biomedical domain has various entities like proteins, genes such as DNA & RNA sequences, drugs, diseases etc. BM-NER presents specific challenges when compared with other domain specific NER such as i) entities vary in size, composition and occurrence ii) does not have standard naming convention to represent the chemical structural information, iii) detecting entity boundaries with precision, iv) ambiguous abbreviations, v) creation and maintenance of domain vocabulary is a tedious task since terms evolve increasingly more, etc.NLP can be used to extract features that are effective for NER and also helps in extracting and disambiguating domain specific entities.

Domain specific NLP plays an important role in identifying entities by adapting the functionalities to the task at hand as domain text varies greatly from generic text in the case of Bio-medical domain. Creation of gold standard datasets and knowledge bases are in favour of supervised machine learning approaches. NER is a natural sequence tagging problem rather than general classification and is a two-step process: 1) Entity boundary detection and 2) Assigning entities pre-defined classes. Various machine learning algorithms are used for NER some of them are Bayesian approaches, Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Support Vector Machines (SVM), Structured Support Vector Machines (SSVM), Conditional Random Fields (CRF)etc. CRF is the most used due to its ability to model multivariate outputs and utilize large set of features for predicting the labels.

The objective of this work is to improve the BNER task with respect to the chemical entities. To overcome the challenges, domain specific NLP is used to extract features and discriminative functions are learned using supervised machine learning algorithm with those features. A novel method to extract biomedical domain entities is proposed and is tested with two different datasets GENIA and CHEMDNER corpus. The approach uses CRF as the tagger and SVM as classifier as both are most effective algorithm for the respective task. The model is tested with different sets of features such as linguistic, orthographic, Morphological, domain features and local context features. The rest of the paper is organized as follows; Related works details the existing methods and process to identify biomedical named entity followed by the proposed methodology. Section Datasets gives the corpus used for named entity recognition and result and evaluation shows the performance of the proposed method. It is followed by discussion on the results that identify venues where improvements can be made and ends with conclusion of the paper.

## Related works

In [7] handle NER as sequence labelling problem and utilizes CRF to tag the Chinese clinical text. To adapt the process to Chinese text, word segmentation, clause level tagging and modified tag set are used. The article [8] utilizes pool-based and uncertainty sampling active learning to annotate clinical text with named entities. [9] uses different features such as linguistic, brown cluster and vector representation on various taggers to tag the clinical text. Tm Chem proposed in [10] applies ensemble of two taggers with varied feature sets along with different pre-processing and post-processing applied on it. Utilize features such as bag-of-word, orthographic features, morphological features, part-of-speech (POS), document structure information, domain knowledge along with different word representations on CRF and SVM machine learning models to tag chemical entities [11]. The article [12] employs hybrid method i.e. rule-based with the use of lexicons and with CRF machine learning algorithm to recognize named entities from Arabic text. In [13, 14] make use of CRF with varied features for medical entity detection. The work proposed in [15] combined SVM, which is used to classify terms as entity and non-entity, with CRF model to assign entity tag to it. The approach in [16] uses dictionary to identify candidate entities and uses neural network and CRF to tag them.

The method in [17] use the vector similarity between the entity class from the knowledge base and the term in the corpus to classify the entity category. The vectors are formed based on the tf-idf values of the terms and Noun phrase chunking is used to select candidate terms. In [18] apply probabilistic generative model to generate features for entity extraction. Word embeddings based on LDA i.e word along with topic, provides different embeddings for different word-topic pair which is used as feature for entity identification. The research in [19] uses lexical resources and search results to identify the boundary of the entities and distributional context to classify them. [20] uses dictionary-based method approach to annotate named entity based on direct match, stemmed match and string edit distance match. Uses external resources like UMLT meta thesaurus for annotate diseases and NCI, MESH, USPMG for identifying medication. Korkontzelos et al. [19] uses dictionary along with aggregate classifier to identify drugs.

Dictionary based approaches [21] utilizes string matching and normalization techniques to match entities in text with the domain based dictionary. In this method, the precision is high but recall is low since only the terms that are matched are extracted. Due to spelling variation and mistakes the string matching methods cannot extract entities efficiently. Also, some entity recognition done using ontology as mentioned in [22]. Also the method suffers from incompleteness that is not all entities are covered and evolve over time, due to which creation and maintenance of domain vocabulary is difficult. Rule-based methods are useful to extract entities that are systematic and follow a pattern. These methods also suffer from covering all patterns and updating patterns based on new entities.

Machine learning algorithms need features that are informative and discriminative so that entities are classified efficiently. As feature selection influence the performance of the algorithms, irrelevant and redundant features are to be neglected and useful features are to be identified using feature selection methods. Features can be classified be as generic and domain dependent features, generic features which can be applied across domain alone are not enough as domain specific features are more effective in identifying domain oriented entities [23, 24]. Based on the algorithm, selected features have to be represented so that the model learns the discriminative function to classify the entities.

## 2. Materials and methods

The objective of this work is to efficiently identify named entities from biomedical literature. Conditional Random field is the most used algorithm for named entity recognition since it combines the capability of discriminative classification and graphical modelling in to one. It considers the context of the input while predicting the sequence labels which is lacking in general classification methods and that makes it a better algorithm for sequence tagging. Support Vector Machine is a state of the art classifier which can perform both linear and non-linear classification. SVM does not consider the context as CRF does and learns a maximum margin classifier to predict the classes. The proposed method utilizes both the algorithms to detect biomedical entities and tag them with their semantic classes.

The drawback of CRF compared to SVM is that it requires more computational space and time. In the proposed method CRF is used to extract entity using the BIO tagging scheme and SVM is used to identify the subtype of the extracted entity. Since SVM is more suitable for only binary classification the multiclass problem is converted to multiple binary classification as it is easy to learn and provides the necessary accuracy. Generally multi-class SVM is constructed using multiple binary class SVM and is carried out by one-vs-rest and one-one pair-wise binary classifier. The one-vs-rest method suffers from class imbalance problem as samples are not equal for both the class. Hence the one-one pair wise classifier is used to identify the subtypes.

To identify ‘N’ different classes N*(N-1)/2 classifiers are constructed and the class is determined by majority voting. The increase in performance of the system comes with the increase in time complexity to learn the model. In order to correctly identify the boundary and also to detect the subtype of an entity, CRF alone is not so efficient since the features used are not so discriminative. Hence the SVM model is used to detect entity subtypes effectively after the entity term extraction.

### Features used

The following sets of features are tested to identify entities by the proposed method:

-Window based context words: the words occurring in the left and right of the given word and the window range from 1 to 3.

-Orthographical information: the special constituents of the given word, such as capital letters, symbols, numbers etc. calculated the number of uppercase and lowercase letters, the number of symbols, number of digits and added as features.

-Roman Numerals and Greek letters: Boolean feature representing the presence of Roman Numerals and Greek letters. It is used as a separate feature as it is domain specific feature.

-Morphological features: prefixes and suffixes present in the term up to a length of five.

-POS tags: POS tag of the given word along with the context words. For each POS, a Boolean feature is added to the feature set.

-Dependency relation feature: Selected dependency relations are used to represent the context. For each relation, Boolean feature is added to the feature set.

-Chemical elements: list of elements and their symbols. Boolean feature identifying the word matches a element or symbol.

-Domain features: presence of prefix and suffix pertaining to drugs or chemicals. Represents characteristics specific to chemicals, including suffixes (e.g. “-yl,” “-oyl,” “-one,” “-ate,” “acid,” etc.), alkane stems (e.g. “meth,” “eth,” “prop” and “tetracos”), trivial rings (e.g. “benzene,” “pyridine” and “toluene”) and simple multipliers ("di,” “tri” and “tetra”)

### SVM and its methodologies

Support Vector Machine (SVM) is a classical machine learning algorithm based on linear model, whose fundamental idea is to change the info space into a high dimensional highlight space by nonlinear change and to locate the ideal straight interface in the new space. As a rule, the higher measurement will prompt the unpredictability of calculation, yet the SVM calculation takes care of the issue in the wake of presenting the piece work, which not exclusively does not expand the computational multifaceted nature, yet additionally maintains a strategic distance from the "Curse of dimensionality".

At the point when the information is straight indivisible, SVM isolates the information directly by mapping to high dimensional element space through bit work. SVM calculation is developed from the ideal isolating surface. Give the preparation a chance to test

$$(p_i, m_i), i \ = 1, \ldots . n, x \in S^d, m \in \{+1, -1\}$$

as the classification mark, to take care of the accompanying quadratic programming issue.

$$min\emptyset(\alpha) = \frac{1}{2}\,||\alpha||^2 = \frac{1}{2}\,(\,\alpha\,.\,\alpha\,)$$

Subject to $m_i[(\alpha.p_i \ + \ b) \geq 1 \ i = 1{,}2 \ldots n$

We derive optimal classification surface as hyper plane:

$$k(p) = \alpha\,.\,p + b = 0$$

At that point use Lagrange streamlining to tackle above issue by changing over the issue to its double issue, to comprehend it with Kuhn-Tucker hypothesis. We can get optimal classification function:

$$\boldsymbol{l}\,(p) = \ sign\ \left\{\sum_{i=1}^{1} \beta_i{}^{*}\ m_i\ (\ p_i.p)\ b^{*}\right\}$$

**Two Types of Linearly Separable Case**

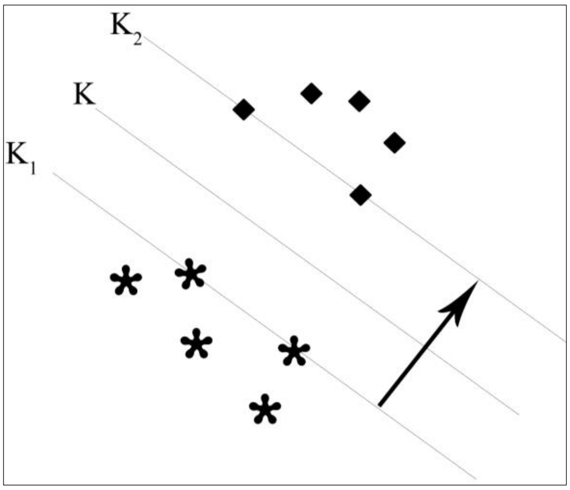

**Figure 1.** Linear classifications

The instance of linear classification as appeared in Figure 1, the five-pointed flowers and the diamond represent two distinct kinds of samples, where K is arranged line. K1 and K2 are straight-lines, which go through various kinds of samples separating the characterized line, and parallel to grouped line K. The separation between line K1 and line K2 is called class interval. The purported optimal separating line alludes to the grouping line which won't just have the option to isolate two sorts of tests effectively, yet additionally to make class interval maximum [25].

Classification line equation can be expressed as:

$$\delta\ .x + b = 0;$$

$$Linear\ sample\ set\ can\ be\ expressed\ as\ (p_i, m_i\ ), i = 1,2, \ldots .n, x \in S^d, m \in \{+1, -1\}$$

Normalize classification line to make linear sample set satisfy the condition:

**Algorithm: Improved supervised SVM algorithm for classifying the samples**

**Step 1:** Input the training data set values $\mathbf{Ds} = (\mathbf{p_i}, \mathbf{m_i}\ )$;

$\mathbf{Linear\ sample\ set\ can\ be\ expressed\ as\ (p_i, m_i\ ), i = 1, 2, \ldots .n, x \in S^d, m \in \{+1, -1\}}$

**Step 2:** Use the SVM classifier to train the sample set S to classify the data models Dm1;

**Step 3:** The data samples can be pre-processed through the SVM sampling methods PrS;

**Step 4:** Categorizing the sampling data's in the corresponding segments and upload the sample set Qs to the segmented region $SG_{reg}$

**Step 5:** Iteration of the sample set Qs can be done until all sample data's were labelled;

**Step 6:** Reusing the complete labelled data sample training set $(\mathbf{p_i}, \mathbf{m_i}\ )$ to get a improved and better classification model Dm2;

**// In few cases different SVM methodologies were used to categorize the samples to get a better output result.//**

**Step 7:** Input the training set to Dm2;

**Step 8:** Output the result

**Task and Evaluation Protocols**

The content extraction is the important task which is extracting the con from scene and web pictures. The content extraction based three things i.e., content location, cropped word acknowledgment, and start to finish acknowledgment. In this content location is used to estimate the zone in the picture or the content available along the vertices. Cropped word acknowledgement is helped to trimmed the related content from the existing picture or web content. Start to finish Recognition, where the goal is to confine and perceive all words in the picture in a solitary advance.

### Nonlinear case

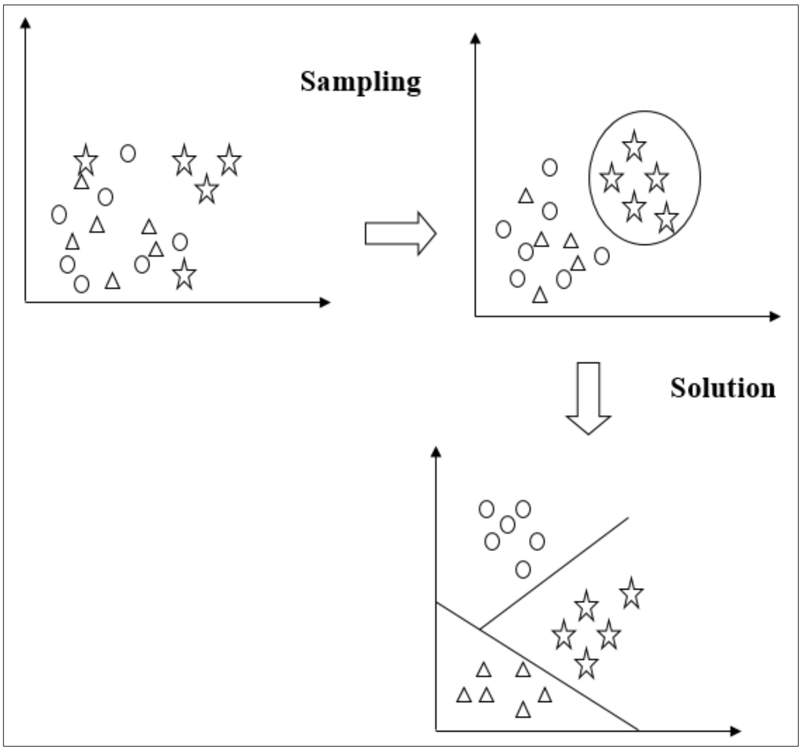

**Figure 2.** Non linear SVM

In the above figure the objects can be classified into different categories. The mixed of objects can be categorized into various different segments to identify each groups in the non linear SVM.The genuine estimation of Support Vector Machine is utilized to tackle nonlinear [26]. The strategy is through a nonlinear mapping to outline test space to a high-dimensional or even vast dimensional component space, with the goal that direct SVM technique in the element space can be connected to take care of the nonlinear arrangement issues in the example space. The nonlinear mapping from the example space to the element space is appeared in the figures.

### Process of text classification

Generally text classification consists of few ways of process. Collecting data’s or data set, pre processing the data’s, feature extraction, classifying the model and the training model were explained in the Figure 3. Text splitting, removing and stopping the text, counting the words of the specified domain and mapping are the few works done by the classifier. In medical scenario the relevant words will be classified using various methodologies.

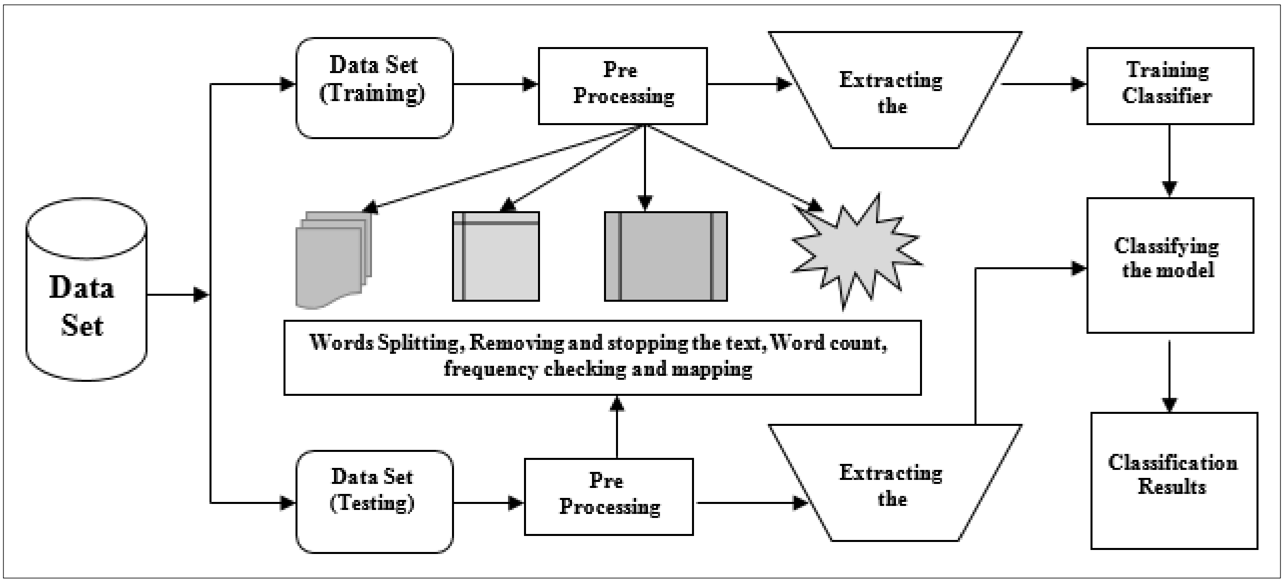

**Figure 3.** General process of text classification

## Pre-processing

The abstracts in the GENIA corpus is divided into sentences using Ling pipe and then basic NLP process such as tokenization, lemmatization, POS tagging and chunking are done using GDep, a biomedical domain dependency parser built using GENIA tagger. The parser produces annotations with BIO tags along with the subtypes for the given abstracts. Similarly sentence tokenization is carried out using Stanford NLP tool for CHEMDNER corpus. Word tokenization is carried out breaking tokens at white space, punctuation's, digits and at case changes. The other features are extracted and then annotations are produced with BIO tags along with the subtypes based on the human annotations.

## Post-processing

It is done to maintain tagging consistency and tag abbreviations effectively as discussed by Leaman et al. [12] if a specific character sequence is tagged more than twice as a chemical mention in an abstract then all other untagged sequence is also tagged as a chemical mention. In case of abbreviations, if the full-form is tagged by the model then its abbreviations are tagged correspondingly. If the abbreviations are tagged and the full-forms are not tagged then the entity tag of abbreviations are removed. As only space tokenization is used entity tagged with unbalanced parenthesis are rejected and taken as false positives.

## Dataset

The GENIA corpus is a collection of biomedical literature which is semantically annotated by humans. The compiled Medline abstracts is used for NER and has five major classes such as protein, DNA, RNA, cell line and cell type and thirty three thousand unique terms. 5-Cross fold validation is performed on the GENIA corpus for evaluating the model on BNER. CHEMDNER corpus is a collection of ten thousand abstracts from various chemistry related documents that are manually annotated with seven different classes - abbreviations, family, formula, identifier, multiple, systematic, trivial. It has around eight four thousand mentions of chemical compounds and drugs entities and is created to aid in the development of named entity recognising tools.

It has three subsets namely 1) a training set containing3500 Medline abstracts annotated with 29478 mentions of chemical entities, 2) a development set composed of 3500 abstracts with 29526 entity mentions and 3) a test set composed of 3000 abstracts, and containing 25351 mentions. To train and improve the model the training and development set is used and to test the model the test set is used.

## 3.Result and discussions

To evaluate the performance of NER applications, the known information extraction measures are used as given below.

$$Precision = TP/(TP + FP)\ (1),\ Recall = TP/(TP + FN)$$

$$F1 = 2 \times Precision \times Recall/(Precision + Recall)$$

where TP refers to true positives, FP to false positives, and FN refers to false negatives.

**Table 1.** GENIA corpus BNER results

| | Proposed Method | | | CRF | | | SVM | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | P | R | F |
| Protien | 84.02 | 93.38 | 88.45 | 75.11 | 59.19 | 66.2 | 86.54 | 17.35 | 28.90 |
| DNA | 93.35 | 87.86 | 90.52 | 83.67 | 74.57 | 78.86 | 82.69 | 23.39 | 36.46 |
| RNA | 96.92 | 89.42 | 93.02 | 90.87 | 97.65 | 94.14 | 84.64 | 14.51 | 24.77 |
| Cell Line | 95.10 | 91.15 | 93.08 | 82.31 | 77.39 | 79.78 | 90.53 | 28.76 | 43.65 |
| Cell Type | 95.10 | 91.15 | 93.08 | 79.61 | 81.55 | 80.57 | 42.35 | 71 | 53.05 |

