Journal of Theoretical and Applied Information Technology
30th June 2019. Vol.97. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EFFECTIVE SEMANTIC TEXT SIMILARITY METRIC USING NORMALIZED ROOT MEAN SCALED SQUARE ERROR

[1]ISSA ATOUM, [2]MARUTHI ROHIT AYYAGARI

[1]Department of Software Engineering, The World Islamic Sciences and Education, Jordan

[2]College of Business, University of Dallas, Texas, USA

E-mail: [1]issa.atoum@wise.edu.jo, [2]rayyagari@udallas.edu

ABSTRACT

The Pearson correlation is a performance measure that indicates the extent to which two variables are linearly related. When Pearson is applied to the semantic similarity domain, it shows the degree of correlation between scores of dataset test-pairs, the human and the observed similarity scores. However, the Pearson correlation is sensitive to outliers of benchmark datasets. Although many works have tackled the outlier problem, little research has focused on the internal distribution of the benchmark dataset's *bins*. A representative and well-distributed text benchmark dataset embody a wide range of similarity scores values; therefore, the benchmark dataset could be considered a cross-sectional dataset. Although a *perfect* text similarity method could report a high Pearson correlation, the standard Pearson correlation is unaware of correlated individual text pairs in a single dataset's cross-section due to outliers. Therefore, this paper proposes the normalized mean scaled square error method, inferred from the standard scaled error to eliminate the outliers. The newly proposed metric was applied to five benchmark datasets. Results showed that the metric is interpretable, robust to outliers, and competitive to other related metrics.

Keywords: *Pearson, Absolute Error, Text Similarity, Correlation, Scaled Square Error, Outliers*

1. INTRODUCTION

Under heavy noise conditions, extracting the correlation coefficient between two sets of stochastic variables is nontrivial [1]. The performance of a Text Similarity (TS) method is most often calculated by the Pearson correlation between the human-mean scores (first variable or the reference), and the method observed scores (second variable). Formally, the performance of a text similarity method is calculated as the covariance of the two variables divided by the product of their standard deviations, which is a figure value from -1 to 1. When the figure is high, it implies a high correlation with the human scores; therefore, the similarity method becomes favorable over another method in a specific task.

Although Pearson correlation has been theoretically approved and used in many domains, the Pearson correlation if taken in isolation may incidentally indicate invalid causation. It was shown that correlation might indicate (*humorously*) that babies are delivered by storks[2]. Similarly, and using the same correlation, it was reported that the consumption of cocoa flavanols results in an acute improvement in visual and cognitive functions [3]. Therefore, the simplicity of a correlation could hide the considerable complexity in interpreting its meaning[4]. Moreover, the application of Pearson correlation, as a linear relationship is limited to predict the correlation in domains that are not normally distributed. For example, it was shown that the Pearson correlation is not a good predictor for the reliability of characteristics of interest[5]. Despite the ever increasing interests in other alternatives [6]–[10], the Pearson correlation is still dominant in domains of text similarity such as those related to the SemEval tasks workshop series [11], [12].

In Spite Of the simplicity and interpretability of the Pearson correlation in the text similarity domain [13]–[15], the cosine similarity is among others getting attention from scholars, especially in word embeddings applications [16], [17]. It was pointed out that Pearson correlation does not provide enough justifiable results in software engineering domain [18].Therefore, the Pearson correlation should be adapted or modified to handle software engineering issues related to software requirements engineering and testing [19]–[21] .

One major problem of Pearson correlation is the outliers. Outliers have a reflective influence on

Journal of Theoretical and Applied Information Technology
30th June 2019. Vol.97. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

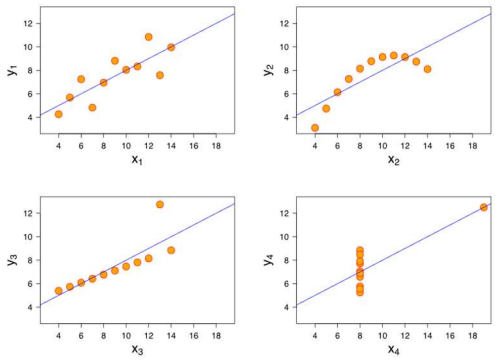

Figure 1 Effect of outliers on Pearson's correlation (Anscombe)

the slope of the regression line, and consequently on the value of the correlation coefficient. The problem is known in the literature as the Anscombe's quartet [22] problem, as shown in Figure 1. The Anscombe's quartet comprises four datasets that have nearly identical Pearson's correlation (0.816), yet they appear very different when graphed. Therefore, datasets distributions should be analyzed to handle outliers.

When a benchmark dataset is designed, it usually works competitively over pairs of text in at least three bins of the dataset that vary in similarity from low (L), medium (M), to high similarity (H). An appropriate similarity method should work well in all cases of dataset scores, L, M, and H. The inherent problem of the standard Pearson correlation is the way of calculation. The standard Pearson correlation does not take into consideration the cross-sectional property of the dataset; instead, it considers all values, including outliers. Therefore, a high Pearson correlation does not guarantee the suitability of the similarity method to its application

Based on the assumption that a useful benchmark dataset is cross-sectional, we claim that there are at least four different similarity methods, low-similarity-method (α), medium-similarity-method (β), high-similarity-method (Ω), and the optimal similarity method (δ). The α method is fair when the dataset (or the cross-section) has low human scores, while the β method is fair when the dataset (or the cross-section) has high human scores. In contrast, the optimal method (δ) should work with all cases of the dataset.

