iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Comparative Analysis of Supervised Approaches for Word Sense Disambiguation Using Text Similarity

Subha Mahajan[1], Rakesh Kumar[2], Vibhakar Mansotra[3]
Department of Computer Science and IT, University of Jammu, Jammu

Abstract: ***The words that are often being correspond to two or more meanings rather than to a single meaning results in semantically-ambiguous words . Measuring the similarity between words, sentences , paragraphs is an important part in information retrieval and word sense disambiguation tasks. One of the biggest challenges in Natural Language Processing is for the system to encompass in what sense a specific word is being used .This paper describes the analysis of text in order to a certain first the similarity in case that exists. Second the effort has been made to resolve the ambiguity in the text. The paper presents the comparison of machine learning approaches in the text similarity analysis. The Naive bayes approach was observed to outperform other approaches including SVM , Max Entropy , Tree , Random Forest and Bagging . Keywords: Text Similarity, Word Sense Disambiguation, Approaches, SENSEVAL, Supervised machine learning algorithms.***

I. INTRODUCTION

In the present world people are mainly depended on the web for searching any kind of content. Search engines have done remarkable job of information retrieval. However, but still the goal of retrieving relevant information is a far cry . When the person is searching information on web he /she does not bother about the ambiguity of a word that whether the content they are retrieving is relevant to them or not. It gets difficult for the user to get relevant information in any language when the word or phrases have more than one interpretation. One step towards realizing this goal is the detection of similarity of texts i.e. determining how close is the meaning of two given texts are . The idea is based on text similarity [1] detection which plays an important role in text related search in tasks such as information retrieval, word sense disambiguation (WSD) , machine translation , Information Extraction and Speech Recognition and others. For example , the phrase “The second hand of the clock is not working “, the word second means a basic unit of time , while in phrase “Ram came second in the class” , the word “second” refers to the position in series .The problem can be reduced up to an extent by the concept of disambiguation of a word. When a word has multiple meaning then it is probably considered an ambiguity. Hence, Word sense disambiguation (WSD) is termed as an open problem of natural language processing with a process of identifying a correct sense of a word in a given context. WSD plays important role in improving the quality of information so as to comprehend in what sense a specific word is being used. WSD was first formulated as a distinct ciphering task during early days of machine translation in late 1940s, making one of the oldest problem of computational semantics. The problem was continued as a challenging task until there was a availability of resources. In 1980 there was prodigious development in the area of WSD research when a large scale lexical resources and corpora came into existence. In 1990s , NLP provided three major developments for WSD :online dictionary WordNet which is organised as a word senses called synsets and used as an online sense inventory ,statistical methodologies which are used as sense classification problems and SENSEVAL which was proposed in 1997 by Resnik and Yarowsky. Further other SENSEVAL evaluation exercises have also been introduced so that researchers can share and upgrade their views in this research area.

II. LITERATURE REVIEW

When the work started on handling of the different languages with automatic means, the problem of ambiguity drew the interest of the researchers at the same time. Work on ambiguity in sense annotation has often focused on techniques to reduce ambiguity in sense inventory .Therefore, we can say that the WSD task is one of the oldest tasks for solving lexical ambiguity . Many of the researchers[2]Mukti Desai and Mrs. Kiran Bhowmick (2013) have surveyed on solving the ambiguity by applying different approaches and techniques of WSD. [3] A. R. Rezapour et al. (2011)have used a K-Nearest Neighbor algorithm of supervised learning method for WSD. The author have done feature extraction which includes the set of words that have occurred frequently in the text and the set of words surrounding the ambiguous word, so as to improve the classification accuracy. [4]Arti Mishra and Meenakshi Pathak (2014) have analyzed the web queries in English language to study the effect on the performance of various

