An Empirical Performance Evaluation of Semantic-Based Similarity Measures in Microblogging Social Media

Noufa N. Alnajran, Keeley A. Crockett, David McLean, Annabel Latham
Department of Computing, Mathematics, and Digital Technology
Manchester Metropolitan University
Noufa.alnajran@stu.mmu.ac.uk,{k.crockett, d.mclean, a.latham}@mmu.ac.uk

***Abstract*—Measuring textual semantic similarity has been a subject of intense discussion in NLP and AI for many years. A new area of research has emerged that applies semantic similarity measures within Twitter. However, the development of these measures for the semantic analysis of tweets imposes fundamental challenges. The sparsity, ambiguity, and informality present in social media are hampering the performance of traditional textual similarity measures as "tweets", have special syntactic and semantic characteristics. This paper reviews and evaluates the performance of topological, statistical, and hybrid similarity measures, in the context of Twitter analysis. Furthermore, the performance of each measure is compared against a naïve keyword-based similarity computation method to assess the significance of semantic computation in capturing the meaning in tweets. An experiment is designed and conducted to evaluate the different measures through examining various metrics, including correlation, error rates, and statistical tests on a benchmark dataset. The potential weaknesses of semantic similarity measures in relation to Twitter applications of textual similarity assessment and the research contributions are discussed. This research highlights challenges and potential improvement areas for the semantic similarity of tweets, a resource for researchers and practitioners.**

Keywords— statistical semantics, semantic similarity, online social network analysis, text similarity, Twitter, WordNet

I. Introduction

Short Text Semantic Similarity (STSS) measures are employed for measuring the degree to which short-texts are subjectively evaluated by humans as being semantically equivalent to each other [1]. Short-texts refer to typical human utterances that are of sentence length ranging from 10 to 25 words [2]. Human generated sentences are prone to forms of text that do not conform to typical grammatical and syntactical rules of a sentence. O'Shea et al. [2] suggested that semantic similarities of these short-texts can be measured through the application of STSS measures. These measurements are gaining prominence as much research in the field of natural language processing (NLP) and artificial intelligence (AI) are emerging in multiple domains. The task of assessing the semantic similarity between short-texts has been a central problem in NLP, due to its importance in a variety of applications. Some of the earliest text similarity applications have been implemented for text classification and information retrieval [3], automatic word sense disambiguation [4], and extractive text summarization [5]. More recent applications of STSS include the incorporation of the measure in a conversational agent to reduce the time associated with the scripting process [6], measuring the similarity between documents [7], and in supervised learning and text classification [8]. Measuring semantic similarity can be performed at various levels, ranging from words, phrases and sentences, to paragraphs and documents. Each of these categories employ different methods and techniques to gauge the underlying meaning at that particular level.

A. Problem Statement

In this paper, the focus is on semantic similarity measures at the short text level. The challenges in determining the degree of semantic equivalence between sentences is attributed to the variations in natural language expressions. In natural languages, a single meaning of a sentence can be expressed in many ways, and therefore the task of measuring the semantic similarity of natural language sentences is very complex. This problem is more prevalent in Online Social Network (OSN) texts due to the informal nature and the high degree of lexical variations used. Areas of work within related fields, such as classification and clustering of tweets face similar issues when identifying similarities in natural language text presented in Twitter [9]. To illustrate some challenges present in Twitter, consider the following tweet [10]: "#qcpoli enjoyed a hearty laugh today with #plq debate audience for @jflisee #notrehome tune was that the intended reaction?" The presence of symbols, spelling mistakes, letter repetitions, e.g. "@jflisee", and abbreviations complicate the process of tokenization and Part-of-Speech [11] tagging required by text analysis tasks. Little research has been conducted in the area of semantic analysis of Twitter data especially in relation to semantically measuring the degree of equivalence between tweets. This may be attributed to the characteristics of such data that make the task significantly more difficult than analyzing general short-text. However, several studies highlighted the potential and significance of developing semantic similarity measures [12] and paraphrase identification techniques [13], [14] specifically for tweets. In the context of Twitter, semantic similarity measures are particularly useful in reducing the challenge of high redundancy and the sparsity inherent in its data. One of the possible approaches to reduce the complexity of dealing with massive data is through integration of these measures in applications of Machine Learning.

This paper addresses the problem of STSS applicability in

the context of Twitter short text messages. As these messages share special lexical and syntactical characteristics, traditional STSS measures, which analyse proper English sentences fail to capture the semantic similarities between these messages. Therefore, this paper sets out to review and empirically evaluate different approaches to STSS measures to compare their performance on a labelled dataset of tweets. This is particularly important for research aiming to adapt or develop new STSS measures that consider the different sorts of noise present in social media data.

B. Research Questions

The paper aims to answer the following research questions:

RQ1. *Which approaches exist that support the identification of semantic similarity between Twitter short text messages?*

RQ2. *What are the challenges present in the language used in Twitter that hinder an effective process of semantic similarity identification?*

RQ3. *How do different kinds of STSS measures perform in relation to human assessments for Twitter short-text Messages?*

C. Contributions and Outline

In this paper, topological-based and statistical-based STSS measures are reviewed and evaluated in terms of performance. Towards accomplishing this purpose, the research investigated in this paper has the following objectives:

1) Provide an overview of the different approaches that can be adapted for identifying sentence-based semantic similarities.
2) Highlight the challenges of the natural language used in Twitter that hamper the performance of semantic similarity measures.
3) Evaluate and compare the performance of various STSS measures in applications of Twitter short text messages.