Figure 2 explains the problem with the four types of similarity methods using our crafted demo dataset. The demo dataset reports 0.7 correlation for

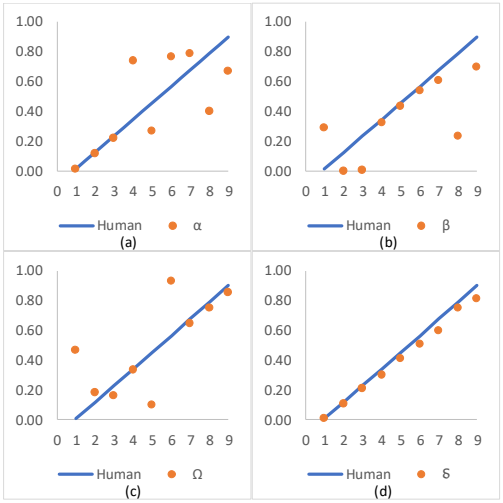

Figure 2 Effect of Similarity method on Pearson's correlation (r=0.7 for a,b,c; r=1.00 for d)

α, β, and Ω methods and 1.0 for the optimum method (δ). On the first hand, an α method (Figure 2a) has a high correlation with text pairs that has low similarity as per human-means (pairs 1-3). On the second hand, an Ω method (Figure 2c) has a high correlation with text pairs that have high similarity as per human-means (pairs 7-9). In contrast, the β method (Figure 2b) has a high correlation with text pairs that has medium similarity as per human-means (pairs 4-6).

Figure 2a is an example with a similarity method that works very well on text pairs that have low similarity while Figure 2b is an example with a similarity method that works very well on text pairs that have a medium similarity, and Figure 2c is an example with a similarity method that works very well on text pairs that are literary similar. The objective is to find a suitable similarity measure that works very well on all benchmark scales. Therefore, a useful method should reduce the errors between actual and observed scores. Therefore, for a task that needs to discover similar text such as plagiarism, the (Ω) is favorable, and for tasks that need to find irrelevant text (irrelevant documents) the method (α) is suitable. Therefore, the standard Pearson method was not able to consider variabilities in text similarity scores. The goal is to choose a method that gives high correlation, such as the *optimum* method in Figure 2d.

Although there are many alternatives to Pearson correlation, most text similarity competitions (e.g., SemEval series [11], [12]) uses Pearson correlation as a standard. Nevertheless, many types of research are pushing toward making a

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

new correlation measure in the text similarity domain. However, most of the ranked correlation methods such as Spearman[7] and the Kendall tau correlation[23] methods suffer from ties and are suitable for datasets that are ranked in nature [24]. Therefore, the aim is to find a method that handles issues of the Pearson correlation and providing alternatives that were not studied deeply in the semantic similarity domain.

Hyndman and Koehler [25] proposed the scaling absolute error methods to scale down observed values in the finance domain. Compared to the relative error methods, the scaling absolute error method is independent of the scale of the observed data, and it can remove the problems of undefined means and infinite variance. Hyndman extended the scaling absolute error method to the Mean Squared Scaled Error (MSSE).

In our context, the absolute error measure is the difference between the text-pair human score and the similarity method observed score. The MSSE is a function of absolute error of human and observed scores concerning the mean variability of observed scores. Consequently, the MSSE should be able to reduce the absolute errors presented in Figure 2. We normalize the MSSE (NMSSE) to a scale between 0 to 1 using the exponent function. The NMSSE, compared to Pearson correlation, ranks text similarity methods based on the target text application task.

Practically, and as a proof of concept, our proposed metric shows the divergence of some commonly cited works. Although the LSA measure of [26] reported good Pearson correlation, it is misjudging text-pairs scores reporting an absolute relative error approaching 80%. Moreover, methods that depend on large corpus tend to overestimate scores of text pairs [27]. The objective of this paper is to propose a new approach that could be used to eliminate data outliers and provide a performance metric to select the best text similarity method.

First, Pearson and its related measures are explained. Next, the proposed metric is explained. Then, the metric is evaluated. After that, we highlight the research implications and limitations. Finally, the paper is concluded.

2. RELATED WORKS

2.1. Pearson Correlation

The Pearson correlation has been proposed long back [28], yet it still applicable as an evaluation metric for many SemEval tasks workshop series [11], [12].

The Pearson correlation is calculated as the covariance of the two variables divided by the product of their standard deviations[29]. In the text similarity domain, the variables are the human-mean scores' group and the related observed test-score group. So, if we have one dataset scores $\{h_1,..., h_n\}$ that represent the human-mean scores of a list of text pairs and another dataset $\{o_1,..., o_n\}$ containing n observed scores (from a text similarity method), the Pearson's correlation coefficient, r, is shown in (1).

$$r = \frac{\sum_{i=1}^{n}(h_i - \bar{h})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{n}(h_i - \bar{h_\iota})^2} \cdot \sqrt{\sum_{i=1}^{n}(o_i - \bar{o})^2}} \quad (1)$$

Where n is the the number of text pairs. h_i, o_i are the i[th] score of human-mean (i.e., reference) and test (observed) scores text pairs. $\bar{h_\iota}$ and $\bar{o_\iota}$ are the mean of the gold standard and test scores respectively.