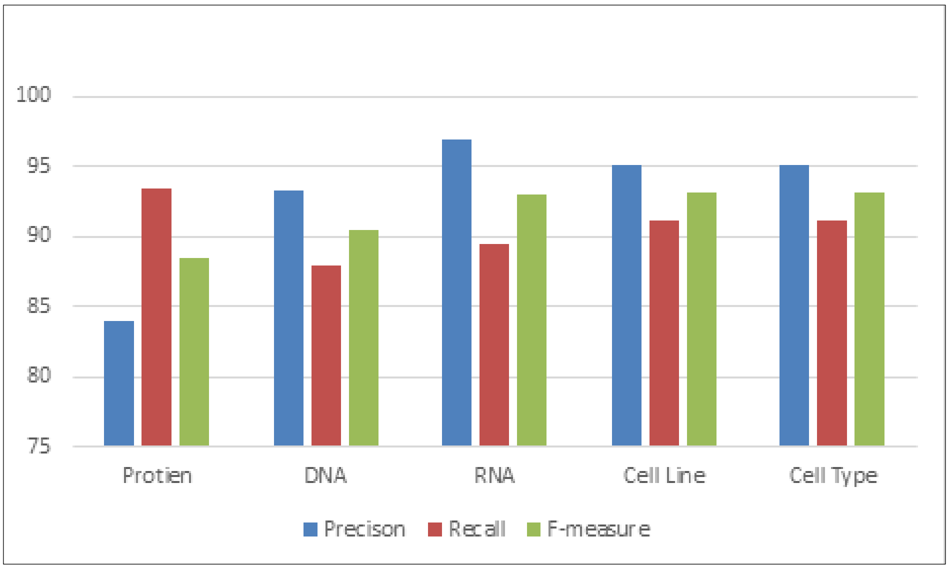

**Figure 4.** Precision Recall and F-measure for various entities in GENIA corpus

Figure 5 describes about the accuracy rate of the different classifiers. Naive Baye's is the standard algorithm in the classification techniques. The text classification method helps to sample the objects in a category using various algorithms. The figure defines that the SVM technique is having highest accuracy rate when compared with Naive bayes, SVD and random forest. Figure 6 describes the duration of the classification method. When comparing all the existing techniques the time taken for the classification is very less in SVM techniques.