search engines .[5] Rekha Jain et al. (2013)have proposed a Dynamic Page Rank algorithm for resolving the ambiguities and also arranging the results according to users need . [6] Gurinder Pal Singh Gosal(2015) have also applied Naive Bayes algorithm, which is the part of a supervised learning techniques on WSD task to disambiguate the senses of different words from available corpora in the SENSEVAL (WSD) in order to observe that the senses of ambiguous word having lesser number of part-of-speeches are disambiguated more correctly. [7] F.B. DianPaskalis and M.L. Khodra(2011) have also implemented WSD on Natural Language Processing (NLP) tool called Lucene using query expansion and relevance feedback using WordNet in order to provide better understanding in the performance of information retrieval system.[8]Neetu Sharma and S. Niranjan(2014) have also attempted to optimize the word sense disambiguation method using a Combinatorial approach of supervised and unsupervised learning algorithm . They have combined Naïve Bayesian of supervised learning algorithm with K-means Clustering of unsupervised learning algorithm with WordNet as database in order to enhance the performance or getting best sense of ambiguous word. [9] Khaled Abdalgader(2016) have incorporated word sense disambiguation technique with in text similarity measure and evaluates the resulting method on the benchmark Microsoft Research Paraphrase Corpus along with wordnet which leads to a significant improvement in performance.[10] Neha Kumari and Sukhbir Kaur(2016) have surveyed different research papers in order to find various methods to calculate the similarity between two sentences to determine whether the sentences are semantically equivalent or not.

III.APPROACHES TO WSD

There are different approaches [11] which are applied to the task of WSD, such as, knowledge-based methods, supervised methods, unsupervised methods to name a few.

A. Knowledge based approach

The Knowledge based approach is based on different types of resources like WordNet, Semcor, SENSEVAL, dictionary or thesaurus etc which are used to get the appropriate sense of the word. The approach provides better wider range of knowledge base which is also known as lexical knowledge base in NLP (Natural Language Processing).These resources help us to provide a sense meaning called gloss and helps in finding the relationship between the word present in the gloss and the word which needs to be disambiguated i.e. WordNet. SENSEVAL is a English Lexical words which is based on HECTOR database and the word which needs to be disambiguated is preceded by a unique identifier (tag sense number).Semcor is another LKB which helps in marking the sense of a word. The approach provides the limitation of Dictionary for sense of target word and also does not provide enough material to build classifier.

B. Supervised Approach

In supervised learning, the system is provided with set of data which is labelled into set of categories and involves learning a function which maps the data into categories. The approach is based on Machine learning techniques to set a classifier from manually or automated sense-annotated datasets .The classification task for assigning correct sense to each word is done by classifiers. The approach consists of two different phases: training phase and testing phase. In training phase a sense annotated training corpus is for the extraction of semantic and syntactic features in order to build a classifier with a help of various machine learning techniques. In the testing phase, the classifier tries to find out the correct sense for the word based on neighbouring words present in the sentence.

C. Unsupervised Approach

This approach use manually created lexical resource instead of sense tagged corpora .This approach is based on the assumption of similarity between the words so as to form clusters . Based on the concept that meaning of a particular word will depend on their neighbouring words. Unsupervised approach partitions the instances of a particular word into number of classes in order to decide whether the instances of word have same sense or not. The objective of this approach is to identity sense clusters . Methods involved are Context clustering which is based on clustering technique in which the context vectors build are grouped into clusters to identify the sense of the word .The method use the concept of vector space as word space and its dimensions as words so as to find the co-occurrence of word (not target) with the target word and then the centroid is calculated of the vector of words occurring in the same context .Another method is word clustering in which words are clustered according to the words which are having similar meaning i.e semantic similarity based on single feature. Also the similarity between the words is given by syntactical dependency. Another method is called graph based method in which a graph is built on some grammatical relationship, in graph weights are assign to the

edge according to the relatedness. An iterative algorithm is applied to get the word with highest degree node and finally minimum spanning tree is applied to disambiguate instance of target word.

D. Semi-supervised Approach

Semi-supervised learning is another machine learning technique which make the use of both type of data, that is, labelled data and unlabelled data, this is because of the problem of lack of training data. Mainly in this approach less amount of labelled data is used as compared to unlabelled data. Bootstrapping algorithm is commonly used semi-supervised learning method for WSD. It works by iteratively classifying unlabelled examples and adding confidently classified examples into labelled dataset using a model learned from augmented labelled dataset in previous iteration. Recent research show about graph based semi-supervised learning algorithms are introduced, which can effectively combine unlabelled data with labelled in learning process by exploiting cluster structure in data. Label propagation algorithm is a graph based semi-supervised learning algorithm (LP algorithm) for WSD.

IV. PROPOSED WORK AND EXPERIMENTAL SETUP

The proposed method involves two phases . In first phase text dissimilarity analysis is done to separate total out of context text from the rest using supervised approaches of Word sense disambiguation based on Naive bayes , SVM , MaxEntropy , Tree , Random Forest , Bagging .In second phase the sense ambiguity analysis is carried out among the similar text using decision list (sense Inventory) and the probability score generated from first phase. The online dataset semantic evaluation which is provided by susssex is used for semantic analysis systems . The evaluation is advanced from SENSEVAL word sense evaluation series .We selected the dataset of SENSEVAL 1 which was endowed in 1998 by sussex .The dataset provides different English-language materials for Word Sense Disambiguation (WSD) evaluation exercise [12]. The dataset consists of 35 words along with dictionary as sense inventory , the words are mapped with WordNet [13] which provides distribution according to the part of speech . It also includes the training, testing and gold standard corpora, we used the standard corpora in which each sense had a numerical unique identifier, as well as a mnemonic. The WordNet is a lexical database which consists of lexical category and consisting of support dictionaries which are used in the analysis and following observation were made.

V. RESULTS AND DISCUSSION

The similarity analysis of the text was carried out using probabilities models involving Naive Byes, SVM, MaxEntropy, Random Forests, Tree and Bagging. The formulas used against each Machine learning algorithm is shown in Table 1.The input parameters during training the model involved 20 text belong to class ‘A ‘ (accident) and other 20 text to class ‘B’ (promise).The test inputs consists of 15 text from class ‘A’ and ‘5’ from class ‘B’ as shown in Table 2. The output observations of each model is tabulated in the Table 3.showing the probability against each text statement, indicating first the similarity between two classes, second the sense of each based on the probability used as correlation score. Following, we discuss general observations of the instances classifications and then describe the results of each analysis. We conclude with a discussion of the how current sense annotation approaches can be improved. The analysis of text was done on Naive Bayes , MaxEntropy , SVM, Tree , Forests, Bagging and following observations were based on the probabilistic models used.

1. The statement 2 “Our care and supervision does not absolve you from responsibility” has somewhat an inclination away from the context term accident as mentioned in Table 2.
2. The tag/ term responsibility is being used in statement 2. Similarly in statement 7 the “tag /term responsibility” is used and in statement 15 and 19 accident has not happened as “tag /term caution”, promise is not used as” tag/term caution” respectively.

TABLE I: Table 1: Showing the formulae of Naive Bayes , MaxEntropy , SVM , Decision Tree , Random Forests, Bagging algorithm.

Sno.	Algorithm	Formulae
1	Naive bayes	$S^{\wedge}=\text{argmax}_{\ s\ E\ senses}\ pr(s\|v_w)$
2	Svm	$[\ \frac{1}{n}\sum_{i=1}^{n} max\ (0,1 - y_i\ (w.x_i\ + b))\ ] + \lambda\| w\|^2$
3	MaxEntropy	$p^* = \text{argmax}\ [\ -\ \sum_{x,y} p\ ^{\sim}\ (x)\ p\ (y\|x)\ \log\ p(y\|x)\ ]$
4	Tree	$\text{argmax}\ [\ i\ (t_p)- P_l\ i(t_l)- P_r\ i\ (t_r)]$

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

5	Forest	$f^{\wedge} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$
6	Bagging	$f^{\wedge} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$

TABLE II
Table 2: Showing input text taken from SENSEVAL dataset.

Sno.	Text	Code	Context	Class
1	The accident happened at 7.45am yesterday near the Evenlode Hotel, outside Eynsham	a1	accident	A
2	Our care and supervision does not absolve you from responsibility for rigging and inspecting the equipment you use and controlling your risk of accident or injury while using it	a2	accident	A
3	Mr Olejnik said he was lucky that the road conditions were fairly dry at the time because otherwise he could have skidded and had a more serious accident	a1	accident	A
4	The electoral register could show four people living in a household and the community charge form only three they may be trying to deceive or their son could have been killed in a motorcycle accident	a1	accident	A
5	A mans been taken to Oxfords John Radcliffe hospital following an industrial accident at Bicester in which his arm was badly injured in machinery	a1	accident	A
6	To be honest I always thought that I would either win this race or there would be an accident	a1	accident	A
7	Coroner Mr Nicholas Gardiner said someone usually supervised tipping trucks but not on the day of the accident	a1	accident	A
8	The judge said the accident had transformed Mrs Kelly from a lively young woman into a bed-ridden invalid	a1	accident	A
9	Mystery surrounding the accident inquiry has deepened because the company had no documentary evidence about how the ship was constructed according to a leaked memorandum dated July 1 1985	a2	accident	A
10	I heard about this awful accident on the radio and Im worried sick	a1	accident	A
11	In the accident and emergency department Anthony Cross the consultant in charge was ventilating the lungs of a man in his 60s who had taken an overdose	a2	accident	A
12	But he stressed he couldn't go so far as to say they played no part in the accident	a2	accident	A
13	The stockbroker from across the street I am told he is dead in a car accident	a2	accident	A
14	We cannot say whether this blood was the result of a shooting incident or because of an accident a spokesman for the prosecutor said last night	a2	accident	A
15	Don't forget that accidents can happen to you too	a3	accident	A
16	The Workers Education Authority based in Brewers, 11Street Oxford has promised to send tutors to the centre for six weeks, starting tomorrow	a4	promise	B
17	The council has also promised not to be a party to any tobacco advertising	a4	promise	B
18	In light of the situation on both sides of the Taiwan Strait, it recently announced the prospect for the recovery of the mainland is high and promising	a5	promise	B
19	Their scorer was the promising Ben Whitehall	a5	promise	B
20	At this stage it is not clear whether Ganshas words constitute a threat or a promise	a4	promise	B

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

TABLE III

Table 3: Showing the probability of the input text using Naive Byaes , MaxEntropy , SVM , Decision Tree , Random Forests , Bagging.

S no.	CLASS	MAXENTROPY	SVM	FORESTS	BAGGING	TREE	NAIVEBAYES
1	1	0.990663	0.895099	0.78	1	0.875	0.99998
2	1	0.77439	0.82563	0.695	1	1	0.95238095
3	1	0.999679	0.966119	0.775	0.96	1	0.95238095
4	1	0.999035	0.930903	0.81	0.96	1	0.99998
5	1	0.999857	0.921135	0.835	1	0.875	0.99998
6	1	0.938784	0.817318	0.705	1	0.875	0.99998
7	1	0.648829	0.713442	0.63	1	0.875	0.99998
8	1	0.982876	0.931207	0.745	1	0.875	0.95238095
9	1	0.964521	0.917285	0.735	1	0.875	0.95238095
10	1	0.978216	0.843197	0.79	1	1	0.99998
11	1	0.997497	0.954403	0.8	1	1	0.99998
12	1	0.97898	0.831937	0.73	1	0.875	0.99998
13	1	0.992258	0.901081	0.77	1	0.875	0.95238095
14	1	0.991534	0.926833	0.785	1	0.875	0.95238095
15	2	0.698754	0.796566	0.625	0.8	0.8	0.95238095
16	2	0.999212	0.940951	0.8	1	1	2.00E-05
17	2	0.999339	0.93405	0.81	1	1	2.00E-05
18	2	0.995833	0.844266	0.64	0.92	1	0.04761905
19	2	0.757897	0.771766	0.65	0.92	1	2.00E-05
20	2	0.98862	0.911602	0.745	1	1	0.04761905

TABLE IV

Table 4: Showing the text context sense against the corresponding code used in the analysis

Sno.	Description	Code
1	Crash	a1
2	Crashmod	a2
3	Happen	a3
4	Vow	a4
19	Ingadj	a5

The text is further classified under the different sense against each tag word based on the probability score as shown in Table 5. For

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

the text with probability score >=0.77 supposed to have code a1 and text with probability score <= 0.77 belong to sense code a2 in class 'A' and >= 0.75 belong to sense code a4 and <= 0.75 belong to sense code a5 in class 'B'.

TABLE V

Table 5: Showing probability score against each text class along with corresponding sense code as mentioned in Table IV

Sno.	Class	Pred_Class	MAXENTROPY	Code
1	A	Sb	0.990663	a1
2	A	Sb	0.77439	a2
3	A	Sb	0.999679	a1
4	A	Sb	0.999035	a1
5	A	Sb	0.999857	a1
6	A	Sb	0.938784	a1
7	A	Sa	0.648829	a1
8	A	Sb	0.982876	a1
9	A	Sb	0.964521	a2
10	A	Sb	0.978216	a1
11	A	Sb	0.997497	a2
12	A	Sb	0.97898	a2
13	A	Sb	0.992258	a2
14	A	Sb	0.991534	a2
15	A	Sa	0.698754	a3
16	B	Sb	0.999212	a4
17	B	Sb	0.999339	a4
18	B	Sb	0.995833	a5
19	B	Sa	0.757897	a5
20	B	Sb	0.98862	a4

