A Thesaurus Based Semantic Relation Extraction for Agricultural Corpora

R. Srinivasan, C. N. Subalalitha

A Thesaurus based Semantic Relation Extraction for Agricultural Corpora *

R.Srinivasan[1][0000−0001−6192−8097] and C.N.Subalalitha[2][0000−0002−8920−707X]

[1,2]SRM Institute of Science and Technology, Kattankulathur 603203, India
srinirvs89@gmail.com
subalalitha@gmail.com

Abstract. Semantic relations exist two concepts present in the text. Semantic relation extraction becomes an essential part of building an efficient Natural Language Processing (NLP) applications such as Question Answering (QA) and Information Retrieval (IR) system. Automatic semantic relation extraction from text increases the efficiency of these systems by aiding in retrieving more accurate information to the user query. In this research work, we have proposed a framework that extracts agricultural entities and finds the semantic relation exist between entities. Entity extraction is done using a Parts Of Speech (POS) tagger, Word Suffixes and Thesaurus without using any of the external domain-specific knowledge bases, such as Ontology and WordNet. Semantic relation exists between entities are done by using Multinomial Naïve Bayes (MNB) classifier. This paper extracts two entities, namely disease and treatment and focuses on two semantic relations namely "Cure" and "Prevent". The "Cure" semantic relation expresses the remedial measure for the diseases that prevail in the crops, and the "Prevent" semantic relation shows the precautionary measures that could prevent the crop from being affected. The proposed approach has been trained with 2281 sentences and tested against 553 sentences and then evaluated using standard metrics.

Keywords: Agricultural Entity · Semantic Relation · Multinomial Naïve Bayes · Feature Extraction.

1 Introduction

Agriculture plays a vital role in the world economy [1]. Several organisations, such as Food and Agriculture Organizations (FAO), International Fund for Agricultural Development (IFAD), National Farmers' Union (NFU), International Federation of Agricultural Producers (IFAP) and many others have been working in the agriculture domain for several years to increase the food production. As a result, there has been a vast increase in agricultural data in an unstructured manner on the World Wide Web (WWW) [2]. This aids in the computational analysis of the data which can bring out fruitful solutions by the computer science

* Supported by SRM Institute of Science and Technology.

researchers, thereby making a significant contribution in solving the problems involved in food production.

Extraction of semantic relations between entities is an intuitive and vital role in Natural Language Processing (NLP) applications. NLP is a sub-domain of Artificial Intelligence which automates the text analysis to build novel applications. Several techniques are used to identify the semantic relations between entities, such as Rule-based approaches, Knowledge-based approaches, Link-based approaches, Machine Learning based approaches and Deep Learning based approaches. The semantic relations between the agricultural entities have not been explored to our knowledge, which is done by the proposed method. In the healthcare sector, the extraction of agricultural entities and semantic relation between entities is an important research topic in the field of agriculture.

The goal of the proposed work is focused on two tasks: The first task is to identify the agricultural entities namely , "disease" and "treatment" using various techniques such as POS tagger, Word Suffixes and Thesaurus. The thesaurus can be extracted from the National Agricultural Library (NAL) [3]. The second task is to identify the semantic relations namely, "Cure" and "Prevent". This proposed work can be a pointer to build agricultural domain-specific Information Retrieval (IR) systems. These types of IR systems can be a significant help for the farmers, agricultural and NLP researchers.

The main contributions of this paper are:

1) Extracting Entities: By using POS tagger, Word suffixes and Thesaurus to identify the "disease" and "treatment" entities and then the sentences are classified into positive or negative sentences. A positive sentence contains both "disease" and "treatment" entities within a sentence, else treated as a negative sentence.

2) Extraction of Semantic relation using Multinomial Naive Bayes (MNB) classifier: Once the sentences are classified, the next task is to identify the semantic relation between the sentences. The semantic relation that is addressed here, cure ,prevent and irrelevant entities. The paper has organized as follows: Section 2 discusses literature survey, Section 3 defines the proposed approach to detect and classify the meaningful sentences, Section 4 examines the evaluation results obtained, and Section 5 describes conclusions and future work.