2.2. Ranking Methods

The Spearman method [7] is considered one of the most cited alternative methods to Pearson correlation; however, it is not used regularity in text similarity domain because it works on ranked data, which is not reasonable in text similarity [24]. Similarly, the Kendall tau correlation[23], which calculates the proportion ranks between datasets, is rarely seen in text similarity domain.

Several other methods measure the gain of a document based on its position in the result list [8]–[10]; however, these methods suffer from ties and are not suitable for scaled text similarity measures[30]. The Hoeffding's D method, a non-parametric measure, measures the difference between the joint ranks and the product of their marginal ranks[31]. The distance correlation as its name implied, is based on the distance (usually Euclidian) to measure the dependence between two variables[32], [33]. The maximal information coefficient (MIC) is a measure of the strength of the linear or non-linear association between two variables[34]; however, it does not perform well in low sample size[35].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.3. Error methods

Error methods are used to quantify the difference or percentage between actual and forecast values. The absolute error computes the amount of error in a trial. The relative error is an extension to the absolute error with relative to the original real value. These methods are easy-to-use [36].

3. PROPOSED METRIC

Equation (2) defines the absolute error (AE_j) of a text pair j, as the difference between human scores (actual, h_j) score and the observed scores (predicted, $\hat{h}_j$) of a text similarity measure.

$$AE_j = |\hat{h}_j - h_j| \quad (2)$$

$$S_{E_j} = \frac{AE_j}{\frac{\sum_{i=1}^{n} \hat{h}_j - \bar{\hat{h}}}{n}} \quad (3)$$

The scaled error (S_{E_j}) for each text pair j is given by equation (3) , where n is the number of text pairs in the benchmark dataset. The $\bar{\hat{h}}$ is the mean of the observed method similarity score. Then the mean scaled square error (MSSE) is defined by equation (4).

$$MSSE = \frac{\sum_{j=1}^{n} (S_{E_j})^2}{n} \quad (4)$$

The lowest value of MSSE is zero when the absolute error of actual and predicted values is zero and is infinity when all predicted values are identical; that is the mean of observed scores ($\bar{\hat{h}}$) equals every predicted value ($\hat{h}_j$). Therefore, we normalize the values of the MSSE between (0,1) to allow a quantitative comparison between different datasets as shown in equation (5). Where the $MSSE$ as shown in equation (4), and e is the exponent value. The NMSSE equals the value of 1 when the error is at the maximum and 0 when the error is very low. Therefore, for ranking similarity methods, the lower NMSSE the better.

$$NMSSE = 1 - e^{-MSSE} \quad (5)$$

4. EVALUATION AND DISCUSSION

4.1. Datasets used in the Experiments

Table 1 shows the set of datasets used in the experiment. The datasets are split into two categories: development (6,427 text pairs) and test datasets (1,909 text pairs). The goal of the split was to support text similarity measures that depended on pre-training or test training[12]; however, in our case, we used both datasets for the selected text measures. We filter datasets from stopwords using the nltk stop words' list.

4.2. Selected Text Similarity Measures

For this paper, the selected text measures illustrate the applicability of the proposed metric over a wide range of text similarity measures, as shown in Table 2.

Table 1 Benchmark datasets

Dataset	Dev.	Test	Total	Description
Demo Crafted Dataset	-	9	9	We prepare this dataset to illustrate similarity measures' problems and to apply the proposed metric on a simple to view dataset.
STS -30	-	30	30	30-sentence pairs collected by Li [37] based on dictionary definitions of words from [38].
SemEval STS	1500	1379	2879	The datasets include text from image captions, news headlines, and user forums which are part of the text similarity tasks of SemEval series [12]
SICK	4927	500	5427	Sentences Involving Compositional Knowledge (SICK) are English sentences from the 8K ImageFlickr and the SemEval 2012 STS MSR-Video Description dataset[39]

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2 Methods used in this experiment

Method	Description
α Method	A demo method used on our crafted demo dataset. An α method is a similarity method used to demonstrate a text similarity method that is leaned toward dissimilar text pairs. The method produces an observed score that is 95% accurate to the human-means for the first three pairs and value at random for the remaining pairs.
β Method	A demo method used on our crafted demo dataset. A β method is a similarity method used to demonstrate a text similarity method that is leaned toward moderately similar text pairs. The method produces an observed score that is 95% accurate to the human-means for the 4-6 pairs and value at random for the remaining pairs.
Ω Method	A demo method used on our crafted demo dataset. An Ω method is a similarity method used to demonstrate a text similarity method that is leaned toward high similar text pairs. The method produces an observed score that is 95% accurate to the human-means for the 6-9 pairs and value at random for the remaining pairs.
δ method	A demo method to show the method that scores the highest Pearson score. The method produces an observed score that is 95% accurate to the human-means scores.
InferSent	InferSent (INF for shorthand), a sentence embedding trained on fastText vectors of Facebook research. INF is BiLSTM with max pooling that was trained on the 570k English sentence pairs of SNLI dataset. [40].
GSE	The universal Google's sentence encoder (GSE) converts any text to a semantic vector. The semantic measure is based on deep learning on the semantic space. We use the Encoder 2 from Google TensorFlow Hub.
TSM	Text Similarity Measure (TSM) is a WordNet measure that calculates the semantic similarity of two sentences using information from WordNet and corpus statistics [27].
WMD	The Word Mover's Distance (WMD) method uses the word embeddings of the words in two texts to measure the minimum amount that the words in one text need to "travel" in semantic space to reach the words of the other text [41]. We use the pre-trained word vectors of Glove (840B tokens) and fastText word vectors W2V (2 million-word vectors).
SIF	The Smooth Inverse Frequency (SIF) uses less weight to solely unrelated words, and so word embeddings are weighted based on the estimated relative frequency of a word in a reference corpus and the common component analysis technique [42]. We use the pre-trained word vectors of Glove (840B tokens) and fastText word vectors W2V (2 million-word vectors).

4.3. NMSSE Illustrated over the Demo Dataset

For illustration and showing various cases of text similarity measures over a wide range of datasets, we use a demo dataset for this experiment. Table 3 shows the list of crafted text pair's scores over four crafted methods α, β, Ω, δ methods as described in Table 2. The table shows the cross-sections of the dataset (bins 1 to 3), and the score for each individual pair using the crafted methods.

Figure 3 shows the Pearson correlation, Spearman, and the proposed NMSSE metric of the data in Table 2. The figure also shows the Pearson correlation for the text pairs 1-3, 4-6,7-9 legend as Pearson_Q1, Pearson_Q2, Pearson_Q3 respectively. Results show that methods that are good to measure no similar text (α method) have high Pearson correlation on the first three text pairs (Pearson_Q1), while methods that are good to measure high similar text (Ω method) has high correlation on the last three text pairs (Pearson_Q3). In the middle between the two methods, the β method shows a high Pearson correlation between the 4-6 pairs (Pearson_Q2).

The reported findings of the three demo methods indicate that the absolute error between human scores and predicted scores is low. Therefore, for a task that needs to discover similar text such as plagiarism the (Ω) is favorable and for semantic tasks (irrelevant documents) that needs to find irrelevant text the (α) method is appropriate.

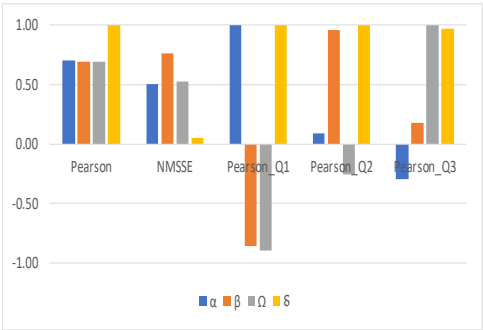

Figure 3 Crafted dataset Pearson correlation

Journal of Theoretical and Applied Information Technology
30th June 2019. Vol.97. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3 The demo similarity methods

Bin	Pair	Human	α Sim.	β Sim.	Ω Sim.	δ Sim.
1	1	0.01	0.01	0.29	0.47	0.01
	2	0.12	0.12	0.00	0.19	0.11
	3	0.23	0.22	0.01	0.16	0.21
2	4	0.34	0.73	0.33	0.33	0.30
	5	0.45	0.27	0.43	0.10	0.41
	6	0.57	0.76	0.54	0.93	0.51
3	7	0.68	0.79	0.61	0.64	0.60
	8	0.79	0.40	0.24	0.75	0.75
	9	0.90	0.67	0.70	0.85	0.81

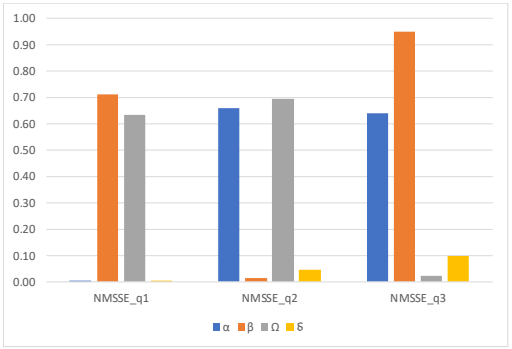

Figure 4 NMSSE performance over sections of datasets

In contrast, the δ method, the best method, has a smooth absolute error except for the outlier shown in the pair number 8. The best method (δ) shows the lowest errors over the dataset.

The unproductive performance of Pearson correlation shown in Figure 3 is illustrated in Figure 4. According to Figure 4, the NMSSE is the lowest for the α method because the α method was doing well in pairs 1-3. The NMSSE also was the lowest

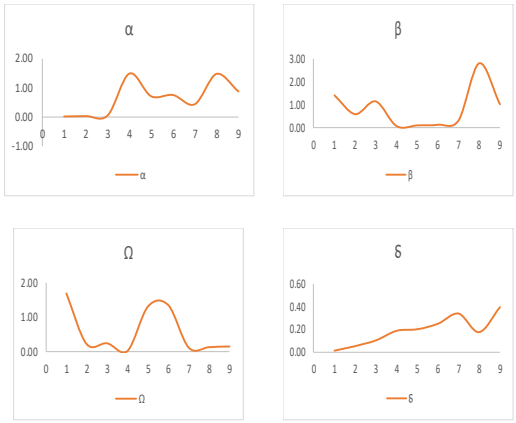

Figure 5 Scaled errors over different measures

for the Ω method since the Ω method is doing well for pairs 7-9. The same thing could be applied to the β method since the β method was doing well for pairs 4-6. The best *optimum* method δ shows lower values for NMSSE for all the three cross-sections of the dataset.

Table 4 shows the statistics of the demo data as per equations (2) – (4). Although α, β are similar in absolute error, they are different in scaled errors because β is higher in the MASE as shown the Figure 4. The root cause of this problem is that as $\hat{h}_j - \bar{\hat{h}}$ increase, the denominator in the equation (3) increase and as a result, the value of the equation is reduced. If $\hat{h}_j \cong \bar{\hat{h}}$, that is the value of the predicted score is like the mean of all predictions, we will get the highest possible error. Although Ω method has the highest test score mean variability ($\sum_{i=1}^{n} \hat{h}_j - \bar{\hat{h}}$), it ranked as the third method using the NMSSE. As shown in Figure 5, the scaled errors are reduced when the method matches the type of the similarity method.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

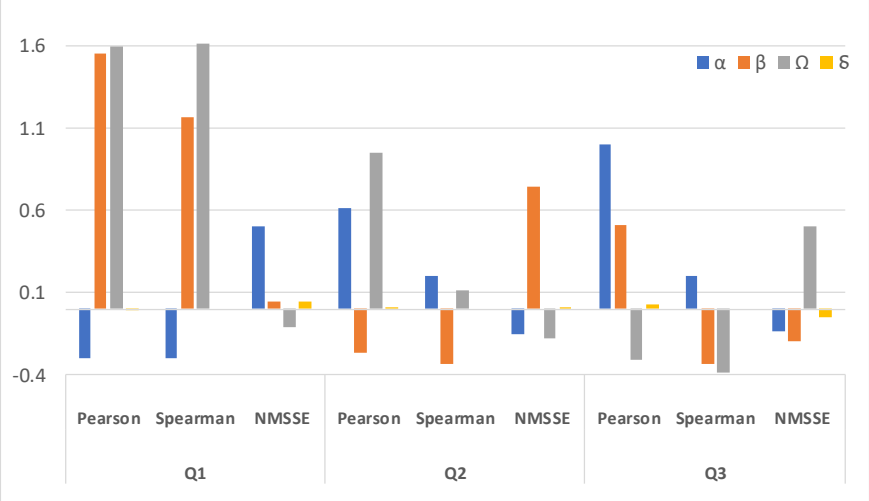

Figure 6 Variability over data segments

Table 4 Statistics of the crafted dataset

	α	β	Ω	δ
$\sum_{j=1}^{9} e_j$	1.52	1.51	1.59	0.39
$\sum_{i=1}^{n} \hat{h}_j - \bar{\hat{h}}$	2.38	1.76	2.4	2.04
$\sum_{j=1}^{9} S_{E_j}$	5.74	7.72	5.35	1.72
MAE	0.17	0.17	0.16	0.04
MASE	0.64	0.86	0.59	0.19
Pearson	0.70	0.70	0.70	1.00
Spearman	0.70	0.67	0.62	1.00
NMSSE	0.47	0.58	0.48	0.17

Furthermore, we calculate the variability between a performance metric (including the NMSSE) on the whole dataset and the value of the compared metric on each section of the dataset Q1, Q2, and Q3. The target is that we should select the performance metric that has the lowest variability; a metric that works well in many situations. Figure 6 shows the variability between Pearson, Spearman, and the proposed NMSSE concerning the three sections of the dataset; pairs 1-3,4-6,7-9 respectively. The lowest variability was in NMSSE for the *best* method, δ. Whereas the Pearson measure shows a higher variability due to outliers in each dataset section. We deduce that NMSSE is effective in scaling data and in removing outliers. However, the NMSSE shows a relatively higher variability in the Q1 dataset because most datasets in this section has low similarity scores that will affect the denominator in equation (3).

4.4. Practical Evaluation of NMSSE

Table 5-7 shows the performance of the NMSE, Pearson correlation, Spearman, and the MAE for the selected methods presented in Table 2. The scores were calculated using the weighted average method based on the number of text pairs in both development and test benchmark datasets. The predicted values and human-mean scores were normalized to be in range 0 to 1 to normalize errors for method.

The NMSSE proposes to rank text similarity methods. As Table 5 shows, if an application is looking for an alternative text similarity method, the GSE is preferred over other methods as they have the lowest NMSSE. The only restriction in this scenario is that the application should be based on any dataset that imitates a similar domain to the SICK dataset. On the STS dataset (Table 6) the SIF method is the best method as it got the lowest NMSSE. However, on the 30-pair dataset (STS-65) shown in Table 7, the SIF had the lowest NMSSE. We emphasize that the proposed metric is smooth-grained with the benchmark dataset, which gives an advantage of our metric over other methods.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 5 Weighted Scores on the SICK dataset (Dev, Test)

	GSE	INF	SIF (W2V)	SIF (GLOVE)	WMD (GLOVE)	WMD (W2V)	TSM
Pearson	0.82	0.76	0.73	0.72	0.64	0.64	0.48
Spearman	0.77	0.70	0.61	0.59	0.59	0.59	0.43
MAE	0.09	0.12	0.16	0.15	0.43	0.43	0.15
NMSSE	**0.43**	0.54	0.52	0.53	1.00	1.00	0.98

Table 6 Weighted Scores on STS dataset (DEV, Test)

	GSE	INF	SIF (W2V)	SIF (GLOVE)	WMD (GLOVE)	WMD (W2V)	TSM
Pearson	0.78	0.75	0.73	0.72	0.55	0.61	0.36
Spearman	0.77	0.74	0.70	0.71	0.55	0.61	0.37
MAE	0.23	0.21	0.18	0.20	0.37	0.38	0.24
NMSSE	0.88	0.92	**0.72**	0.83	1.00	1.00	0.94

Table 7 STS65 scores

	GSE	INF	SIF (W2V)	SIF (GLOVE)	WMD (GLOVE)	WMD (W2V)	TSM
Pearson	0.78	0.80	0.80	0.73	0.69	0.74	0.52
Spearman	0.80	0.79	0.77	0.79	0.63	0.68	0.47
MAE	0.27	0.39	0.12	0.16	0.47	0.42	0.35
NMSSE	0.91	1.00	**0.37**	0.50	1.00	1.00	1.00

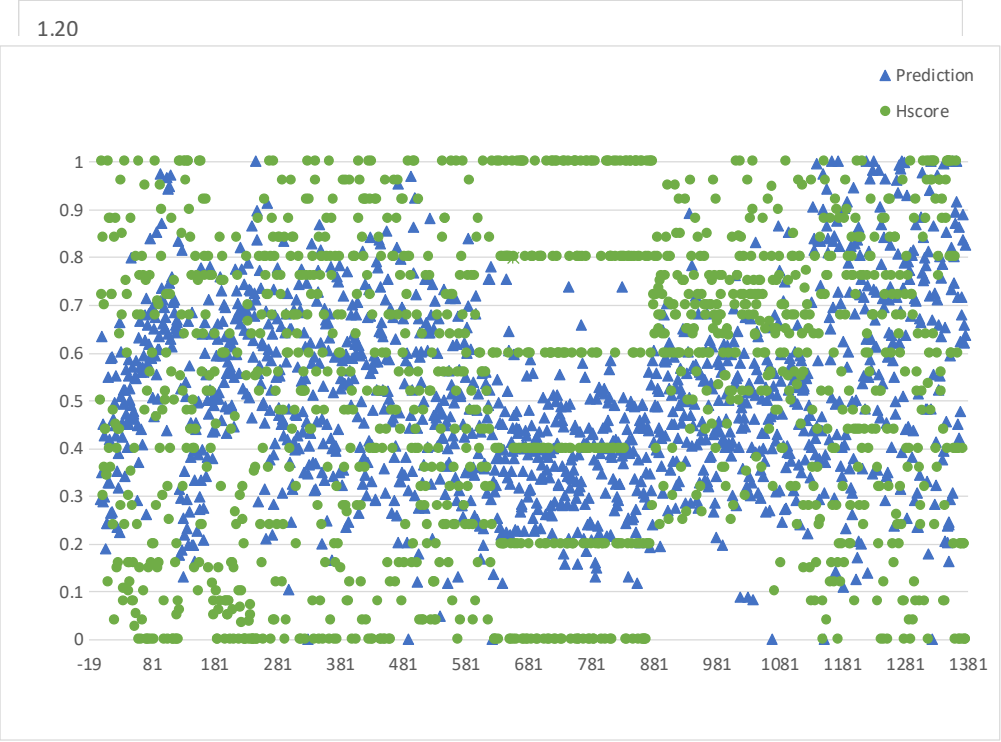

Figure 8 NMSEE of WMD over STS dataset

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The application of the NMSSE handles the problematic issues of Pearson correlation, as shown in Figure 7. The figure shows the weighted scores over all the five benchmark datasets. The leaders are the GSE and the INF methods as they have the lowest SSE compared to other methods. Over the datasets, the traditional edge counting method TSM method outperformed the frequency (SIF) and word distance method (WMD) due to the addition of knowledge from WordNet exploited by the TSM. We noticed that the WMD method got the highest NMSSE due to the scaled error value which was (10^{-6}); consequently, the NMSSE will be high as the denominator of equation (2) becomes low. The root cause of the low scaled error was due to the predicted values of the WMD method that had a mean of 0.5; In other words, the average of the difference between the prediction of the scores and the mean of the prediction approaches zero. Figure 8 shows the WMD method scores and the human scores for the 1380 text pairs of the STS test benchmark dataset. The figure shows that the WMD is overestimating or underestimating scores by almost a constant value. Therefore, the WMD got the lowest NMSSE.

4.5. Comparing NMSSE with Related Methods

To our knowledge, no complete performance metric could be used for the text similarity domain. We carry out a comparison between the proposed NMSSE and other methods over the following criteria:

A. Interpretability: a useful performance metric should be easy to use and interpret; therefore, its output can be easily compared within a predefined scale.
B. Dependency: a useful metric should find the dependency between the human scores and the predicted scores.
C. In-group relationship: a useful metric should indicate how each value in the group is related to each other. As the human scores in a benchmark dataset have a range of values between 0 to 5, the predicted scores should have similar consistent behavior.
D. Robustness to outliers: performance metrics should resolve outliers' issues without affecting the ultimate performance metric score.
E. scale: a performance metric that has a numeric value (e.g., 0 to 1) is quantifiable when compared to other values resulted from other related applications.

Table 8 shows the comparison of our metric and a list of selected metrics where the ☑ stands for the availability of the criterion while ☒ stands for a non-applicable criterion. Although most of the compared methods are interpretable (A), they suffer from outliers (D). The MAE can be made interpretable by getting the relative or percentage error. The drawback of the MAE is that it does not take into consideration the in-group predicted scores (C), and it does not provide a standard scale (E). We underline that we are not looking to replace Pearson correlation but to add extra information that could be utilized to researchers in natural language processing and machine learning communities.

Table 8 Comparison of the proposed metric and related approaches

Criterion	NMSSE	Pearson	Ranking Methods	MAE
A	☑	☑	☑	☒
B	☑	☑	☑	☑
C	☑	☑	☑	☒
D	☑	☒	☒	☒
E	☑	☑	☑	☒

5. IMPLICATIONS

The implication of this research is theoretical and practical. The new measure suggests a re-look to the ongoing usage of the Pearson correlation for a long time. In practice, applications should select the similarity method with the lowest possible normalized error. Although the scaled error method was borrowed from a non-related domain (the finance domain), the new proposed normalized scaled square error could be used in other domains where outliers play a significant effect in natural language processing task. Since the proposed metric is robust to outliers and provides an interpretable scaled value, it would be practical in comparing text in domains such as plagiarism detection and text entailments.

6. LIMITATION

Despite the fact that the proposed method is superior in ranking and text evaluation, researchers need to do more research before generalizing results. The method was applied to five

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

datasets only, and it was not applied practically in any semantic text similarity task.

The research direction should target to generalize the results with text similarity by annotating current and new datasets to allow the comparison of the proposed approach with other alternatives. Therefore, further experiments are needed to test the situations where we would prefer the Pearson correlation over the proposed normalized means scaled square error method.

In the future, the proposed approach should be evaluated using simulations and applying the proposed method on a large empirical dataset.

7. CONCLUSION

This paper proposes a new semantic similarity metric that could be used to compare and rank semantic similarity methods. The proposed metric reduces dataset noise by scaling absolute error with the mean of the absolute difference of observed scores with observed mean scores. The metric was compared with Pearson, Spearman, and the Mean Absolute Error. Results showed that the new proposed normalized scaled square error is effective in reducing skewness and is applicable in domains with different observed scores. In the future, we plan to run several simulations over the new metric and evaluate the metric with extra-large benchmark datasets.

REFERENCES

[1] N. Moriya, "Noise-Related Multivariate Optimal Joint-Analysis in Longitudinal Stochastic Processes," *Prog. Appl. Math. Model.*, pp. 223–260, 2008.

[2] T. Höfer, H. Przyrembel, and S. Verleger, "New evidence for the theory of the stork," *Paediatr. Perinat. Epidemiol.*, vol. 18, no. 1, pp. 88–92, 2004.

[3] D. T. Field, C. M. Williams, and L. T. Butler, "Consumption of cocoa flavanols results in an acute improvement in visual and cognitive functions," *Physiol. Behav.*, vol. 103, no. 3–4, pp. 255–260, 2011.

[4] R. Aggarwal and P. Ranganathan, "Common pitfalls in statistical analysis: The use of correlation techniques," *Perspect. Clin. Res.*, vol. 7, no. 4, p. 187, 2016.

[5] O. Hryniewicz and J. KArpińsKi, "Prediction of reliability--the pitfalls of using Pearson's correlation," *Eksploat. i Niezawodn.*, vol. 16, 2014.

[6] F. Serinaldi, A. Bárdossy, and C. G. Kilsby, "Upper tail dependence in rainfall extremes: would we know it if we saw it?," *Stoch. Environ. Res. risk Assess.*, vol. 29, no. 4, pp. 1211–1233, 2015.

[7] C. Spearman, "The proof and measurement of association between two things," *Am. J. Psychol.*, vol. 15, no. 1, pp. 72–101, 1904.

[8] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000, pp. 41–48.

[9] J. Kekäläinen, "Binary and graded relevance in IR evaluations—comparison of the effects on ranking of IR systems," *Inf. Process. Manag.*, vol. 41, no. 5, pp. 1019–1033, 2005.

[10] D. Katerenchuk and A. Rosenberg, "RankDCG: Rank-Ordering Evaluation Measure," *CoRR*, vol. abs/1803.0, 2018.

[11] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, D. Cer, and A. Gonzalez-Agirre, "Semeval-2012 task 6: A pilot on semantic textual similarity," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 2012, no. 3, pp. 385–393.

[12] D. M. Cer, M. T. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation," *CoRR*, vol. abs/1708.0, 2017.

[13] S. Nithya, A. Srinivasan, M. Senthilkumar, and others, "Calculating the user-item similarity using Pearson's and cosine correlation," in *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, 2017, pp. 1000–1004.

[14] I. Atoum, "A Novel Framework for Measuring Software Quality-in-use based on Semantic Similarity and Sentiment Analysis of Software Reviews," *J. King Saud Univ. - Comput. Inf. Sci.*, p. , 2018.

[15] I. Atoum, A. Otoom, and N. Kulathuramaiyer, "A Comprehensive Comparative Study of Word and Sentence Similarity Measures," *International Journal of Computer Applications*, vol. 135, no. 1. Foundation of Computer Science (FCS), NY, USA, pp. 10–17, 2016.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[16] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 2014, vol. 12, pp. 1532–1543.

[17] Y. Li, L. Xu, F. Tian, L. Jiang, X. Zhong, and E. Chen, "Word embedding revisited: A new representation learning and explicit matrix factorization perspective," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina*, 2015, pp. 3650–3656.

[18] I. Atoum, "A Scalable Operational Framework for Requirements Validation Using Semantic and Functional Models," in *Proceedings of the 2Nd International Conference on Software Engineering and Information Management*, 2019, pp. 1–6.

[19] M. R. Ayyagari and I. Atoum, "CMMI-DEV Implementation Simplified:A Spiral Software Model," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 4, pp. 445–450, 2019.

[20] M. R. Ayyagari, "iScrum: Effective Innovation Steering using Scrum Methodology," *Int. J. Comput. Appl.*, vol. 178, no. 10, pp. 8–13, May 2019.

[21] I. Atoum, "Requirements Elicitation Approach for Cyber Security Systems," *i-manager's J. Softw. Eng.*, vol. 10, no. 3, pp. 1–5, 2016.

[22] F. J. Anscombe, "Graphs in Statistical Analysis," *Am. Stat.*, vol. 27, no. 1, pp. 17–21, 1973.

[23] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[24] J. D. Gibbons and M. Kendall, "Rank correlation methods," *Edward Arnold*, 1990.

[25] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Int. J. Forecast.*, vol. 22, no. 4, pp. 679–688, Oct. 2006.

[26] J. O'Shea, Z. Bandar, K. Crockett, and D. McLean, "A Comparative Study of Two Short Text Semantic Similarity Measures," in *Agent and Multi-Agent Systems: Technologies and Applications*, vol. 4953, N. Nguyen, G. Jo, R. Howlett, and L. Jain, Eds. Springer Berlin Heidelberg, 2008, pp. 172–181.

[27] I. Atoum and A. Otoom, "Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 7, no. 9. The Science and Information (SAI) Organization Limited, pp. 124–130, 2016.

[28] R. S. (Great Britain), *Proceedings of the Royal Society of London*, no. v. 58. Taylor & Francis, 1895.

[29] J. Lee Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *Am. Stat.*, vol. 42, no. 1, pp. 59–66, 1988.

[30] L. A. Goodman and W. H. Kruskal, "Measures of association for cross classifications," in *Measures of association for cross classifications*, Springer, 1979, pp. 2–34.

[31] W. Hoeffding, "A non-parametric test of independence," *Ann. Math. Stat.*, pp. 546–557, 1948.

[32] G. J. Székely, M. L. Rizzo, and others, "Brownian distance covariance," *Ann. Appl. Stat.*, vol. 3, no. 4, pp. 1236–1265, 2009.

[33] G. J. Székely, M. L. Rizzo, N. K. Bakirov, and others, "Measuring and testing dependence by correlation of distances," *Ann. Stat.*, vol. 35, no. 6, pp. 2769–2794, 2007.

[34] D. N. Reshef *et al.*, "Detecting novel associations in large data sets," *Science (80-).*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[35] R. Heller, Y. Heller, and M. Gorfine, "A consistent multivariate test of association based on ranks of distances," *Biometrika*, vol. 100, no. 2, pp. 503–510, 2012.

[36] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, 2009.

[37] Y. Li, D. Mclean, Z. Bandar, J. D. O. Shea, and K. Crockett, "Sentence Similarity Based on Semantic Nets and Corpus Statistics," vol. 18, no. 8, pp. 1–35, 2006.

[38] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Commun. ACM*, vol. 8, no. 10, pp. 627–633, Oct. 1965.

[39] M. Marelli *et al.*, "A SICK cure for the evaluation of compositional distributional semantic models.," in *LREC*, 2014, pp. 216–223.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[40] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 670–680.

[41] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International Conference on Machine Learning*, 2015, pp. 957–966.

[42] S. Arora, Y. Liang, and T. Ma, "A Simple but Tough-to-Beat Baseline for Sentence Embeddings," in *International Conference on Learning Representations*, 2017.