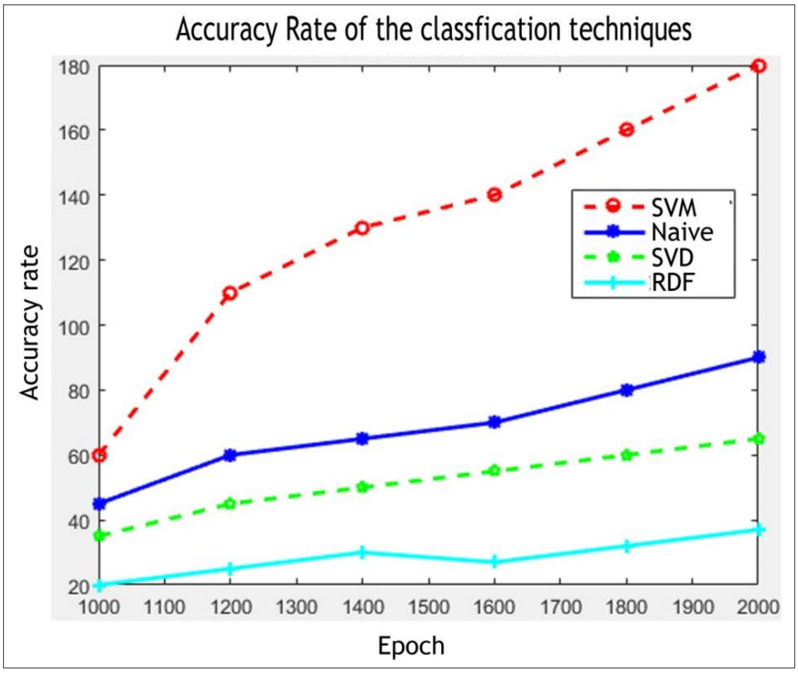

**Figure 5.** Accuracy rate of the classifiers

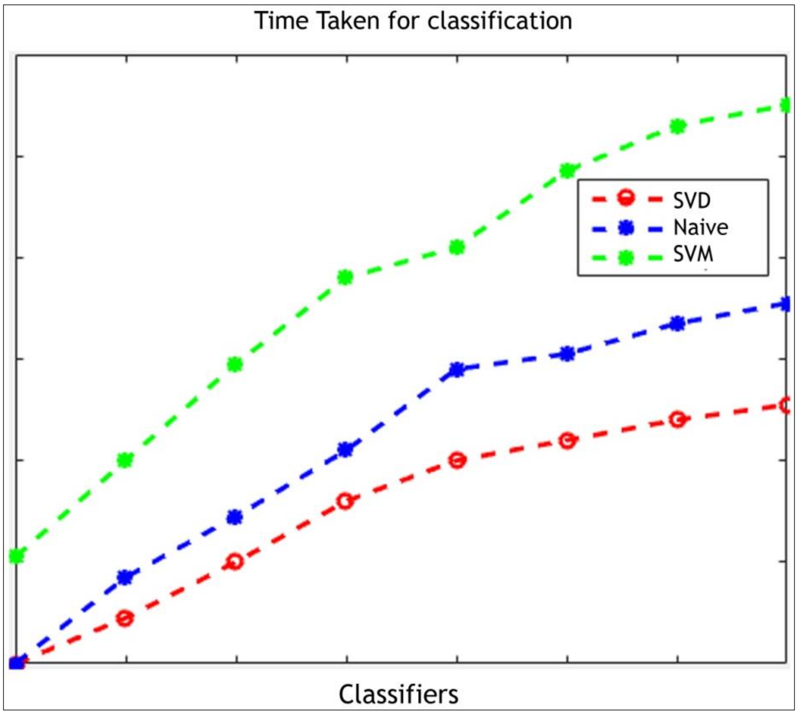

**Figure 6.** Time taken for the classifying process

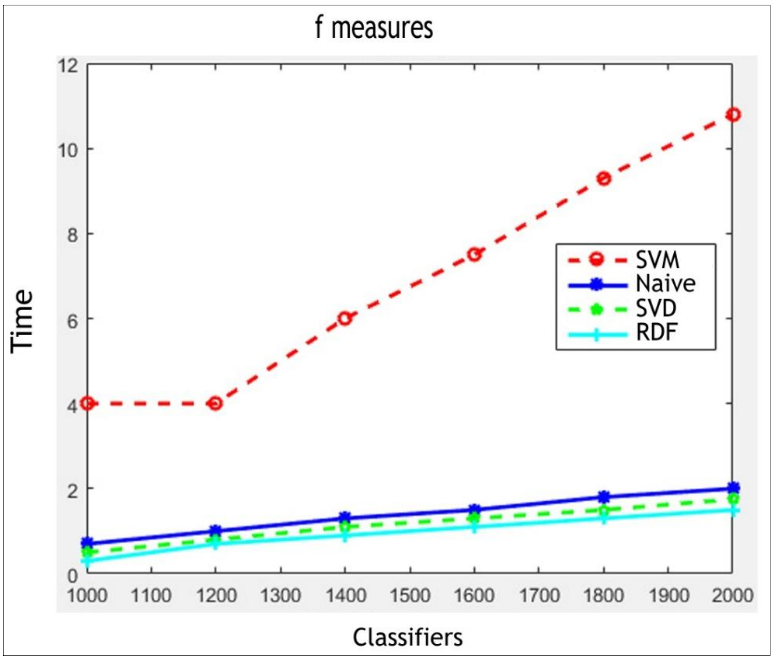

**Figure 7.** f-measure values of the classifiers

The Figure 7 describes the f measures value of the existing technique. The f measure of SVM is greater than other compared techniques. Figure 8 defines the loss percentage of the word while classification. The loss of words in the classification technique leads to poor result in the process flow. Our proposed technique is having very less loss percentage. The losing of words in classification technique is quite low when compared with other techniques.

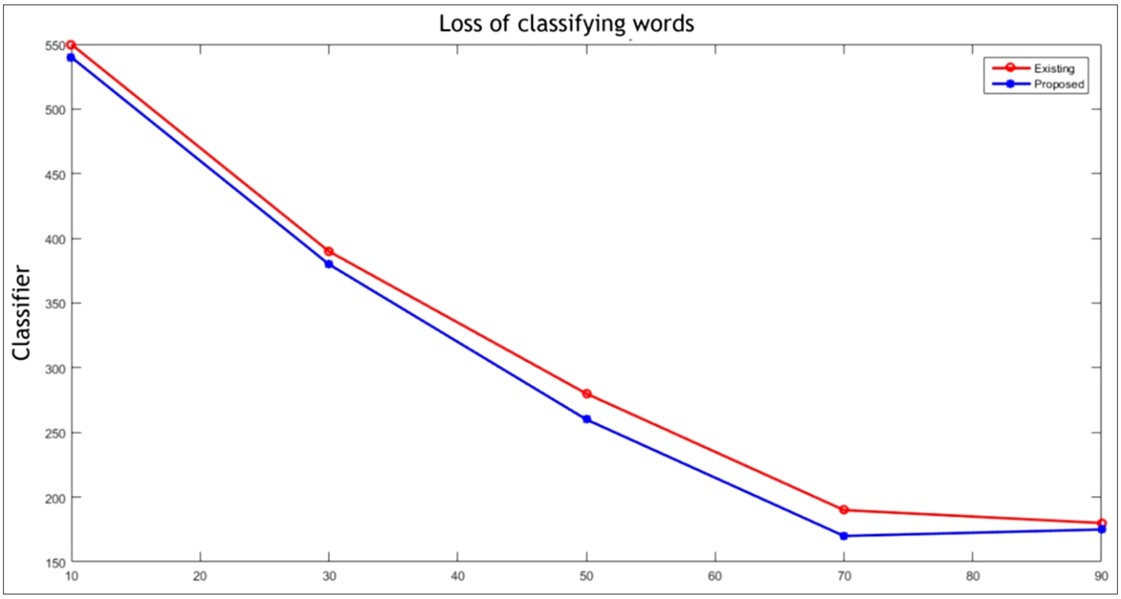

**Figure 8.** Loss percentage of the words in classification techniques

**Table 2.** CHEMDNER corpus BNER results

| | **Proposed Method** | | | **CRF** | | | **SVM** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| **ABBR.** | 90.67 | 93.39 | 92.01 | 80.33 | 82.82 | 81.55 | 74.73 | 77.85 | 76.26 |
| **Family** | 86.11 | 88.69 | 87.38 | 76.26 | 78.62 | 77.42 | 70.94 | 73.90 | 72.39 |
| **Formula** | 84.29 | 86.81 | 85.53 | 80.88 | 83.39 | 82.12 | 75.25 | 78.38 | 76.78 |
| **Identifier** | 86.42 | 89.01 | 87.69 | 83.73 | 86.32 | 85.01 | 77.89 | 81.14 | 79.48 |
| **Multiple** | 48.42 | 49.87 | 49.13 | 42.90 | 44.23 | 43.55 | 39.91 | 41.57 | 40.72 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Systematic** | 93.65 | 96.45 | 95.04 | 88.25 | 90.98 | 89.59 | 82.10 | 85.52 | 83.77 |
| **Trivial** | 92.87 | 95.65 | 94.24 | 85.09 | 87.73 | 86.39 | 79.16 | 82.46 | 80.78 |
| **Overall** | **83.20** | **85.70** | **84.43** | **76.78** | **79.15** | **77.95** | **71.43** | **74.40** | **72.88** |