2 Literature Survey

The author proposed a platform for extracting medical entities and identifying relationship from texts. Genia tagger has been used to identify the entities and determine semantic types using MetaMap [4]. The linguistic patterns had been used along with the help of domain knowledge to extract 16 types of medical entities. A precision of 75.72% and a recall of 60.46% is obtained using linguistic patterns.

Xin has presented a novel approach for semantic relation extraction, which combines both the pairwise relation and the link-based relation within words [5]. The pairwise and Link-based approach is used to measure the relationship

between the phrases. The author combines the approach mentioned above to generate a document clustering model. The author proposed a two-phase method which includes entity identification and relationship integration [6]. In the entity identification phase, the ML algorithms, namely Support Vector Machine (SVM), and Decision Tree (DT) are combined using statistical features. Relationship identification between the entities has been made by clustering the semantic entities. Min-Ling Zhang et al. (2009) have proposed a feature selection mechanism which incorporates Multi-Label Naïve Bayes (MNB) to improve its performance [7]. Principal Component Analysis (PCA) has been used to remove irrelevant and repeated features. Naive Bayes classifiers had been widely used for various natural language processing tasks [8].

Koichi Takeuchia et al. (2005) have used Support Vector Machine (SVM) to extract biological entities like scientific names, protein names, genes and viruses [9]. The entities are obtained from various combinations of features, such as orthographic features, context window and head noun features. Ensemble technique combined with fuzzy logic to extract the disease entities with the help of orthographic features. It provides the promising result of 94.66%, 89.12%, 84.10%, and 76.71% of F-measure for various corpora [10].

Oana Frunze et al. (2011) has used various ML methodologies that are suitable for identifying the health-care information [11]. It extracts sentences from published papers that have the mention of diseases and treatments and identifies the semantic relations between the entities. Three semantic relations, namely Cure, Prevent and Side-effect have been identified. The main idea of this work is carried out the semantic relationships in biomedical text and identifying the best Machine Learning algorithm for their dataset.

Archana Chaudhary et al. (2016) have presented a hybrid ensemble technique that combines more than one machine learning algorithm for diagnosing the oil-seed disease [12]. The oil-seed dataset is developed from different sources, which include 24 nominal attributes. It provides humidity, soil moisture, temperature and symptoms to diagnose the oil-seed disease.

As far as NLP is concerned, agricultural semantic extraction from texts has never been attempted to the best of our knowledge. In Machine Learning point of view also, agricultural domain-specific text data remain unexplored. The proposed system differs from the state of the art by attempting the semantic relation classification using MNB classifier on agricultural domain-specific documents. This proposed system might be a future pointer to develop many other useful applications for agricultural domain-specific texts using both advanced NLP and ML techniques. The next section describes the proposed methodology.

3 Proposed Methodology

3.1 System architecture

The architecture of the proposed method is shown in Figure 1. The two steps are performed in this work provide the interface for information extraction framework that is accomplished to identify and extract agricultural information. The

first task determines the meaningful sentences on diseases and treatments topics, while the second one performs a classification of sentences according to the semantic relations that exist between diseases and treatments.

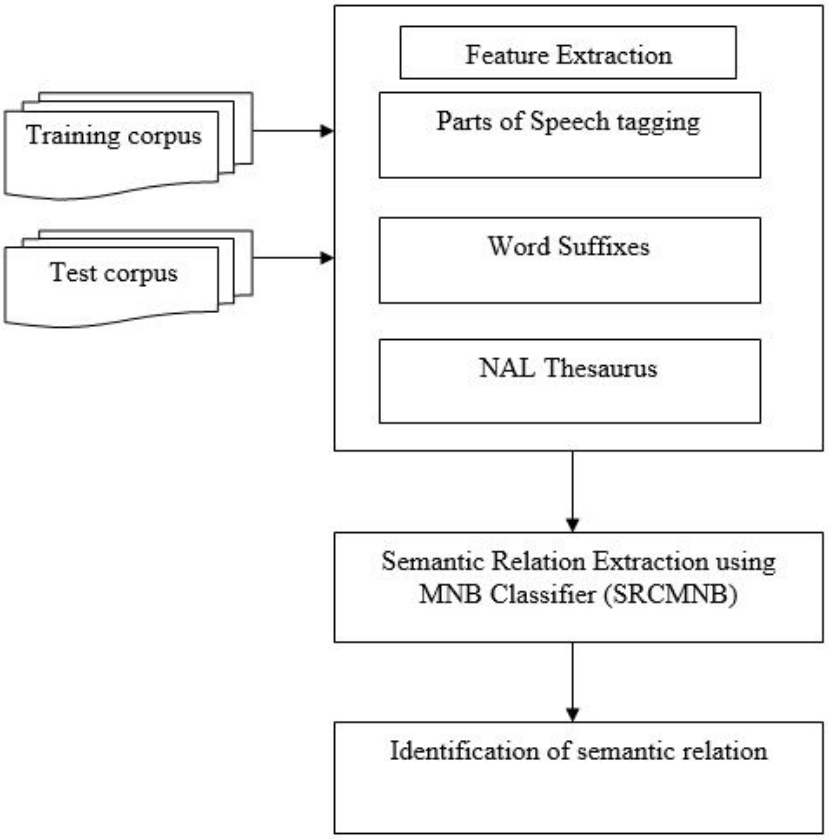

Fig. 1. Architecture for Proposed Methodology.

In the first step (Feature Extraction and Sentence Selection) agriculture sentences are collected from various online resources which include web pages, research articles and abstracts that discuss diseases and treatments. It extracts meaningful information that contains both disease and treatment information. The second step (Semantic Relation using MNB classifier) identifies semantic relations in the sentences from the meaningful information (e.g., the output of the first step). This research work focuses on two semantic relations, namely "cure" and "prevent".

3.2 Dataset details

We have created our in-house agricultural data set by web scraping, and the same is available at [13] named as agriculture. Table 1 details the in-house agricultural data and the same has been used for our research. The numbers in parentheses which describe the count of training and testing data. For example, prevent relation consists of 173 sentences, 140 is used or training set and 33 is used for testing set.

Table 1: Agriculture Dataset

Total Sentences	**Semantic Relation**	**Example**	**Disease identified**	**Treatment Identified**
Cure (701,160)	Treatment cures Disease	Wiping plants with soapy water is to heal aphids	Aphids	Soapy water
Prevent (140,33)	Treatment avoid Disease	Soluble Boron (0.01 per cent) can be done at monthly intervals to tackle ringspot in papaya	Ringspot	Soluble Boron
Irrelevant (1458,360)	No disease and treatment relation	Agriculture is the process of producing food	No disease name	No treatment name

Since it is an in-house dataset,it should be tested using inter-rater reliability tool. The inter-rater reliability is test validity tool to measure the score given by the human experts. Human experts will classify the sentences based on their meaning. In our dataset, human experts to check the sentences fall into cure category or prevent category. It is not necessary to check the irrelevant category. Irrelevant category may not contain about the agricultural entities. Cure and Prevent category contains the agricultural entities that are used to classify the sentences. The dataset consist of 2834 sentences, 1034 sentence fall into either cure category or prevent category.

Table 2: Percentage agreement across multiple annotators

Sentence	**Annotator1**	**Annotator2**	**Annotator 3**	**% agreement**
sentence 1	0	0	0	100
sentence 2	1	1	1	100
sentence 3	0	1	1	66.66

.....	.....	....	.....	....
.....	.....	....	.....	....
.....	.....	....	.....	....
sentence 1034	1	1	1	100
inter-rater reliability				89.23

Table 2 describe the percentage agreement for the cure and prevent sentences. In Table2, 0 represents cure category and 1 represents prevent category. For example, the sentence 1 all the annotator classified the sentence as 0, whereas 0 represents cure category and agreement is 100 percentage. Similarly, inter-rater reliability is calculated for all the cure and prevent category sentences.