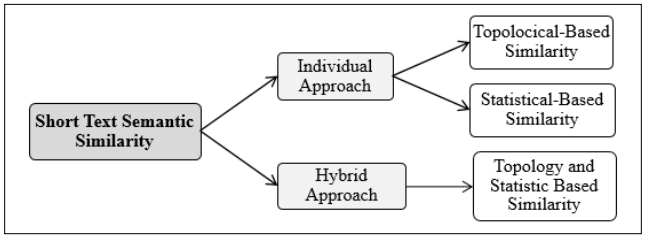

Fig. 1. Outline of STSS approaches

A hybrid semantic similarity is a more recent approach which is composed of a combination of different implementations of STSS measures. The resource of integrated information provided in this paper shall provide insights on the relevant issues and perspectives that should be considered in future proposals, and therefore facilitate the development of future works that aim to contribute to the field of Twitter NLP and social media analysis. Fig. 1 summarizes the similarity approaches studied in this paper.

The remainder of the paper is organized as follows: Section II describes the methods that are used in the review part. Section III describes the three categories of STSS measures under consideration. Section IV discusses the challenges presented in Twitter that hinder the performance of these measures and observations derived from the reviewed approaches. Section V explains the experimental methodology in terms of design, hypothesis, dataset and sample size, feature set, and experiment analysis and evaluation metrics. In section VI, the experiment results and analysis using correlations, Mean Squared Error (MSE) and inferential statistical analysis are presented and explained. Section VII discusses the experiment results and observations taking into consideration the current settings in which the experiment took place. Finally, the conclusion and further directions are provided in Section VIII.

II. Methods

A. Inclusion Criteria

The inclusion criteria for the contributions reviewed in this research are as follows:

1) Contributions to enhance the semantic textual analysis of Twitter short text messages through the development of semantic similarity measures.
2) Contributions to determine latent topics in textual data obtained from Twitter through potential semantic similarity processes for topic modelling, such as Latent Dirichlet Allocation (LDA), which is further elaborated in Section III.B.

III. Short Text Semantic Similarity Measures

STSS measures are generally divided, in terms of their core functionality and attributes, into three categories: topological, statistical, and hybrid.

A. Topology-Based STSS

The semantic similarity between short-texts can be gauged through defining a topological similarity, which is based on using knowledge bases such as ontologies. The distance between terms and concepts are determined by means of these resources. Calculating the topological similarity between ontological concepts can be done either by using the edges and their types (edge-based) or the nodes and their properties (node-based) as data sources. Liu and Wang [15] presented a topological measure for computing the semantic similarity between short texts based on the structural and semantic relationships in a predefined hierarchical concept tree (HCT), without requiring any additional corpus information. A major drawback of this approach is that it does not take into account the word's sequence in which it appears in the sentence. For instance, the sentences *the cat chased the dog* and *the dog chased the cat* would be considered identical.

Another drawback is related to the scalability and performance of the current state-of-the-art semantic measures libraries. The authors in [16] argue that these

drawbacks are due to using naïve graph representation models, which fail to capture the intrinsic structure of the represented taxonomies. Consequently, topological algorithms that are based on naïve models suffer from degraded performance due to demanding high computational cost. This complexity problem is derived from the caching strategy adopted by current semantic measures libraries. This strategy stores all nodes' ancestors and descendants within the taxonomy, which significantly increases memory usage leading to scalability problems concerning the taxonomy size. Moreover, the dynamic resizing of the caching data structures, further memory allocation, or the integration with external relational databases will raise performance issues.

Current state-of-the-art is a new representation model for taxonomies, along with a new software library based on it [16]. This model is claimed to properly encode the intrinsic structures and bridges the aforementioned gaps of scalability and performance. It is an adaptation of the half edge representation in the field of computational geometry [17] in order to represent and interrogate large taxonomies in an efficient manner.

1) Applications of topology-based STSS in Twitter Analysis: Rudrapal et al. [18] proposed a method for measuring the semantic similarity between Bengali tweets using the Bengali WordNet developed by Das and Bandyopadhyay [19]. The Bengali model computes the semantic similarity score of a pair of tweets through the use of a lexical based method. It is built on the basis of analyzing common words similarity among tweets. This approach may be used for English tweets, bearing in mind that Bengali tweets are less noisy in nature compared to English, and therefore requires less comprehensive pre-processing. This is because people tend to use fewer abbreviated words (e.g. "great" instead of "gr8"), character repetition (e.g. "heeeey" for "hey"), etc. in Bengali tweets. Another approach to applying topological STSS which is based on knowledge bases is provided in [20]. The authors utilized the English WordNet ontology [21] to estimate the semantic score between microblogs and recommended the top similar microblog records to the user. In their approach, the authors computed the similarity between sentences based on the similarity of the pairs of words contained in the corresponding sentences. Furthermore, the semantic similarity between two word senses is captured through path length, in which the taxonomy is treated as an undirected graph and the distance is calculated between them based on WordNet. The performance of this approach was compared to a statistical based approach, which will be presented and discussed in Section III.B. Findings suggested that this topological-based approach performed better than the statistical-based one in terms of precision. Further research aimed at comparing the performance of several models for determining topic coherence in relation to a Twitter dataset with human assessments has been conducted in [22]. Among the utilized models, the approach employed an individual thesaurus and corpus based measures to determine the semantic similarity between terms within extracted topics from the Twitter dataset. The topics were identified through Latent Dirichlet Allocation (LDA) (described further in Section III.B) and each topic was represented by the top ten words ranked according to their probabilities in the term distribution. Any two words from these top ten form word pairs of a topic and the topic coherence is measured by averaging the semantic similarity of all word pairs in that topic. In this approach, the semantic similarity was computed by using individual measures on WordNet and statistical measures on Wikipedia and a Twitter corpus containing 30,151,847 processed tweets. Three path length based methods were used to calculate the lexical similarity between words in WordNet, LCH [23], JCN [24], and LESK [4]. LCH finds the shortest path between concepts in WordNet. This path length is then scaled by the maximum length observed in the "is-a" hierarchy, in which the two concepts occur. JCN, on the other hand, includes the information of the least common subsumer in addition to the shortest path length. Finally, LESK incorporates information from WordNet glosses, where it finds overlaps between the glosses of the two concepts under consideration, in addition to the concepts that directly link to them. This WordNet based approach will be referred to in the subsequent section, where comparisons are made.