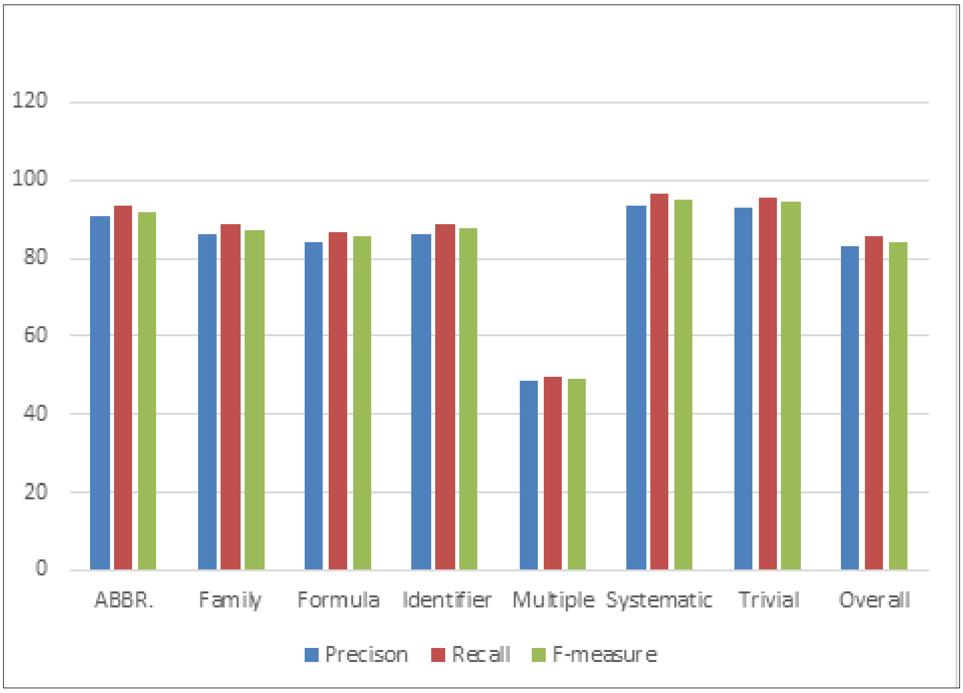

**Figure 9.** Precision Recall and F-measure for various entities in CHEMDNER corpus

It is seen from the results that with SVM classifier the recall is low and with CRF both precision and recall are moderate. The combined use of both the algorithms is effective and increases both precision and recall. This increase is also attributed to post-processing after entity extraction as it helps in resolving ambiguities and reduce false positives. In GENIA corpus there is decline in recognizing protein entities and this is due to class imbalance in the training data. Multiple entities are not recognized in CHEMDNER corpus due to coordination ellipsis.

The common issue in detecting entities is identifying the boundary of an entity mention which drastically reduces the model performance. The root of the issue is the modifier, both adjective and noun modifiers that are added to the head word which can be a part of entity mention. Hence to classify them accurately dependency relation between the word and the context word are used. The dependency context seems to provide discriminative features to identify whether the modifiers are part of the entity. The other issue is efficiency to tag unseen entities by the learned model using the context of the word.

In case of biomedical entities in GENIA corpus, in detecting entity tag for unseen word the context of the word plays a vital role. But in case of entities in CHEMDNER corpus, certain entities are formed following a pattern and the composition of word itself can be used to identify the type and hence unseen entities can be found using these discriminative features. Hence both the features are combined in this method for identifying entities across the domain. It is further observed that to classify entities in to subtypes in CHEMDNER corpus the features are not so discriminative. The trivial and family class entities have unambiguous boundaries but other types have uncertainty in determining the ending of a mention and start of a new one. SVM model used cannot fully distinguish between the subtypes based on the features as some set of entities have close resemblance to each other.

As there is a crucial necessity to identify the completeness of the entity subtypes, rule-based or dictionary based method can be incorporated for identifying subtypes in case of ambiguity. The coordination ellipsis is where one or more of the conjuncts is not complete and is missing a part of the constituent. This type of coordination ellipsis is difficult to find due to the various structure of ellipsis formed by entity mentions. In this case the proper identification of multiple class entities is deterred due to the ellipsis as shown in the result. The incorporation of dependency context does not provide discriminative feature to identify these entities. The learned model often extracts mentions that are non-elliptical with ease but to extract mentions with ellipsis long distance context of the word has to be incorporated or domain lexicons to be used.

Other type of error is from the ambiguous abbreviations and is resolved using the post-processing steps. But the error is not fully eradicated as there are abbreviations that start with non-alphabets. Hence the detection of abbreviations should also take in to account the problem and modify the model to suit the extraction of such types. Finally, the error due to missing annotation due to guidelines is also found and they can be rectified using semi-supervised models that can tag missed and evolving entities.

## 4. Conclusions

The proposed method combines models with different strengths for identifying entities and subtypes. The performance of the model is tested with varied feature sets with post processing to examine its efficiency on BNER task. The result obtained shows the efficiency of the model to extract entities from both corpuses. The model’s performance is slightly better in GENIA corpus rather than CHEMDNER corpus. This is because of the boundary detection problem in the corpus along with the classification error. The analysis of result exposed various errors in the BNER that provides venue for improvement which would benefit the task. Due to significant evolution in biomedical entities and increasing volume of literature it would be inefficient to use only supervised methods as it requires golden standard data for learning. The future work can be to adopt semi-supervised methods for BNER since it utilizes unlabelled corpus and provide generic solution which will boost the performance of the task.

### References

1.LIU, S., TANG, B., CHEN, Q., WANG, X., 2015. Drug Name Recognition: Approaches and Resources. *Information*, *6*(4), pp.790-810

2.WANG, X., YANG, C., GUAN, R., 2018. A Comparative Study for Biomedical Named Entity Recognition. *International Journal of Machine Learning and Cybernetics*, *9*(3), pp.373-382.

3.DIPPONG, T., MIHALI, C., VOSGAN, Z., AVRAM, A., BERINDE, Z., DUMUTA, A., Comparative Analysis Regarding the Chemical and Microbiological Characteristics of Some Red Wine Assortments Produced in Two Romanian Viticultural Areas. *Rev. Chim.*, **71**(1), 2020, 411-415.

4.CALINESCU, M., STOICA, C., NITA-LAZAR, M., 2019. Complex Compounds of Sm (III) with Chlorhexidine Synthesis, Characterization, Luminescent Properties and Antibacterial Activity. *Rev. Chim.*, **70** (1), 2019, 6-12.

5.SELVAKUMAR, K., SAIRAMESH, L., 2021. User Query-Based Automatic Text Summarization of Web Documents Using Ontology. In International Conference on Communication, Computing and Electronics Systems (pp. 593-599). Springer, Singapore.

6.MARRERO, M., URBANO, J., SÁNCHEZ-CUADRADO, S., MORATO, J., GÓMEZ-BERBÍS, J.M., 2013. Named Entity Recognition: Fallacies, Challenges and Opportunities. *Computer Standards & Interfaces*, *35*(5), pp.482-489.

7.WANG, Y., YU, Z., CHEN, L., CHEN, Y., LIU, Y., HU, X., JIANG, Y., 2014. Supervised Methods for Symptom Name Recognition in Free-text Clinical Records of Traditional Chinese Medicine: an Empirical Study. *Journal of Biomedical Informatics*, *47*, pp.91-104

8.CHEN, Y., LASKO, T.A., MEI, Q., DENNY, J.C., XU, H., 2015. A Study of Active Learning Methods for Named Entity Recognition in Clinical Text. *Journal of Biomedical Informatics*, *58*, pp.11-18.

9.PEREZ, A., WEEGAR, R., CASILLAS, A., GOJENOLA, K., ORONOZ, M., DALIANIS, H., 2017. Semi-supervised medical entity recognition: A study on Spanish and Swedish Clinical Corpora. *Journal of Biomedical Informatics*, *71*, pp.16-30.

10.LEAMAN, R., WEI, C.H., LU, Z., 2015. tmChem: A High Performance Approach for Chemical Named Entity Recognition and Normalization. *Journal of cheminformatics*, *7*(1), p.S3.

11.TANG, B., FENG, Y., WANG, X., WU, Y., ZHANG, Y., JIANG, M., WANG, J., XU, H., 2015. A Comparison of Conditional Random Fields and Structured Support Vector Machines for Chemical Entity Recognition in Biomedical Literature. *Journal of cheminformatics*, *7*(S1), p.S8.
12.HKIRI, E., MALLAT, S., ZRIGUI, M., 2017, November. Integrating Bilingual Named Entities Lexicon with Conditional Random Fields Model for Arabic Named Entities Recognition. In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on* (Vol. 1, pp. 609-614). IEEE.
13.HERWANDO, R., JIWANGGI, M.A., ADRIANI, M., 2017, September. Medical Entity Recognition Using Conditional Random Field (CRF). In *Big Data and Information Security (IWBIS), 2017 International Workshop on* (pp. 57-62). IEEE.
14.MIFTAHUTDINOV, Z., TROPSHA, A., TUTUBALINA, E., 2017. Identifying Disease-related Expressions in Reviews Using Conditional Random Fields.
15.ZHU, F., SHEN, B., 2012. Combined SVM-CRFs for Ciological Named Entity Recognition with Maximal Bidirectional Squeezing. *PloS one*, *7*(6), p.e39230.
16.BASALDELLA, M., FURRER, L., TASSO, C., RINALDI, F., 2017. Entity Recognition in the Biomedical Domain Using a Hybrid Approach. *Journal of Biomedical Semantics*, *8*(1), p.51
17.ZHANG, S., ELHADAD, N., 2013. Unsupervised Biomedical Named Entity Recognition: Experiments with Clinical and Biological Texts. *Journal of Biomedical Informatics*, *46*(6), pp.1088-1098.
18.EL BAZI, I., LAACHFOUBI, N., 2017. Arabic Named Entity Recognition Using Topic Modeling. *context*, *230*.
19.XU, J., GAN, L., CHENG, M., WU, Q., 2018. Unsupervised Medical Entity Recognition and Linking in Chinese Online Medical Text. *Journal of Healthcare Engineering*, *2018*.
20.QUIMBAYA, A.P., MÚNERA, A.S., RIVERA, R.A.G., RODRÍGUEZ, J.C.D., VELANDIA, O.M.M., PEÑA, A.A.G., LABBÉ, C., 2016. Named Entity Recognition over Electronic Health Records Through a Combined Dictionary-based Approach. *Procedia Computer Science*, *100*, pp.55-61.
21. KORKONTZELOS, I., PILIOURAS, D., DOWSEY, A.W., ANANIADOU, S., 2015. Boosting drug named entity recognition using an aggregate classifier. *Artificial Intelligence in Medicine*, *65*(2), pp.145-153
22.SELVAKUMAR, K., RAMESH, L.S., KANNAN, A., 2015. Enhanced K-means clustering algorithm for evolving user groups. Indian Journal of Science and Technology, 8(24), p.1-8.
23.AMBIKA, M., MANGAYARKARASI, N., GOPALSAMY, R., RAMESH, L.S., SELVAKUMAR, K., 2021. Secure and Dynamic Multi-Keyword Ranked Search. International Journal of Operations Research and Information Systems (IJORIS), 12(3), pp.1-10.
24.DALIANIS, H., 2018. Basic Building Blocks for Clinical Text Processing. In *Clinical Text Mining* (pp. 55-82). Springer, Cham.
25.DOĞAN, R.I., LEAMAN, R., LU, Z., 2014. NCBI Disease Corpus: a Resource for Disease Name Recognition and Concept Normalization. *Journal of biomedical informatics*, *47*, pp.1-10.
26.YONGYI CHEN, XIAODING YU, XUEHAO GAO, HANZHONG FENG. A New Method for Non-Linear Classify and Non-Linear Regression: Introduction to Support Vector Machine [J] (in chinese). Quarterly Journal of Applied Meteorology, 2004,15(03):345-354.

Manuscript received: 7.04.2021