3.3 Feature Extraction

Parts of Speech Stanford POS tagger is used in our model to extract the entities. A Part-Of-Speech Tagger (POS Tagger) is a software that reads a text and assigns parts of speech to each word such as noun, adjective, adverb, verb, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. At a first step, a sentence is passed into POS tagger and finds the parts of speech. Noun always refers to the entities and identifies the pattern for the sentence.
Example 1(Training Data)

- Soluble Boron (0.01 per cent) <treatment name> can be done at monthly intervals to tackle ringspot <disease name> in papaya
- POS of Treatment Name: Soluble <Adjective> Boron <Noun>
- POS of Disease Name: ringspot <Noun>
- Pattern identified for extracting treatment name:Adjective, Noun(feature)

In Example 1, we identified the feature for extracting treatment name. The feature is Adjective followed by a Noun is probably a treatment name. This step is used to extract the feature of treatment entities. Patterns are considered a feature by applying the method as mentioned above. Once the features are extracted, stored in two different feature set namely Disease feature set and Treatment feature set. Most frequent feature are listed in Table 3.

Table 3: Features of Disease and Treatment Entities using POS

Entity type	**Example**	**Feature**
Disease	Aphids	Noun
Disease	Late Blight	Noun, Noun
Treatment	Liquid copper	Adjective, Noun
Treatment	Wiping plants with soapy water	Verb, Noun, Preposition, Adjective, Noun

Word Suffixes Word suffixes are a group of characters added to the ending or in between the texts. In our paper, word suffixes play an important role to extract disease or treatment entities. The first step of feature extraction is done by POS tagger. It extracts the patterns from the tagged entity that helps to identify the entity in the test data. But not all the patterns which obtain the proper disease and treatment names. The output of the first step is passed into the second step. The second step verifies the entities is disease name or treatment name. Word suffix identifies the minimal amount of disease and treatment entities. Some of the disease names follow the common word ending suffixes and are listed in Table 4.

Table 4: Features of Disease and Treatment Entities using Word Suffixes

Entity type	**Example**	**Suffix**	**Feature**
Disease	Leaf Spot	Spot	Spot
Disease	Late Blight	Blight	Blight
Treatment	Psychodynamic Therapy	Therapy	Therapy
Treatment	Overdiagnosis	Diagnosis	Diagnosis

NAL Thesaurus The next step is done by analysing the definitions given in the NAL thesaurus that contains the seed words and meanings. The analysis done by NAL is a collection of descriptions of agricultural entities developed in conjunction with the formation of the National Agricultural Library(NAL) Thesaurus. NAL thesaurus Contains over 255,390 terms, including 148,282 descriptors in English and Spanish. The 2017 edition contains 5,223 definitions, and the NAL Thesaurus Staff composes it. Features are extracted from POS Tagging and Word Suffixes. Entities are passed through the NAL Thesaurus. NAL Thesaurus verifies that the entity is present or not.

Let S = w1, w2,. . . . . , wn where S is a sentence and w represents the words which may contain the definition of a disease or treatment name is a context word in NAL thesaurus. If a sentence carries both the entities, the sentence is labelled as a positive sentence. If a sentence not containing any of the information or non-relevant information, labelled as a negative sentence.

3.4 Semantic Relation Classifier based on Multinomial Naïve Bayes (MNB)

MNB is a supervised Machine Learning model widely used as a probabilistic classifier model for classifying the text. MNB classifiers are widely used for text classification process [14]. MNB classifier is based on Bayes' theorem with strong independent attribute assumption [15]. MNB classifier assumes not only the independent attributes but also finds the frequency of the all attribute belong to the same class. Frequency count of an attribute is important that decides the

probability of an attribute falls into class [16]. Once it classifies the positive, the next step is to identify the semantic relation occurs in a sentence by using MNB. Rosario and Hearst described the nine semantic relations in bio-science text [17]. The proposed work describes the two semantic relations such as "cure" and "prevent" in the agricultural texts. By applying MNB classifier to extract the semantic relations from the identified positive sentence.

function TRAINMNB(D,C)
For each class c belongs to C
Calculate prior[c]
Ns = total number of positive sentences in D
Nc = total number of positive sentences from D belongs to class c

$$prior[c] = \frac{N_c}{N_s} \tag{1}$$

Calculate conditionalprobabilty[w,c]
W ← All the words of positive sentences in D
tot[c] ← append(S) for S belongs to D with class c
u ← number of unique words of D
For each word w in W
count(w,c) ← no of occurrences of w in tot[c]

$$conditionalprobabilty[w, c] = \frac{count(w, c) + 1}{\sum_{w^t inW} count(w^i, c) + u} \tag{2}$$

return W, prior, conditionalprobability

function TESTMNB(testdata, W, prior, conditionalprobability, C)
for each class c ∈ C
sum[c] ← prior[c]
for each position of j in testdata
word ← testdata[j]
if word ∈ W
sum[c] ← sum[c] * conditionalprobabilty[word,c]
return argmax_c sum[c]

4 Results and Discussion

Accuracy, Recall, Precision and F-score metrics have been used to evaluate the proposed system[18]. A confusion matrix is developed before evaluating the classification model[19]. Table 5 describes the confusion matrix used in the proposed system.

Table 5: Confusion Matrix

True Positive (TP)	False Positive (FP)
False Negative (FN)	True Negative (TN)

In Table 5, TP denotes no of sentences that are correctly predicted as positive. FP means no of sentences misclassified as positive. FN denotes no of sentences misclassified as negative. TN denotes no of sentences that are correctly predicted as negative. Accuracy is the ratio of the number of correctly predicted sentences to the total sentences. Accuracy is calculated by using the following equation

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

Recall defined the ratio of the correctly predicted relevant sentences to the overall observation in the relevant sentences. The following equation is used to calculate the recall.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

Precision is defined as the ratio of correctly predicted relevant sentences to the total predicted relevant sentences and can be calculated as follows.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

F-score is the weighted harmonic mean between precision and recall.

$$F - score = \frac{2 * recall * precision}{recall + precision} \tag{6}$$

4.1 Evaluation and Performance Analysis of the sentence selection module

Table 6 describes the dataset used for sentence selection. Sentences which contain information about both "disease" and "treatment" are labelled as positive sentence. As mentioned above in section 3.2, sentences which contain "disease" or "treatment" information or sentences containing neither of this information are considered as negative sentence.

Table 6: Dataset used for Sentence Selection

Sentence	**Positive**	**Negative**
Training	1141	1458
Testing	193	360

POS tagger extracts 139 features from the training sentences. Word suffixes extract 30 features from the training data. By combining the features extracted from the various techniques and mapped with the NAL Thesaurus. POS taggers are not sufficient for extracting disease and treatment entities. Table 4 "Aphids"

is a word refers to a noun. But all the nouns are not disease or treatment entities. Precision value is low by applying POS tagger alone. Entities retrieved by the system are more, but actual disease and treatment entities are less. The next step is to combine POS tagger with word suffixes. Word suffixes are to verify the entities are disease or treatment. All the disease entities are not terminated with word suffixes. For example, an accelerometer is word occurs in test data. By applying POS tagger, an accelerometer is a noun and checks the word suffixes. Word suffixes having "cele" are considered as a feature."cele" means swelling that considered as a feature. An accelerometer falls into this positive category, but it does not contain a disease or treatment entities.

The next step is to apply NAL Thesaurus to verify the entity is correct or not. By applying POS tagger, word suffixes and thesaurus gives good accuracy. Table 7 describes the evaluation metrics for sentence selection. After applying the POS, Word Suffixes, and Thesaurus to the test data, 144 sentences are predicted as positive sentences , 266 sentences are identified as negative sentences, and the remaining sentences are misclassified. Accuracy is less due to entities are not repetitive in nature. Most of the entities are unique in our dataset. Due to this reason, accuracy falls down to 74.14%. By applying equation 3 through 6, the values of Recall, Precision, F-score and Accuracy are calculated.

Table 7: Evaluation Metrics for Sentence Selection

Method	**Recall**	**Precision**	**F-score**	**Accuracy**
POS + Word Suffixes + Thesaurus	60.50	74.61	66.73	74.14

4.2 Output for Semantic Relation

The second step is to identify the sentence that contains a cure, prevent or irrelevant semantic relation. The second step is a three way classification technique to identify the semantic relation from the first step. In Table 8, which describes Recall, Precision, F-Score value for each classification algorithm. The output of the table 8 depends on the output of the previous table. In Table 7, the overall accuracy is obtained as 71.80%. 410 sentences are correctly classified as true positive and true negative in the first step.

Table 8: Performance Comparison for Semantic Relation(first step followed by second step)

Classification algorithm	**Disease** **R/P/F-score**	**Treatment** **R/P/F-score**	**Other** **R/P/F-score**
MNB	83.01/96.95/89.44	92.31/80.00/85.72	96.97.97/89.96/93.33
SVM	90.83/83.9/87.23	100/61.54/76.19	90.53/97.00/93.65
Random Forest	85.09/82.20/83.62	93.33/42.42/58.33	89.32/96.91/92.96

Table 8, describes the Recall, Precision, F-score value of all category. For example, "immune" is the name of the disease as well as name of the treatment. The word "immune" followed by "disorder", then it is a disease entity. The word "immune" followed by "therapy" or "inhibitors", then it is a treatment entity. Based on the meaning or, it is considered to be a disease entity or treatment entity. After applying Support Vector Machine (SVM), Random Forest, Multinomial Naive Bayes(MNB), MNB outperforms the best result. It calculates the probability value of each and every word and find out the overall probability of the sentences. Due to word count of each word MNB obtains the best F-score value in the semantic relation.

Table 9: Performance Comparison for Semantic Relation(second step independent of first step)

Classification algorithm	**Disease** **R/P/F-score**	**Treatment** **R/P/F-score**	**Other** **R/P/F-score**
MNB	81.31/89.94/85.41	76.92/66.67/71.43	95.97/91.99/93.94
SVM	88.13/81.50/84.69	84.62/55/66.67	89.35/96.09/92.6
Random Forest	89.71/62.24/73.49	90.48/52.78/66.67	79.23/96.76/87.12

Table 9 describes the Recall, precision, F-score value of second step which is independent of the first step. Naïve Bayes algorithm outperforms the best result for this dataset in all aspects. The first task is followed by the second task which produces the F-score value of 93.94%. In other case, second task is independent of the first task and produces the F-score value of 93.33%.

5 Conclusion

This paper presented a novel approach to classify agricultural documents based on Multinomial Naïve Bayes classifier. The first task identifies the disease and treatment relation with the help of POS Tagger, Word Suffixes and Thesaurus. The second task is to identify the semantic relationships between text present in a sentence with the help of MNB method. The study focused on two semantic relations between disease treatment relation in the agricultural text. The accuracy of 71.9% and F-score of 71.43% in the semantic relation are achieved. The performance of an MNB approach shows a better accuracy for the classification process. In future, we are researching how to identify the other types of agricultural relations that relate the agricultural entities in text. Future work can overcome ambiguous entities using different techniques and tries to improve accuracy. This kind of semantic relation extraction will lead to building useful IR Systems and QA systems like agricultural chatbots. Also, various ML algorithms can be tried to achieve better accuracy of the dataset.

A Appendix

The term NER was introduced in 1996 at the Message Understanding Conference to refer the entities [20]. NER is defined as to identify and classify the information elements called Named Entities [21]. Biomedical Named Entity Recognition (BNER) is used to determine the biological entities such as protein names, genes, disease name in biomedical texts [9] [22]. Please note that the first paragraph of a section or subsection is not indented. The first paragraph that follows a table, figure, equation etc. does not need an indent, either.

References

1. Alston, Julian and Pardey, Philip.: Agriculture in the Global Economy. The Journal of Economic Perspectives **28**, (2002)
2. Sander J.C. Janssen and Cheryl H. Porter and Andrew D. Moore and Ioannis N. Athanasiadis and Ian Foster and James W. Jones and John M. Antle: Towards a new generation of agricultural system data, models and knowledge products: Information and communication technology. Agricultural Systems **155**, 200–212 (2017)
3. National Agricultural Library - Thesaurus, https://agclass.nal.usda.gov/download.shtml. Last accessed 4 Feb 2019
4. Ben Abacha, Asma and Zweigenbaum, Pierre.: Automatic extraction of semantic relations between medical entities: A rule based approach. Journal of biomedical semantics **2**(5), S4 (2011)
5. Xin Cheng and Duoqian Miao and Can Wang.: A link-based approach to semantic relation analysis. Neurocomputing **154**, 127–138 (2015)
6. Dingxian Wang and Xiao Liu and Hangzai Luo and Jianping Fan.: A novel framework for semantic entity identification and relationship integration in large scale text data. Future Generation Computer Systems **64**, 198–210 (2016)
7. Min-Ling Zhang and José M. Peña and Victor Robles.: Feature selection for multi-label naive Bayes classification. Information Sciences **179**(19), 3218–3229 (2009)
8. Altheneyan, Alaa Saleh and Menai, Mohamed El Bachir.: NaïVe Bayes Classifiers for Authorship Attribution of Arabic Texts. J. King Saud Univ. Comput. Inf. Sci. **26**(4), 473–484 (2014)
9. Koichi Takeuchi and Nigel Collier.: Bio-medical entity extraction using support vector machines. Artificial Intelligence in Medicine **33**(2), 125–137 (2005)
10. Balu Bhasuran and Gurusamy Murugesan and Sabenabanu Abdulkadhar and Jeyakumar Natarajan.: Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. Journal of Biomedical Informatics **64**, 1–9 (2016)
11. O. Frunza and D. Inkpen and T. Tran.: A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts. IEEE Transactions on Knowledge and Data Engineering **23**(6), 801–814 (2011)
12. Archana Chaudhary and Savita Kolhe and Raj Kamal.: A hybrid ensemble for classification in multiclass datasets: An application to oilseed disease dataset. Computers and Electronics in Agriculture **124**, 65–72 (2016)
13. Agricultural dataset, https://drive.google.com/file/d/1b1TfA25dqXFxdH6U9eW2MP2S12ae8IaI/view?usp=sharing Last accessed 4 Feb 2019
14. Zhang, Wei and Gao, Feng.: An Improvement to Naive Bayes for Text Classification. Procedia Engineering **15**, 2160–2164 (2011)

15. John, George H. and Langley, Pat.: Estimating Continuous Distributions in Bayesian Classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338–345. UAI'95, Location (1995)
16. Bermejo, Pablo and Gámez, José and Puerta, Jose.: Improving the performance of Naive Bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets. Journal **38**, 2072–2080 (2011)
17. Rosario, Barbara and Hearst, Marti.: Classifying Semantic Relations in Bioscience Texts. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pp. 430–437. Barcelona, Spain (2004)
18. Powers, David.: Evaluation: From precision, recall and fmeasure to roc, informedness, markedness and correlation. Journal of Machine Learning Technologies **2**, 37–63 (2007)
19. Zhang, Wei and Gao, Feng.: An Improvement to Naive Bayes for Text Classification. Procedia Engineering **15**, 2160–2164 (2011)
20. Chinchor, Nancy A..: Overview of MUC-7. In: Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia (1998)
21. Mónica Marrero and Julián Urbano and Sonia Sánchez-Cuadrado and Jorge Morato and Juan Miguel Gómez-Berbís.: Named Entity Recognition: Fallacies, challenges and opportunities. Computer Standards & Interfaces **35**(5), 482–489 (2013)
22. Mourad Gridach.: Character-level neural network for biomedical named entity recognition. Journal of Biomedical Informatics **70**, 85–91 (2017)