B. Statistical-Based STSS

Statistical approaches determine the semantic similarity between short texts through calculating words co-occurrence frequencies based on a large corpus of text. Deerwester et al.'s Latent Semantic Analysis (LSA) is the prominent statistical-based semantic similarity measure, which is provided as a method for information retrieval [25]. LSA, which is sometimes referred to as Latent Semantic Indexing (LSI), is based on the distributional hypotheses that words similar in meaning will occur in similar contexts [26]. Therefore, calculating word similarity can be derived from a statistical analysis of a large text corpus. The set of unique terms and documents (short-texts in this context) in the corpus are used to generate a high dimensional matrix of terms occurrences. This term-document matrix is commonly decomposed by the application of a matrix factorization algorithm such as Singular Value Decomposition (SVD). The incorporation of SVD into LSA reduces the dimensionality of the single frequency matrix through approximating it into three sub matrices, term-concept matrix, singular value matrix, and concept-document matrix. The SVD process in LSA preserves the important semantic information while reducing noise presented in the original space. It has been found that SVD has improved the effectiveness of word similarity measures [27].

LDA is a semantic topic extraction model that is based on probabilities [28]. LDA is a significant extension of LSA, where terms are grouped into topics, in which most of these terms exist in more than one topic [29]. Despite the commonalities between LDA and LSA, each of the

algorithms generate distinct models. While LSA uses SVD in which the maximum variance across the data is determined for a reduced number of dimensions, LDA employs a Bayesian model. This model considers each document as a mixture of underlying topics and every topic is modeled as a mixture of term probabilities from a vocabulary. Moreover, even though LDA and LSA outputs may be used in similar scenarios, the values of their outputs represent completely different quantities, with different ranges and meanings. LSA generates term by concept and document by concept correlation matrices, with values ranging between -1 and 1 with negative values denoting inverse correlations. On the other hand, LDA generates term by topic and document by topic probability matrices, in which probabilities range from 0 to 1. LDA has an advantage over LSA, which is its ability to tackle the problem of disambiguation and therefore has higher accuracy. This is done by comparing a document to two topics and determining which of them is closer to the document, across all combinations of topics that seem broadly relevant. This direct interpretation of similarities and differences between the most effective statistical semantic measures is important for the challenging process of understanding which measure may be most appropriate for a given text analysis task.

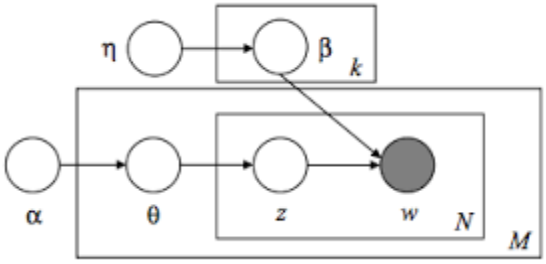

Fig. 2. LDA graphical model [28]

In recent years, there has been an increase in approaches proposing to compose word vectors by using neural language models, which have a core of trained neural networks [30]. Given a sequence of initial words, early neural models were designed to predict the next word in the sentence [31] (e.g. text input auto-completion). While these models can be trained with a variety of techniques to achieve different tasks, they share a common feature of having at their core a dense vector representation of words that can be exploited for computing similarity. This representation is commonly referred to as "neural word embedding", in which their effectiveness varies with regard to the chosen technique and corpus for similarity computation.

1) Applications of statistical-based STSS in Twitter analysis: Steiger et al. used LDA to assess the semantic similarity among tweets [32]. A corpus of 20.4 million processed tweets was created as the lexical resource for which LDA performed its semantic probabilistic model. The application of LDA reduced the semantic dimensions through clustering co-occurring words into topics. Each topic is referred to by labeling it with the highest probability associated words (>0.03). In their adopted approach of LDA, Steiger et al. assumed each tweet α contains a random number of topics, and each topic is characterized by a word distribution β (see Fig. 2). For an individual word w within each tweet, z is the corresponding associated topic. The topic distribution for the overall number of tweets M is denoted by θ, each being of length N. The main challenges encountered, were the estimation of the posterior parameter and the computation of variables such as the number of topics k. However, this study has several limitations that need to be further addressed. Some pitfalls within the bag-of-words (BOW) assumption of LDA caused words to be assigned to various topics while they should be associated with the same topic. Moreover, taking into consideration the syntactical structure (e.g. n-grams) would allow for word orders to be associated to several topics, and therefore better handle semantic complexities. Further, this study did not include the author-topic model [33] (i.e. all tweets of the same user are treated as a single document) due to missing benchmarking process.

Another study that used LDA to gauge the semantic similarity in the context of Twitter data, includes the work presented in [20], in which a corpus of 548 tweets is used. In this approach, each tweet (microblog) is represented as a topic vector, and consequently, the similarity calculation between tweets is equal to the dot product of the two corresponding topic vectors. This statistical method of assessing the semantic similarity was evaluated and compared to the performance of the topology based approach explained earlier in Section III.A. The results showed that the topological-based approach performed better than the topic-based one in terms of precision.

LSA and Pointwise Mutual Information (PMI) statistical approaches were used on Wikipedia and a background dataset of tweets as corpora. SVD was applied to reduce LSA space to 300 dimensions. The empirical evaluation showed that the PMI based measure using Twitter corpus worked better than PMI using Wikipedia, and it best matched the human ground truth ranking of topic coherence on Twitter among all semantic similarity measures used. This might be due to the generic and formal nature of Wikipedia that may prevent capturing specific terms and trends used in Twitter.

C. Hybrid-Based STSS

Some of the topological methods of estimating the semantic similarity may incorporate a statistical function of term frequency in a corpus in order to determine the value of a concept [34-38]. However, their fundamental component of determining the degree of semantic equivalence remains based on a predefined ontology. The similarity computation might also be composed of a combination of statistical and topological methods.

STASIS [35] is an effective measure that estimates the semantic similarity between short sentences based on topological information derived from WordNet ontology and

statistical information obtained through the use of the Brown corpus [39]. This measure calculates the overall semantic score of similarity between two sentences based on a function of multiple factors. These factors include the path between two synsets in the ontology, depth of the subsumer in the hierarchical semantic nets, and information content derived from the Brown corpus. STASIS forms a word order vector composed of unique words contained in both sentences. The combination of syntactic word order and semantic information determines the overall similarity. Although the proposed method does not consider word sense disambiguation for polysemous words as this would scale up the measure's complexity, it still performs well as per the experimental results.

During the last few years, many state-of-the-art STSS approaches have used linear combinations of measures. For example, six topology-based and two statistical-based measures were tested in [40], for the related task of paraphrase identification. In this work, the efficacy of applying topological-based word similarity measures was explored in comparison to texts. They reported that the two approaches are comparable to corpus-based measures such as LSA. The authors of [41] proposed a method that uses a combination of mandatory (string and semantic word) and optional (common word order) similarities. Evaluated on a dataset of 30 sentence pairs, this method outperformed the correlation obtained in [35]. Moreover, a hybrid approach was proposed in [34] where the authors combined a statistical-based semantic relatedness measure over the complete sentence in addition to a topology-based semantic similarity scores that were computed for the words that share similar syntactical role labels in both sentences. These calculated scores performed as the features that were fed to machine learning models such as BOW to predict a single similarity score given two sentences. Results of this method showed a significant improvement of a hybrid measure compared to corpus-based measures taken alone. UKP (Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures) [38], is a similarity detection system that showed reasonable correlation results. It implemented a string similarity, a semantic similarity, and text expansion mechanisms and measures related to structure and style. These multiple text similarity measures were combined through the use of a simple regression model based on training data.

1) Applications of hybrid-based STSS in Twitter Analysis: Das and Smith presented an approach for measuring the semantic similarity between pairs of tweets through identifying whether the two hold a paraphrase relationship [36]. The probabilistic model incorporates syntax and lexical semantics to compute the similarity between two sentences by using a logistic regression model, with eighteen features based on n-grams. The system builds a binary classification model for identifying paraphrase through using precision, recall, and F1-score of n-gram tokens from sentence pairs. The model is capable of determining whether there exists a semantic relationship between a pair of tweets. However, it may be improved by principled combination with more standard lexical approaches.

SemSim is a hybrid based semantic textual similarity system, composed of several modules designed to handle the automatic computation of the degree of equivalence between pieces of multilingual short-text *[37]*. The system was developed to handle general short texts segments and has been tested on a tweets dataset. The system is composed of a module for calculating the semantic similarity of words and another one for pairs of short-text. The former is the core of the system that computes the semantic similarity based on a combination of HAL and WordNet. The semantic textual similarity module uses the semantic word similarity model to calculate the similarity between pairs of short-text. Keywords similarities are calculated through the word similarity module after aligning multiple terms in one sentence to a single term in the other sentence. The words are then paired and the overall similarity score is computed through the semantic textual similarity (STS) module. Generally, SemSim demonstrated a good performance in terms of correlation, but performed poorly in the case of tweets. This is attributed to the absence of some words in the vocabulary, and the top definitions of other words are not always reliable as they may be less prominent.

This section highlighted current state-of-the-art algorithms to distinguish areas of improvement and stimulate creativity towards the development of new approaches. **RQ1** has been explored through discussing settings and features of the aforementioned algorithms in the context of Twitter text analysis. To the best of our knowledge, STSS measures have not been previously reviewed with regard to social media data. Tackling **RQ1** paves the way towards **RQ2** which investigates weaknesses of applying current STSS measures on the noisy and challenging social data and calls for improvement in research and practice. These challenges and weaknesses are further emphasized in the subsequent section.

IV. STSS Challenges in Twitter

One of the most difficult aspects of NLP is to establish the understanding and reasoning of the underlying meaning of the text. The challenge of measuring the semantic similarity increases when there is a reduced quantity and quality of text. In terms of social media data, particularly Twitter, the task becomes much harder due to many inaccuracies that may be present in the short pieces of text. These inaccuracies include:

1) Poor grammatical and syntactical structure due to the character limit which encourage the frequent use of abbreviations and irregular expressions [9].
2) Misspellings, out-of-vocabulary words, and acronyms.
3) Lots of redundant information as people tend to repost some original messages.
4) Conventions such as hashtags and other metadata that may interrupt the potential meaning in a text.

Due to these inaccuracies, computers face difficulties in understanding the intended meaning or associating the semantic similarity between pairs of tweets. This is especially true in a tweet which expresses sarcasm, such as "*I enjoy waiting forever for my appointment*", which is common in social media. Therefore, the automation of this process through computation is a challenging task as there are general conventions (hashtags, mentions, URLs, and etc.) and improper English, such as spelling mistakes (e.g. *bcuz* instead of *because*), shared on this communication platform. Many approaches to STSS measures have been based upon adaptation of existing document similarity methods of general English, with no comprehensive consideration of the language used in Twitter. As such, these methods are less applicable to the problem domain of Twitter analysis.

Several key points with regards to the challenges of the STSS approach in social media datasets, particularly Twitter, have been observed:

1) Topological-based approaches use ontologies to capture the semantic similarity between concepts. These approaches often demonstrate scalable and acceptable performance, however, when applied in the context of social media, their performance degrades. This is due to the informal terms used in these sites that are absent from these English dictionaries. To minimize this problem, some approaches suggest using external informal dictionaries for dealing with out-of-vocabulary tokens.
2) Statistical-based methodologies are not effective for measuring the semantic similarity for short and sparse text as they are for long and rich text. However, they tend to perform better when the utilized corpus consists of the same domain than the case of general corpus, such as the Brown corpus. This is due to the fact that these corpora contain information from traditional media and therefore may fail to capture specific terms and trends dynamically propagated through social media networks.
3) Although not many hybrid based systems were developed for the intended approach, it can be observed that these approaches outperform single measures of determining the semantic similarity between short segments of texts. However, they tend to consume high computational resources.

Moreover, it has been observed that a robust pre-processing and feature extractor function that is able to normalize and extract Twitter specific text features may significantly improve the performance of STSS measures in the context of social media data [42], [43], [11].

V. Experiment Methodology

As demonstrated in Section III, STSS measures differ according to their core body of components and functionality. Therefore, an experiment was designed and implemented in order to evaluate the validity of different semantic versus non-semantic STSS when applied in the context of Twitter OSN. These experiments require a dataset that is subjectively annotated with human ratings of the actual similarity score by a predefined class of annotators. Part of the SemEval-2014 shared task comprises a published annotated news tweets training and testing dataset [44]. A corpus of the training data was built for weighting the terms and for the statistical analysis performed by LSA.

This section describes the experiment conducted to evaluate the level of effectiveness of the measures explained in Section III. The results of the measures were normalized as each measure scores on different scale. The empirical evaluation of the measures were made through several statistical analysis and tests in order to answer **RQ3**. These are further elaborated in the subsequent sections.

A. Hypothesis

The hypothesis to be tested relates to the accuracy of the similarity measure compared to typical human cognition similarity assessment, which is as follows:

$H0_a$ - *The similarity measure deployed can accurately approximate human cognition of semantic interpretation. That is, there is no statistically significant difference between the actual (human) and predicted (measure) values.*

$H0_b$ - Actual and predicted values are numerically close.

$H1_a$ - *The similarity measure is unable to produce a relatively accurate similarity judgment. That is, there is a statistically significant difference between the actual (human) and predicted (measure) values.*

$H1_b$ - Actual and predicted values are numerically not close.

B. Experiment Design

An implementation of the measures under consideration was developed and the outcome was evaluated against a benchmark. The experiment carried out was set to test the correlation between the similarity scores of the human judges and results of the implemented measures. The experimental analysis outcome will provide insights on the direction and potential measure improvement that can be addressed through further research.

The effectiveness of the designed experiment is tested through a representative random sample of the SemEval-2014 dataset. The analysis of the experiment results will be used in further research towards approximating human cognition in similarity assignment and adjusting features and measure's parameters to maximize its accuracy.

C. Dataset and Sample Size

SemEval-2014 is a collection of computational semantic analysis tasks intended to explore the nature of meaning in language. It carried out several semantic tasks, including evaluation of compositional distributional semantic measures through entailment and multilingual semantic textual similarity in Twitter. Multiple datasets were published for

system training and testing in order to unify the evaluation and allow for a fair comparison of all contributions. However, as this experiment is aimed at evaluating the capability of a measure to capture the semantic between pairs of tweets, it is necessary have a dataset that is labelled with human ratings. Part of the published trial datasets is a tweet-news dataset containing 750 annotated pairs [44]. The gold standard implements a 5-point Likert scale to interpret the degree of similarity between pairs, as defined by Agirre [45].

D. Experiment STSS Measures

1) Weighted keyword-based similarity: The first implemented similarity approach is based on shared keywords rather than semantic similarity. Given the corpus that was generated from the evaluation dataset, each document (tweet) is represented by a vector of weighted terms in that corpus. Each term is then represented by the number of its occurrences in the document multiplied by its frequency of occurrence in the whole corpus as in

$$tf - idf_{t,d} = \mathrm{tf}_{t,d} \cdot \log \frac{N}{\mathrm{df}_t} \qquad (1)$$

Where, $tf_{t,d}$ is the total number of occurrences of t in d, df_t is the total number of documents containing i, and N is the total number of documents in the corpus. Finally, the cosine of the two vectors (representation of the two short-texts under consideration) yields the similarity value.

2) LSA: Several statistical-based similarity measures have been reviewed and LSA was nominated as it has been reported to outperform LDA in a system that measures the similarity between movies based on their metadata [46]. Although the movies dataset is different than a dataset of tweets, it might uncover potential insights as both datasets share mutual prominent factor, which is the short-text content. There has not been found any equivalent or similar study that was performed on a Twitter dataset.

3) STASIS: STASIS is selected as it accounts for word order as part of its system components. STASIS assigns the similarity score based on a combination of the syntactic and semantic ratio of similarity. Hence, it may have potential capabilities for the domain under consideration. However, this measure was tested on a dataset of short formal English sentences that utilizes WordNet and the Brown corpus, whereas the data under consideration has lots of informality and out of dictionary terms. Therefore, it is necessary to determine and evaluate its applicability through experiments.

E. Feature Set

A feature extractor module has been implemented to parse the text input and generate a set of features that represents the given tweet. In the conducted experiment, the input was represented by the set of weighted unigrams that are presented in a tweet, which are non-function words. The term weights were calculated according to (1).

F. Experimental Analysis and Evaluation Metrics

The data gathered from each run was collected and subsequently analyzed to explore the findings from the experiment. The experiment results are evaluated through several measures to ensure that they are thoroughly analyzed. These measures include the Pearson correlation coefficient, Spearman's rank correlation coefficient, MSE, and a statistical hypothesis test. These are further elaborated in Section VI.

VI. Experiment Results and Analysis

This section discusses the result of the evaluation metrics.

A. Rational for the Selection of Evaluation Measures

Correlation coefficient: Pearson correlation has been a common practice for assessing the performance of STSS systems through computing the correlation between human judgments and machine assigned semantic similarity scores [1]. Systems that record higher correlations are generally considered "accurate", and would often be among the top choices for the system designer of an STSS based evaluation task. However, this common practice of STSS evaluation through Pearson correlation has been questioned previously.

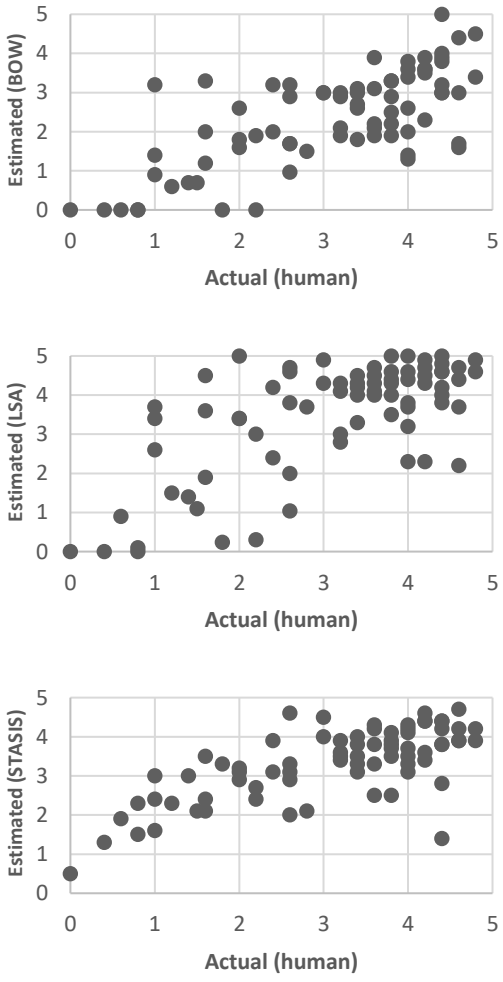

Fig. 3. Correlation scatterplots between actual and estimated values

Zesch [47], reported several limitations of the Pearson correlation,

1) Sensitive to outliers.
2) Limited to measuring linear relationships.
3) The two variables need to be approximately normally distributed.

Zesch recommended the usage of Spearman's rank ρ correlation coefficient as it is not sensitive to outliers, non-linear relationships, and non-normally distributed data. However, most evaluation methods of STSS systems only report the Pearson correlation. Nevertheless, the experiment results were evaluated via computing both Pearson and Spearman's correlation coefficient to avoid uncertainty.

Although Pearson and Spearman's tend to perform different calculations, both outcomes are interpreted in the same way that is mentioned above. Correlation scatterplots between the measures and human annotations are shown in Fig. 3, where each point represent a pair in the dataset.

1) MSE: Agirre [1] mentioned in SemEval-2013 discussion: "*Evaluation of STS is still an open issue*" and in addition to the Pearson correlation, *"...other alternatives need to be considered, depending on the requirements of the target application*". Therefore, it is reasonable to compute the average error rate between the actual and estimated values, and assess the STSS measures accordingly.

TABLE I. TEST SET RESULTS ON SEMEVAL-2014

Measure	*r*	*ρ*	*MSE*
Weighted_BOW	0.7102	0.6517	1.4009
LSA	0.6753	0.5692	1.3304
STASIS	0.7086	0.6567	0.8168

The least MSE results are the closest to human judgments. The results on the SemEval-2014 dataset with gold standards are summarized in Table 1, showing Pearson's *r*, Spearman's *ρ*, and MSE.

B. Statistical Test

Selecting an appropriate statistical technique for testing the hypothesis is the most difficult part when conducting research [48]. This is attributed to the lack of a universal methodology that clearly guides researchers on the right statistical test choice [49]. The challenge of this choice refers to the variations in the nature of research, as it depends on the type of research questions that need to be addressed. In terms of the STSS measures, it also depends on the scale of similarity assignment, the variables to be analyzed, the underlying assumptions for specific statistical techniques, and the nature of the data itself [48].

Parametric tests are inferential statistical analysis based on assumptions regarding the population and require numerical score [50]. Non-parametric techniques do not employ such strict requirements nor do they make distribution assumptions, and therefore sometimes referred to as distribution free tests. These tests are most often used with categorical and ordinal data as they do not require the data to be normally distributed and are not based on a set of assumptions about the population [51].

The "Test of normality" is investigated to test the distribution of the data. It is generally agreed that significant values greater than 0.05 indicate that the data is similar to a normal distribution, otherwise it is not normally distributed.

TABLE II. TEST OF NORMALITY

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.(p)	Statistic	df	Sig.(p)
Human	.145	75	.000	.924.	75	.000
BOW	.125	75	.006	.963	75	.028
LSA	.188	75	.000	.840	75	.000
STASIS	.105	75	.039	.946	75	.003

Table 2 presents the results of the normality test. As the data is not normally distributed, a nonparametric test will be utilized for the data analysis. Hence, the Wilcoxon Signed Rank Test will be used to test the hypothesis. This test is the nonparametric alternative to the repeated measure *t*-test, however, Wilcoxon converts scores to ranks and compares them instead of comparing the means of the two systems under study. It can be concluded that the differences between the two scores is statistically significant, if the significance level (p-value) is equal to or less than .05 [48].

In addition to classifying the data in terms of normality, inferential statistical analysis tests were carried out to investigate whether the similarity results obtained from each measure are any close to human judgments.

C. Inferential Statistical Analysis

Wilcoxon Signed Rank was used to test the following hypothesis:

H0$_a$: $\mu d = 0$ (No significant difference between the actual and measured values)

H1$_a$: $\mu d \neq 1$ (Significant difference between the actual and measured values)

1) Hypothesis Result: A Wilcoxon Signed Rank test was established on each measure paired with the gold standard, were actual refers to human judgments and estimated refers to similarity measurements.

TABLE III. WILCOXON SIGNED RANK TEST RESULTS

Test Statistics			
Actual	**Predicted**	**Z**	**Asymp. Sig.**
Human annotation	Weighted_BOW	-5.633	.000
	LSA	-3.125	.002
	STASIS	-2.320	.020

The results demonstrated that for each of the similarity measures tested to evaluate the accuracy of the measures in the context of Twitter short-text, there is a statistically significant difference (p-value < 0.05) between the similarity obtained by the measures and the gold standard (accept **H1$_a$** and reject **H0$_a$**). Consequently, this means that the actual and predicted values are numerically not close (accept **H1$_b$** and reject **H0$_b$**). The results of the statistical analysis are present in Table 3. The evaluation methods are further discussed in Section VII.

VII. Discussion

The goal of the evaluation criteria utilized to gauge the performant of the STSS measures are twofold. The first part involved employing metrics to assess and compare the accuracy between measures under investigation in relation to the gold standard. Whereas the next part involved performing an inferential statistical analysis to test how close are the measures to human judgment.

The evaluation using Pearson correlation demonstrated the highest result for the weighted BOW (0.7102) and the lowest for LSA (0.6753). However, these results might not be reliable as the data contained outliers, such as a tweet that is composed of two words or even one, in which Pearson correlation is sensitive. Therefore, the correlations were better represented using Spearman's rank, which employs rankings instead of the actual scores. The results on the SemEval-2014 dataset based on Spearman's showed that there is no strong correlation for the three measures; however, STASIS and the weighted BOW approach were more correlated to human judgments than LSA, with STASIS slightly higher. However, the intrinsic common evaluation based on only correlation in the differentiation between STSS systems might be ill suited as mentioned earlier in Section VI. Therefore, the need of an additional evaluation measure has led to calculating the MSE in order to find out which one had the least error rate. STASIS had an average error of 0.8168, LSA 1.3304, and weighted BOW recorded 1.4009 when compared with the gold standard. It can be concluded that the semantic-based measures performed better than the keyword-based, although LSA was not substantially less than the weighted BOW (0.1), but STASIS was less by 0.6.

The inferential analysis revealed negative statistics not only for the keyword-based approach, but also for the statistical and for hybrid based approaches. The Wilcoxon Signed Rank test showed that there is a significant difference between the similarity scores obtained by the three measures, and the gold standard. This is attributed to the dataset that these measures were applied to. While the evaluated measures may be effective in approximating the human ratings in different settings of short-text data, it is evident that the challenges present in Twitter language (discussed in section IV) are hampering the accuracy and effectiveness of these measures. These require further research to enhance the performance of the semantic similarity measure.

The analysis of the results are useful in guiding further work of measure adaptation to deal with the textual challenges present in Twitter. This can be achieved through examining cases where the measure performed poorly and adjusting parameters, such as redesigning the feature set in a way that had better capture a tweet's semantical structure.

VIII. Colclusion and Future Work

This paper presents the work conducted to address the research questions provided in Section I.B. The evaluation of different STSS measures revealed insights for the development of new STSS measures to overcome the weaknesses of existing ones in capturing the semantics of Twitter data.

The experimental results showed evidence that, although the evaluated measures may produce high correlations when dealing with proper English text, the nature of most short-textual data propagated in social media, are hindering the performance of these measures. Thus, it is imperative to adapt the components of such measures in a way that can understand the modern natural language generated in Twitter. This is particularly useful for applications of Machine Learning handling social media data.

Towards proceeding with future research, the preliminary evaluation revealed key information regarding the accuracy of STSS measures compared to a non-semantic based measure in the context of Twitter data. The main observations are summarized as follows:

- The features used in the implemented experiment are not adequate to handle the challenges presented in the language and structure of Twitter data, and therefore additional preprocessing and features need to be utilized.
- Semantic-based measures performed better than the keyword-based measure in detecting the degree of semantic equivalence between pairs of tweets.
- While STASIS performed better than LSA, they are both potential contenders for estimating the semantic similarity between tweets and therefore require further investigation, as some of their components may be integrated and utilized for developing a Twitter-specific semantic similarity measure.

Further research continue on towards determining new methodologies for adapting and developing scalable and robust STSS measures that can handle the unstructured and noisy microblogging data.

References

[1] Agirre, E., et al. *SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation*. in *SemEval@ NAACL-HLT*. 2016.

[2] O'Shea, J., et al., *A comparative study of two short text semantic similarity measures.* Agent and Multi-Agent Systems: Technologies and Applications, 2008: p. 172-181.

[3] Rocchio, J.J., *Relevance feedback in information retrieval.* The SMART retrieval system: experiments in automatic document processing, 1971: p. 313-323.

[4] Lesk, M. *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone*. in *Proceedings of the 5th annual international conference on Systems documentation*. 1986. ACM.

[5] Salton, G. and C. Buckley, *Term-weighting approaches in automatic text retrieval.* Information processing & management, 1988. **24**(5): p. 513-523.

[6] O'Shea, K., Z. Bandar, and K. Crockett, *A conversational agent framework using semantic analysis.* International Journal of Intelligent Computing Research (IJICR), 2010. **1**(1/2).

[7] Lin, Y.-S., J.-Y. Jiang, and S.-J. Lee, *A similarity measure for text classification and clustering.* IEEE transactions on knowledge and data engineering, 2014. **26**(7): p. 1575-1590.

[8] Albitar, S., S. Fournier, and B. Espinasse. *An effective TF/IDF-based text-to-text semantic similarity measure for text classification*. in *International Conference on Web Information Systems Engineering*. 2014. Springer.

[9] Alnajran, N., et al. *Cluster Analysis of Twitter Data: A Review of Algorithms*. in *9th International Conference on Agents and Artificial Intelligence*. 2017. SCITEPRESS.

[10] Farzindar, A. and D. Inkpen, *Natural language processing for social media.* Synthesis Lectures on Human Language Technologies, 2017. **10**(2): p. 1-195.

[11] Gómez-Adorno, H., et al., *Improving feature representation based on a neural network for author profiling in social media texts.* Computational intelligence and neuroscience, 2016. **2016**: p. 2.

[12] Guo, W. and M. Diab. *A simple unsupervised latent semantics based approach for sentence 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. Association for Computational Linguistics.

[13] Zanzotto, F.M., M. Pennacchiotti, and K. Tsioutsiouliklis. *Linguistic redundancy in twitter*. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2011. Association for Computational Linguistics.

[14] Xu, W., A. Ritter, and R. Grishman. *Gathering and generating paraphrases from twitter with application to normalization*. in *Proceedings of the sixth workshop on building and using comparable corpora*. 2013.

[15] Liu, H. and P. Wang, *Assessing Text Semantic Similarity Using Ontology.* JSW, 2014. **9**(2): p. 490-497.

[16] Lastra-Díaz, J.J., et al., *HESML: A scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset.* Information Systems, 2017. **66**: p. 97-118.

[17] Botsch, M., et al., *Openmesh-a generic and efficient polygon mesh data structure.* 2002.

[18] Rudrapal, D., A. Das, and B. Bhattacharya. *Measuring Semantic Similarity for Bengali Tweets Using WordNet*. in *Proceedings of the International Conference Recent Advances in Natural Language Processing*. 2015.

[19] Das, D. and S. Bandyopadhyay. *Developing Bengali WordNet affect for analyzing emotion*. in *International Conference on the Computer Processing of Oriental Languages*. 2010.

[20] Chen, X., et al. *Recommending Related Microblogs: A Comparison Between Topic and WordNet based Approaches*. in *AAAI*. 2012.

[21] Miller, G.A., *WordNet: a lexical database for English.* Communications of the ACM, 1995. **38**(11): p. 39-41.

[22] Fang, A., et al. *Topics in tweets: A user study of topic coherence metrics for Twitter data*. in *European Conference on Information Retrieval*. 2016. Springer.

[23] Leacock, C. and M. Chodorow, *Combining local context and WordNet similarity for word sense identification.* WordNet: An electronic lexical database, 1998. **49**(2): p. 265-283.

[24] Jiang, J.J. and D.W. Conrath, *Semantic similarity based on corpus statistics and lexical taxonomy.* arXiv preprint cmp-lg/9709008, 1997.

[25] Deerwester, S., et al., *Indexing by latent semantic analysis.* Journal of the American society for information science, 1990. **41**(6): p. 391-407.

[26] Harris, Z.S., *Mathematical structures of language.* 1968.

[27] Landauer, T.K. and S.T. Dumais, *A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.* Psychological review, 1997. **104**(2): p. 211.

[28] Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation.* Journal of machine Learning research, 2003. **3**(Jan): p. 993-1022.

[29] Crossno, P.J., et al. *Topicview: Visually comparing topic models of text collections*. in *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on*. 2011. IEEE.

[30] Christoph, L., *Measuring Semantic Similarity and Relatedness with Distributional and Knowledge-based Approaches.* 2016.

[31] Mnih, A. and G.E. Hinton. *A scalable hierarchical distributed language model*. in *Advances in neural information processing systems*. 2009.

[32] Steiger, E., et al., *Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data.* Computers, Environment and Urban Systems, 2015. **54**: p. 255-265.

[33] Zhao, W.X., et al. *Comparing twitter and traditional media using topic models*. in *European Conference on Information Retrieval*. 2011. Springer.

[34] Aggarwal, N., K. Asooja, and P. Buitelaar. *DERI&UPM: Pushing corpus based relatedness to similarity: Shared task system description*. 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. Association for Computational Linguistics.

[35] Li, Y., et al., *Sentence similarity based on semantic nets and corpus statistics.* IEEE transactions on knowledge and data engineering, 2006. **18**(8): p. 1138-1150.

[36] Das, D. and N.A. Smith. *Paraphrase identification as probabilistic quasi-synchronous recognition*. in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. 2009. Association for Computational Linguistics.

[37] Kashyap, A., et al., *Robust semantic text similarity using LSA, machine learning, and linguistic resources.* Language Resources and Evaluation, 2016. **50**(1): p. 125-161.

[38] Bär, D., et al. *Ukp: Computing semantic textual similarity by combining multiple content similarity measures*. 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. Association for Computational Linguistics.

[39] Francis, W.N. and H. Kucera, *Brown corpus.* Department of Linguistics, Brown University, Providence, Rhode Island, 1964. **1**.

[40] Mihalcea, R., C. Corley, and C. Strapparava. *Corpus-based and knowledge-based measures of text semantic similarity*. in *AAAI*. 2006.

[41] Islam, A. and D. Inkpen, *Semantic text similarity using corpus-based word similarity and string similarity.* ACM Transactions on Knowledge Discovery from Data (TKDD), 2008. **2**(2): p. 10.

[42] Duong, P.H., H.T. Nguyen, and N.-T. Huynh. *Measuring Similarity for Short Texts on Social Media*. in *International Conference on Computational Social Networks*. 2016. Springer.

[43] Demirsoz, O. and R. Ozcan, *Classification of news-related tweets.* Journal of Information Science, 2016: p. 0165551516653082.

[44] Guo, W., et al. *Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media*. in *ACL (1)*. 2013.

[45] Agirre, E., et al. *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. Association for Computational Linguistics.

[46] Bergamaschi, S. and L. Po. *Comparing LDA and LSA Topic Models for Content-Based Movie Recommendation Systems*. in *International Conference on Web Information Systems and Technologies*. 2014. Springer.

[47] Zesch, T., *Study of semantic relatedness of words using collaboratively constructed semantic resources*. 2010, Technische Universität.

[48] Pallant, J., *SPSS survival manual*. 2013: McGraw-Hill Education (UK).

[49] Kinear, P.R. and C.D. Gray, *SPSS for Windows made simple: release 10*. 2001: Psychology Press.

[50] Gravetter, F.J. and L.B. Wallnau, *Statistics for the behavioral sciences*. 2016: Cengage Learning.

[51] Nolan, S.A. and T. Heinzen, *Statistics for the behavioral sciences*. 2011: Macmillan.