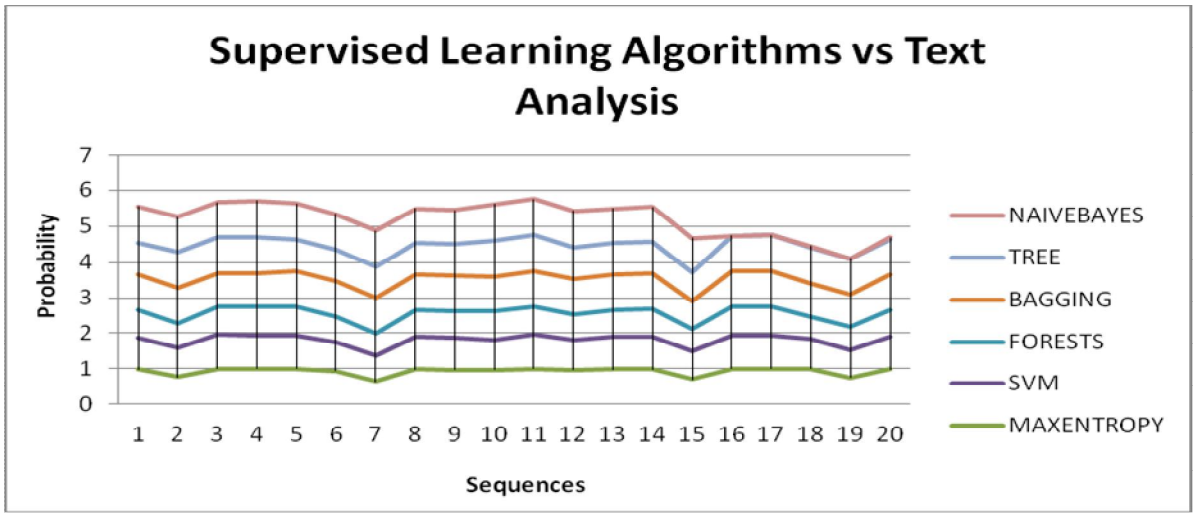


Figure1. Showing Performance Evaluation of machine learning algorithms including Naive Bayes, MaxEntropy, SVM, Decision Tree, Bagging, Random Forests used in text similarity and sense disambiguity analysis.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653

The text 2 , text 7, although defined under the tag/term 'accident' have inclination little away from the tag /term. Context /sense accidents is being used in the training model . Therefore the probability of the mentioned terms is <0.9 as predicted by the algorithms Max_Entropy , SVM , Bagging ,Forests and Tree.

The text 15, text 19 ,although defined under the tag/term 'promise' have inclination little away from the tag /term .Context /sense promise is being used in the training model . Therefore the probability of the mentioned terms is <0.8 as predicted by the algorithms Max_Entropy , SVM , Bagging ,Forests and Tree as shown in Figure1.

TABLE VI

Table 6: Showing the Accuracy of models used.

Sno.	Algorithm	Accuracy
1	NaiveBayes	100%
2	MaxEntropy	98%
3	SVM	92%
4	Tree	88%
5	Forests	85%
6	Bagging	95%

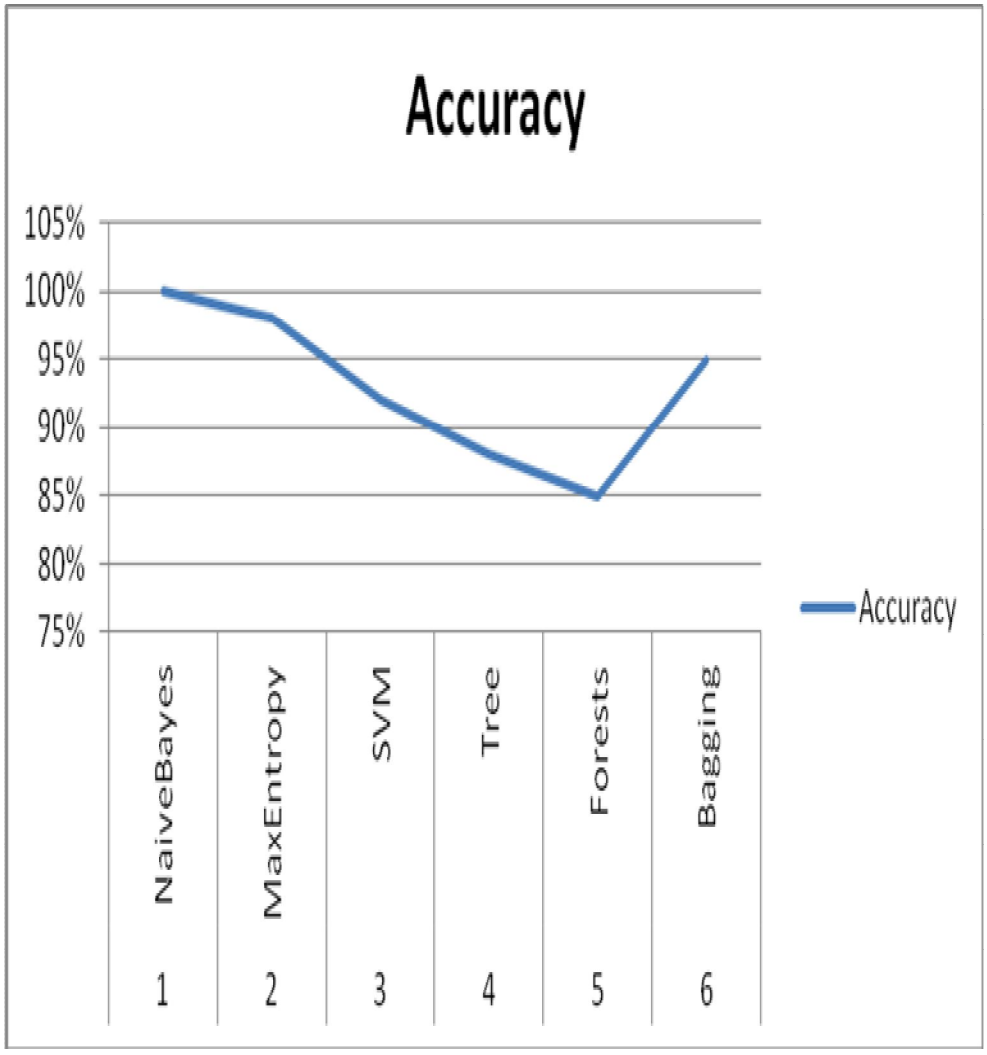


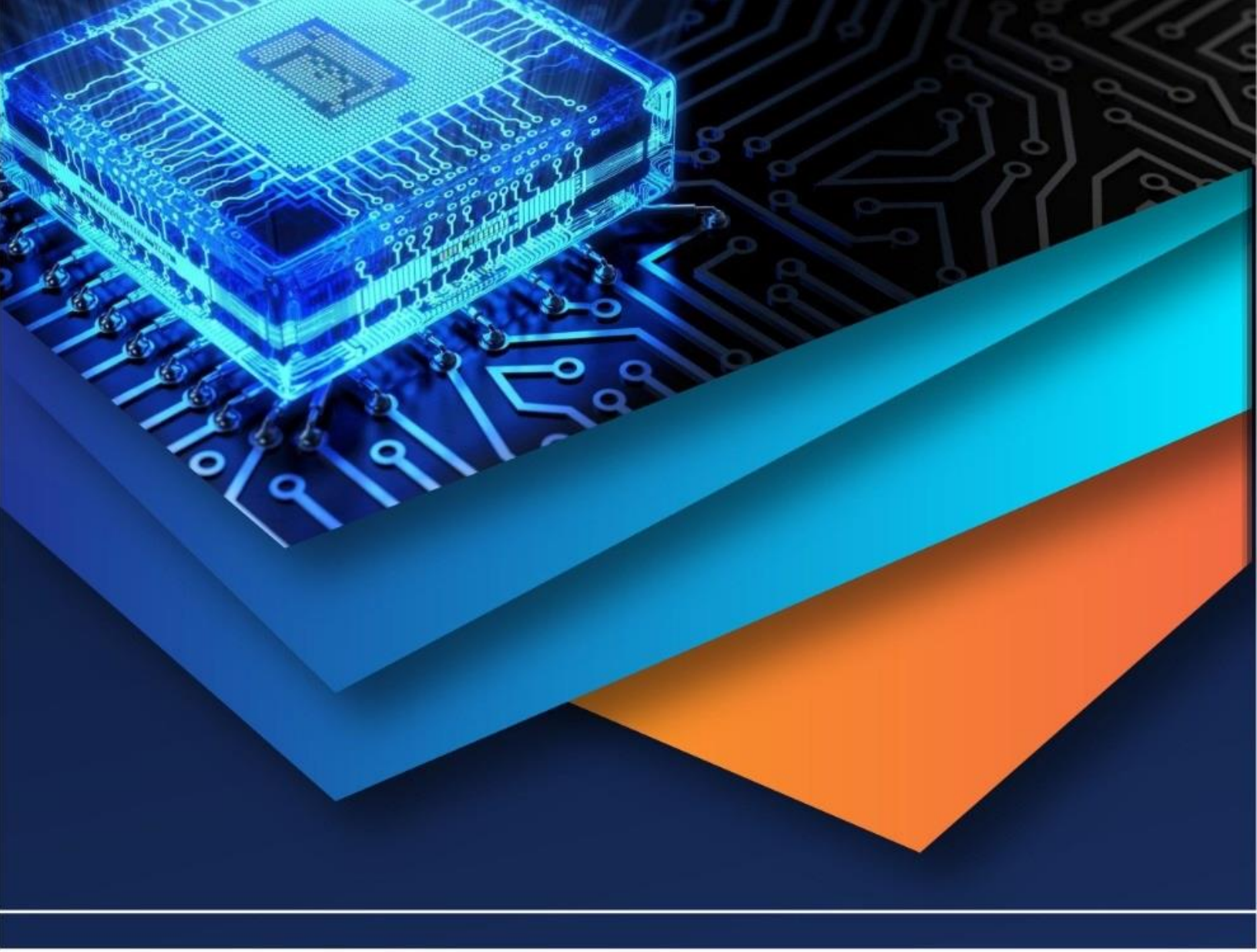
Figure 2. Showing the accuracy of different machine learning algorithms.

VI. CONCLUSION

The Observation leads to the conclusion that machine learning algorithm can be used in successful analysis of text context . The six main algorithms including Naive Bayes , MaxEntropy , SVM, Tree , Forests , Bagging was used and Naive Bayes was found to out perform all other with maximum accuracy of 100%.

REFERENCES

[1] Wael H. Gomaa and Aly A. Fahmy ," A Survey of Text Similarity Approaches",International Journal of Computer Applications (0975 – 8887) Volume 68– No.13, April 2013.
[2] Mukti Desai, Mrs. Kiran Bhowmick," Word Sense Disambiguation",International Journal of Engineering Science Invention Volume 2 Issue 10 October. 2013
[3] A. R. Rezapour, S. M. Fakhrahmad and M. H. Sadreddini ,"Applying Weighted KNN to Word Sense Disambiguation ",Proceedings of the World Congress on Engineering 2011 Vol III WCE 2011, July 6 - 8, 2011, London, U.K
[4] Arti Mishra, Meenakshi Pathak, "A Survey on Impact of Word Sense Ambuiguity on Search Engine's Performance" ,Journal of Basic and Applied Engineering Research Volume 1, Number 8; October, 2014.
[5] Rekha Jain, Rupal Bhargava, Sulochana Nathawat, G.N Purohit ," Sense Disambiguation in Information Retrieval ",International Journal of Emerging Technology and Advanced Engineering, Certified Journal, Volume 3, Issue 4, April 2013.
[6] Gurinder Pal Singh Gosal ,"A Naïve Bayes Approach for Word Sense Disambiguation", International Journal of Advanced Research in Computer Science and Software Engineering , Volume 5, Issue 7, July 2015
[7] F.B. Dian Paskalis , M.L. Khodra , "Word Sense Disambiguation In Information Retrieval Using Query Expansion" ,2011 International Conference on Electrical Engineering and Informatics 17-19 July 2011,
[8] Neetu Sharma, S. Niranjan ,"An Optimized Combinatorial Approach of Learning Algorithm for Word Sense Disambiguation" ,International Journal of Science and Research (IJSR) Volume 3 Issue 6, June 2014
[9] Khaled Abdalgader ,"Text-Fragment Similarity Measurement using Word Sense Identification", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 24 (2016)
[10] Neha Kumari, Sukhbir Kaur,"Online Assessment of Similarity between Sentences in Question Analogous System: A Review Paper",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 5, May 2016
[11] Abhishek Fulmari, Manoj B. Chandak," A Survey on Supervised Learning for Word Sense Disambiguation" ,International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013
[12] Yoong Keok Lee and Hwee Tou Ng , "An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation" Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002
[13] Samhith.K,Arun Tilak.S, G.Panda , "Word Sense Disambiguation using WordNet Lexical Categories", International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY