SEMANTIC TEXTUAL SIMILARITY USING MACHINE LEARNING ALGORITHMS

V Sowmya[1], K Kranthi Kiran[2], Tilak Putta[3]
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

Abstract
Sentence similarity measures plays a key role in text-related research and applications in areas like as text mining, natural language processing, information extraction, etc. Semantic Textual Similarity (STS) measures the degree of semantic equivalence between two text fragments, though the sentence pair has different words. The text fragments are word phrases, sentences, paragraphs or documents. This paper describes various regression techniques of supervised model used to analyze the impact of syntactic and semantic features in calculating the degree of STS. The empirical evaluation done on SemEval-2017 datasets.

Index Terms*:* Regression Techniques, Semantic, Semantic Textual Similarity, Supervised model, Syntactic.

I. INTRODUCTION

Machine Learning (ML) [17] is a sort of artificial intelligence (AI) that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The main goal of ML is to build algorithms that can receive input data and use statistical analysis to predict an output data. Machine learning algorithms are classified as supervised and unsupervised. Supervised algorithms have both input and desired output. And the unsupervised algorithms have input, but not desired output.

By using unsupervised algorithms (clustering, etc.), Researchers may discover the unknown data [1]. And another kind of machine learning is reinforcement learning [2]. This paper presents supervised learning. Again supervised learning characterized into classification, regression, recommendation. The regression model [18] is a predictive modeling technique which examines the relationship between *dependent* (target) and *in dependent variable (s)* (predictor). And it is as well a supervised problem; the outputs are continuous rather than discrete. Regression techniques are mostly driven by three metrics, They are:

i. Number of independent variables,
ii. Type of dependent variables
iii. And shape of regression line.

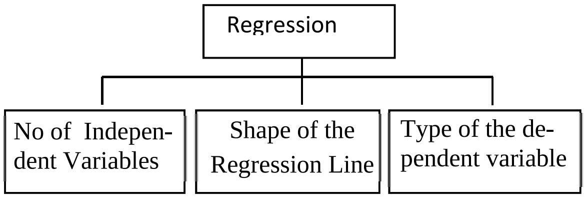
Regression
No of Indepen-dent Variables
Shape of the Regression Line
Type of the de-pendent variable

Fig. 1 Types of regression model

As per this paper describes, the algorithm of regression techniques is: boosting, bagging, random forest, simple vector machines, multi linear regression, rpart etc.

This paper organized in seven sections. Section 2 described about the related work in STS. Section 3 described about the regression techniques. Section 4 described about the lexical, syntactic and semantic features. Section 5 described about the experimental work for preprocessing of the datasets and model building. Section 6 described about the attainment of the results is depicted. Section 7 described about the conclusion of this work.

II. LITERATURE SURVEY

Dutton D [3] describes a brief review of what ML includes and various classification algorithms and the recent attempt for improving classification accuracy—ensembles of classifiers. Boosting was introduced by Schapire [4] as a method for boosting the performance of a weak learning algorithm. Sent2Vec performs the mapping using the Deep Structured Semantic Model (DSSM) proposed in [5], or the DSSM with Convolutional - pooling Structure (CDSSM) proposed in [6]. Wael H. Gomaa [13] are three text similarity approaches were discussed; String-based, Corpus-based and Knowledge-based similarities. In string based approach they explained about N-grams, Longest Common SubString (LCS).

V. Sowmya [16] proposed, the impact of PE on semantic textual equivalence is depicted using Pearson correlation, calculated between system generated scores and the human annotations. Iqbal Muhammad [12] mentioned strength and weakness of classification algorithms of supervised learning. J. R. Quinlan [14] presents the results of applying both techniques to a system that learns decision trees and testing on a representative collection of datasets. While both approaches substantially improve predictive accuracy, boosting shows the greater benefit. All datasets that are taken from The Semantic Evaluation (SemEval) – 2017. SemEval – 2017 [15] is the eleventh workshop in the series of International Workshops on Semantic Evaluation. The Semantic Evaluation (SemEval) series of workshops focuses on the evaluation and comparison of systems that can analyze diverse semantic phenomena in text with the aim of extending the current state of the art in semantic analysis and creating high quality annotated datasets in a range of increasingly challenging problems in natural language semantics. SemEval provides an exciting forum for researchers to propose challenging research problems in semantics and to build systems/techniques to address such research problems.

III. MODEL

A. Supervised Learning Algorithms

The computers have no ability of Learning from the past experiences, as it is an attribute of humans. In supervised machine learning, our main goal is to learn a target function that will be used to predict the values of a class. The process of applying supervised ML to a real-world problem is described in below figure.

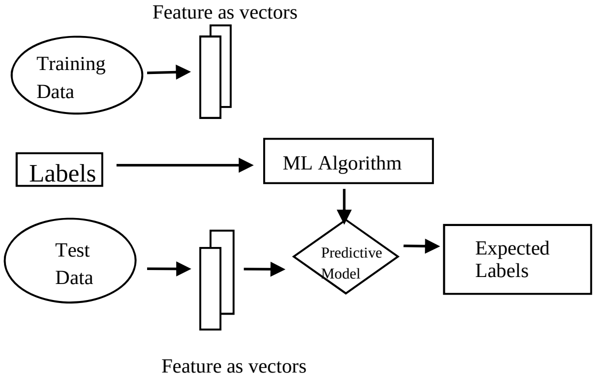

Fig.2 Supervised Machine Learning Model

The first step of the supervised learning is dealing with the dataset. An appropriate expert could suggest a better selection of features in order to perform a better training on the data set. Ultimately, in most cases the datasets contain noise and missing feature values, and therefore it requires the pre-processing the data set. In the next step, data preparation and data preprocessing is a key function in Supervised Machine Learning (SML). A number of techniques have been introduced by different researchers to deal with missing data issue. Hodge & Austin [7] have conducted a survey of contemporary techniques for outlier (noise) detection. Karanjit & Shuchita [8] have also discussed different outlier detection methods which are being used in different machine learning.

How it works: This algorithm consists of a target or dependent variable which is to be predicted from a given set of predictors (independent variables). Using this set of variables, generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data.

B. Regression Techniques

i. Support Vector Regression

Support Vector Machines (SVMs) are a set of supervised learning methods which have been used for classification, regression and outlier's detection [12]. The main goal of SVM is to construct the hyper plane in a high dimensional. Hyper plane refers to a subspace one dimension less than its ambient space. Whether there is 3-dimensional space, then its hyper plane is 2-dimensional planes. By hyper plane a good separation is achieved that has the largest distance to the nearest training-data point of any class.

ii. rpart

Recursive partitioning (report) is a statistical method for multivariable analysis. It creates a decision by dividing it into sub-populations based on some divided independent variables. The procedure is called recursive because each sub-population may be divided an indefinite number of times. After a particular stopping criterion is reached, the splitting process will be carried on.

iii. Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

iv. Bagging

Bagging [9] also called as Bootstrap aggregating. It is an ensemble *learning* method *of Bootstrapping, which* makes individuals for its ensemble by training each regression model on a random redistribution of the training set.

v. Boosting

Boosting approach is *similar to Bagging. It focuses on the performance of the learning algorithm and it also concentrates on instances that have not been correctly learned. It is a* machine learning ensemble algorithm mainly used to reduce bias, and also variance in supervised learning, and a family of machine learning algorithms which convert weak learners to strong ones. By using Boosting algorithm will get better accuracy in classification or regression models.

vi. Multi-Linear Regression

The multiple vector based values predictor variables are called as multi linear regression. Maximum regression models involve multiple predictors, and basic illustrations of linear regression are often idiom in terms of the multiple regression model.

$$a_i = x_0 + x_{1b_i} + x_{2b_i^2} + \varepsilon_i,\ i = 1, \dots, n.$$

Where: a_i = dependent variable

b_i = independent variable

x_o, x_1 = parameters

$\mathrm{\varepsilon_i}$ = ε is an error term and subscript i is an index.

IV. Feature

Numerous Regression techniques exist to combine sets of lexical and semantic information to build a learned model. Lexical and syntactic features are used to measure the structure and similarity between a sentence pair.

A. *The lexical and syntactic features are:*

i. Length features

Length features of sentence pairs are captured using the cardinality of a set. The set of a sentence is defined as the unique words present in that sentence. For example, consider A and B are two sets and which contains the unique words from the different sentence pair called S1 and S2.

It generates some features like $|A|, |A|, |A - B\ |,\ \frac{|A-B|}{|B|}, |B - A|, \frac{|B-A|}{|A|}\ |A \cap B|,$ $|A \cup B|$.

ii. Longest Common Sequence (LCS)

Based on the length of the longest common word sequence (lcws), LCS algorithm will find the similarities between two strings of a sentence pair S1 and S2 respectively.

$$Lcs(s_{1,}\ s_2\ ) = \frac{len(lcws(\ s_1,\ s_2))}{\min(len(s_1),\ len(s_2))}$$

iii. N-gram features

It divides the sentence into n items/grams. N-gram similarity algorithm compare the N-grams from each character/word in a sentence pair. It finds the distance by separating the similar n-grams among many of n-grams.

$$J(A, B) = \frac{|A\ \cap B|}{|A\ \cup B|}$$

Where, J is Jaccard similarity,

A, B are the set of corresponding N-grams.

iv. Word Order Similarity

Li et al. [11] describes the word order similarity measure as the normalized difference of word order between a sentence pair. The equation of word order similarity is shown below

$$sim_{wo}(s_1, s_2) = 1 - \frac{||r_1 - r_2||}{||r_1 + r_2||}$$

Here r_1 and r_2 are word order vector of sentence s_1 and s_2, respectively.

Word order vector is a feature vector whose feature set comes from words that appear in a sentence pair. The index position of the words in the corresponding sentence is used as term weights for the given word features. That is, each entry in the word order vector r_i is derived from computing a word similarity score between a word feature $'w'$ with all the words in the sentence s_i. An index position of the word in si that gives the maximum word similarity score to w is chosen as its term weight.

v. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF), is a numeric measure that is use to give the score of a word in a document based on how much times it appear in that document. The impression for this measure is: Whether a particular identifies very often in one particular document, then it will be treated as high score. When a particular word appears in too many other documents, then it is possibly not an unique identifier, so suites for a lower score to that word. The math formula for this measure:

$$tfidf(t, d, D) = tf(t, d) * idf(t, D)$$

Where ‘t’ = the terms
‘d’ = document
‘D’ = collection of documents

vi. Phrase Entity Alignment (PEA)

A phrase entity [16] is conceived by the combination of zero or one determiner, zero or more adjectives and a noun. Phrase entities are extracted from each sentence in the sentence pair. The semantic similarity is computed between each pair of sentence phrase entities using WordNet and brown corpus.

B. Semantic Feature

Semantic features deal with the meaning of the words in the sentences. The approach being used to calculate semantic similarity will be classified as corpus-based models and knowledge-based models.

i. Knowledge based feature

Knowledge based feature as knowledge-based models try to find how similar two terms are based on some sort of semantic network like an ontology.

ii. Corpus based feature

Corpus-based models are used to find the similarity from a large collection of documents (corpus).

iii. Sent2Vec feature

Sent2Vec model used to maps the similarity a pair of short text strings. It uses two features to find semantic similarity between a sentence pair. They are Deep Structured Semantic Model(DSSM). Similarly, Convolutional pooling Deep Structured Semantic Model (CDSSM).

DSSM: DSSM is a Deep Neural Network (DNN) [5] modeling technique for representing text strings (sentences, queries, predicates, entity mentions, etc.) in a continuous semantic space and modeling semantic similarity between two text strings It has many applications as information retrieval and web search ranking, question answering etc.

CDSSM: To identify contextual structures of Convolutional Deep Structure Semantic Model, the important word n-gram level while DSSM treats them as a bag of words (n-gram) [6].

V. EXPERIMENT

A. Data set:

The data set was taken from The Semantic Evaluation (SemEval) 2017. The data set contains the data as sentence pairs. The Datasets English – English, Arabic – English, Arabic – Arabic, Spanish – Spanish, Turkish – English has 250 sentence pairs. And Turkish – English have 500 sentence pairs. Each pair was assigned similarity scores in the range [0, 5] by multiple human annotators (0: no similarity, 5: identicality) and the average of the annotations was taken as the final similarity score. Each dataset is divided at 80% of data is taken as training data and 20% of data is taken as testing pairs. All data sets were translated into English using Google Translator.

B. *Data preparation and data pre-processing*

The data preprocessing steps are often having a significant impact on the generalization performance of a supervised ML algorithm.

- ✓ Selection is the process of identifying and removing as much irrelevant and redundant information.
- ✓ Removing noisy and unreliable data.
- ✓ Identify the missing values and replace with meaningful value.
- ✓ Data pre-processing includes data cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set.

C. *Model building*

The regression algorithms are used to build the model. The similarity measured by lexical and syntactic features of semantic textual similarity. All regression algorithms are implemented by R using seed as 9.

VI. RESULT

The empirical evaluation is done on SemEval-2017 datasets. From the results it has been observed that, for Arabic – Arabic dataset Bagging Regression model and for Arabic – English, Spanish – Spanish, English – English, Turkish – English datasets Boosting Regression model are given the best accuracy.

Regression models	**Datasets**				
	Arabic – Arabic	**Arabic - English**	**Spanish-Spanish**	**English - English**	**Turkish - English**
Bagging	**0.744**	0.668	0.798	0.796	0.266
Boosting	0.738	**0.729**	**0.819**	**0.811**	**0.49**
Multi Linear	0.723	0.681	0.738	0.804	0.394
RPart	0.638	0.587	0.76	0.748	-0.04
Random Forest	0.734	0.657	0.816	0.789	0.277
SVM	0.688	0.617	0.787	0.795	0.13

VII. CONCLUSION

This paper describes, various regression techniques of supervised model are used to analyze the impact of syntactic and semantic features in calculating the degree of STS. The empirical evaluation is done on SemEval-2017 datasets. By observations Boosting algorithm is giving more accuracy for Arabic – English, Spanish – Spanish, English – English, Turkish – English datasets.

REFERENCES

[1] Jain, A.K., Murty, M. N., and Flynn, P. (1999), Data clustering: A review, ACM Computing Surveys, 31(3): 264–323.

[2] Barto, A. G. & Sutton, R. (1997). Introduction to Reinforcement Learning. MIT Press.

[3] Dutton D, Conroy G (1996) A review of machine learning. Knowl Eng Rev 12:341–367.

[4] Schapire, R.E., Freund,Y., Bartlett, P., & Lee,W.S. (1997). Boosting the margin: A new explanation for the effectiveness of voting methods. In D. Fisher (Ed.), Machine Learning: Proceedings of the Fourteenth International Conference (pp. 322–330). Morgan Kaufmann.

[5] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In CIKM 2013. [PDF]

[6] Jianfeng Gao, Patrick Pantel, Michael Gamon, Xiaodong He, Li Deng, Yelong Shen. 2014. Modeling interestingness with deep neural networks. In EMNLP, Oct, 2014. [PDF]

[7] Victoria J. Hodge and Jim Austin, "A Survey of Outlier Detection Methodologies", Artificial Intelligence Review, Vol. 22, No. 2, pp. 85-126, 2004.

[8] Karanjit Singh and Shuchita Upadhyaya, "Outlier Detection: Applications and Techniques", International Journal of Computer Science Issues, Vol. 9, Issue. 1, No. 3, pp. 307-323, 2012.

[9] L. Breiman, Bagging Predictors. Machine Learning, 24(3) (1996) 123-140.

[10] Wael H. Gomaa and Aly A. Fahm, "A Survey of Text Similarity Approaches" Volume 68– No.13, April 2013.

[11] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In CIKM 2013. [PDF]

[12] Iqbal Muhammad and Zhu Yan, SUPERVISED MACHINE LEARNING APPROACHES: A SURVE, DOI: 10.21917/ijsc.2015.0133

[13] 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.

[14] J. R. Quinlan, Bagging, Boosting, and C4.5, From: AAAI-96 Proceedings. Copyright © 1996, AAAI (www.aaai.org). All rights reserved.

[15] 11th International Workshop on Semantic Evaluations (SemEval-2017), http://nlp.arizona.edu/SemEval-2017/SemEval-2017.pdf

[16] Vangapelli sowmya, Bulusu Vishnu Vardhan, Mantena S.V.S. Bhadri Raju, Improving Semantic Textual Similarity with Phrase Entity Alignment, International Journal of Intelligent Engineering and Systems, Vol.10, No.4, 2017.

[17] MachineLearning, http://whatis.techtarget.com/definition/machine-learning

[18] Analytics Vidhya, https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression