2015 IEEE 27th International Conference on Tools with Artificial Intelligence

Lexical Semantic Relatedness for Twitter Analytics

Yue Feng[1], Hossein Fani[1,2], Ebrahim Bagheri[1], Jelena Jovanovic[3]
[1]*Laboratory for Systems, Software and Semantics* (LS^3), *Ryerson University*
[2]*University of New Brunswick,* [3]*University of Belgrade*

***Abstract*—Existing work in the semantic relatedness literature has already considered various information sources such as WordNet, Wikipedia and Web search engines to identify the semantic relatedness between two words. We will show that existing semantic relatedness measures might not be directly applicable to microblogging content such as tweets due to *i)* the informality and short length of microblogging content, which can lead to shift in the meaning of words when used in microblog posts, *ii)* the presence of non-dictionary words that have their semantics defined/evolved by the Twitter community. Therefore, we propose the Twitter Space Semantic Relatedness (TSSR) technique that relies on the latent relation hypothesis to measure semantic relatedness of words on Twitter. We construct a graph representation of terms in tweets and apply a random walk procedure to produce a stationary distribution for each word, which is the basis for relatedness calculation. Our experiments examine TSSR from three different perspectives and show that TSSR is better suited for Twitter analytics compared to the standard semantic relatedness techniques.**

***Keywords*-Semantic Relatedness, Microblogging, Twitter, Random Walk, Information Retrieval, Semantic Similarity**

I. Introduction

Semantic relatedness is defined as any form of lexical or functional association between two words that points to contextual or semantic similarity of those two words regardless of their syntactical differences [1]. Semantic relatedness measures are increasingly used in information retrieval to augment syntactical matching between the query and document spaces with semantic comparison and matching that may reveal relatedness of non-syntactically-related content [2]

Researchers have already used many different information and knowledge sources in order to compute semantic relatedness between two words. These sources include WordNet, Wikipedia, and Google search results, just to name a few. For instance, Gabrilovich and Markovitch [3] have developed a widely-used semantic relatedness method, known as Explicit Semantic Analysis (ESA), which measures semantic relatedness of words based on their co-occurrence in the same Wikipedia articles, while the method proposed by Cilibrasi et al. [4] for computing semantic relatedness considers the frequencies of page overlaps within Google search results for the given pairs of query keywords.

Empirical research has already shown that results of many of these semantic relatedness techniques have reasonable correlation with subjective interpretation of relatedness of two words [5]. In other words, given that these techniques rely predominantly on stable information sources, such as encyclopedic information on Wikipedia, to draw their conclusions regarding the relatedness of words, they are able to closely model the most common underlying semantic relatedness of two words. For this reason, existing semantic relatedness techniques can be effectively applied in traditional information retrieval tasks such as finding information in Web pages and News articles. However, with the emergence of popular microblogging services such as Twitter that have unique characteristics, e.g. short length and informality, semantic relatedness measures need to be revisited to make them suitable for information retrieval tasks in such contexts.

In our empirical work, we observed that word interpretation and usage can be different depending on the communication medium, i.e., people tend to use the same words to convey different meaning depending on whether they are using them on Twitter or Wikipedia. To illustrate this Table I lists five pairs of words that have been found to be highly related by the ESA technique given that they were frequently seen together on Wikipedia; however, they were never seen together in our Twitter dataset with 10 million tweets. For instance, according to the ESA metric, the words *precedent* and *law* are highly related, but out of the 2,135 and 94,185 times that these two words were observed in our dataset, they never co-occurred in our Twitter dataset. A similar trend can be observed in Table II where the words that are not highly semantically related by ESA were highly correlated in Twitter. For instance, the words *movies* and *popcorn* appeared very frequently together in our Twitter dataset whereas their ESA-based semantic relatedness is far from being high. These observations directly lead to the rationale for the work reported in this paper:

1. The communication and writing style on Twitter are quite different from traditional communication media, in that it promotes short and informal communication. This leads to the appearance of many new words that do not necessarily have explicit linguistic semantics, e.g., tweetup and attwaction.
2. Tweets often include hashtags that further qualify the purpose that the user intended to convey. Many of the hashtags have emerged through an informal social exchange and the semantics are only understood

1082-3409/15 $31.00 © 2015 IEEE
DOI 10.1109/ICTAI.2015.41

within the context of the communities on Twitter that use them, e.g. #baddayattheoffice and #shareforshare.
3. The meaning of some words can shift when used in informal Twitter conversations; therefore, the meaning of a word when used on Twitter would be different from its meaning in regular usage, e.g. Yoyo is a popular playing object (toy), but when used on Twitter, the word refers to "you're on your own".

Table I
SAMPLE WORD PAIRS WITH HIGH RANK IN ESA, LOW RANK IN TSSRPAIRS ARE FROM THE WORDSIMILARITY-353 COLLECTION AND THE RANKS ARE THE ESTIMATED SIMILARITY RANK FROM THE 353 PAIRS, BY TSSR AND ESA.

#	Word1	Word2	ESA Rank	TSSR Rank
1	decoration	valor	87	347
2	aluminum	metal	73	275
3	precedent	law	96	279
4	psychology	Freud	34	150
5	physics	proton	86	280

Table II
SAMPLE WORD PAIRS WITH LOW RANK IN ESA, HIGH RANK IN TSSR

#	Word1	Word2	ESA Rank	TSSR Rank
1	cup	coffee	167	7
2	love	sex	195	33
3	drink	eat	98	65
4	life	lesson	244	39
5	movies	popcorn	309	21

Based on these issues, performing information retrieval tasks on microblogs, such as search for relevant tweets, finding similar tweets and identifying trending topics, requires customized semantic relatedness measures that take the above considerations into account. Therefore, when directly applied in the context of Twitter messages, the existing semantic relatedness techniques do not necessarily lead to the expected performance improvements.

Our work presented in this paper is focused on the development of a technique for measuring word semantic relatedness on Twitter, called Twitter Space Semantic Relatedness (TSSR). We construct a Twitter word dependency graph by exploiting hashtags and word co-occurrences in tweets. The graph is then used to extract a unique stationary distribution for each word by applying a random walk process. The similarity of two words on Twitter is then calculated based on the similarity of their corresponding stationary distributions. Our experimental results show that TSSR is able to effectively model the semantic relatedness of words when evaluated on common benchmark datasets and is able to specifically capture the shift in the semantics of words when used on Twitter. Furthermore, when used in the context of Twitter search, TSSR is able to perform better compared to state of the art semantic relatedness techniques.

The rest of this paper is organized as follows: Section 2 reviews the related work, which is the followed by Section that defines the problem we are addressing in this paper, formally defines all relevant concepts, and then presents the proposed method in detail. Section 4 reports on the extensive evaluation of the proposed work. Finally, Section 5 concludes the paper with pointers to possible future work.

II. RELATED WORK

Quality and applicability of a semantic relatedness technique rely on the information source the technique relies upon since this is what determines the types of lexical relations, the coverage of words, and the accuracy of the included information. Researchers have mainly used three types of information sources to build semantic relatedness techniques: 1) linguistically constructed sources; 2) collaboratively constructed sources; and 3) Web based content.

Linguistically constructed resources, such as WordNet [16] and GermaNet [17][18], are systematically constructed by trained linguists. Many semantic relatedness techniques based on WordNet [16] or GermaNet have been developed. For instance, Jiang et al. [19] employed the information content value of two concepts as well as the information content value of the concepts subsumer in WordNet to compute their semantic relatedness. Patwardhan et al. [20] used the co-occurrence information as well as the definitions of words in WordNet to build gloss vectors corresponding to each word; then they applied the cosine similarity function on the two vectors to obtain the relatedness between two words. Resnik [13] applied the notion of information content on the IS-A taxonomy of WordNet to measure the semantic relatedness. Hirst and St-Onge [15] established the relatedness between two words based on WordNet by finding a path that is neither too long nor that changes direction too often. The work by Hughes and Ramage [7] applies Markov chain theory to measure semantic relatedness based on the graph extracted from WordNet where the nodes are words or concepts, and the edges are formed by relational links between words or concepts.

Collaboratively constructed information sources are semantic resources constructed by user communities who do not have training or expertise in linguistics. One typical instance is Wikipedia. Gabrilovich and Markovitch [3] developed a new methodology called Explicit Semantic Analysis (ESA) that directly uses Wikipedia. In ESA, input texts are represented as weighted vectors of concepts, each concept corresponding to one Wikipedia article; the elements of the vectors are TF-IDF values of a term in the underlying article. Besides ESA, Strube et al. [12] took advantage of Wikipedia articles and category tree to compute semantic relatedness. In their work, they applied Wikipedia measures that were originally designed for WordNet. Articles are retrieved from Wikipedia by querying for word pairs. Disambiguation pages obtained for each word in a pair are used for disambiguation.

The categories related to the retrieved articles are used to compute semantic relatedness by, for instance, considering the length of the shortest-path or the length of the path that maximizes information content.

Web based information resources have received wider attention in the past few years. For instance, in order to overcome the problem that traditional document similarity methods face, which is their poor performance on short text snippets, Sahami and Heilman [21] have introduced a new approach for computing semantic relatedness by leveraging Web search results for enhancing short snippets. Top ranked words based on the TF-IDF measure from the search results are used to build a vector for each input word. Such vectors are then used to compute the degree of semantic relatedness between two words. In another approach, Radinsky et al. [22] hypothesized that by studying the similarity of word usage patterns over time, a great deal of relatedness information can be discovered to enhance the semantic relatedness results. Thus, they proposed Temporal Semantic Analysis (TSA), which considers temporal information of resources. In their method, each word is represented as a weighted vector of concept time series derived from a historical archive such as NY Times archive. Then semantic relatedness of a pair of words is computed by finding the similarity between the two times series representing the words. Cilibrasi et al. [4] have proposed a method that relies on the information retrieved from a Web search engine. The motivation behind their work is that similar words when used as search queries will result in similar Web page results. Therefore, the number of search results returned by a Web search engine for three different queries, namely w1, w2, w1 and w2, is used to formalize the normalized Google distance (NGD). Semantic relatedness is the inverse of NGD.

III. Proposed Approach

The objective of our work is to develop a semantic relatedness measure between two words, regardless of whether they have explicit semantics (e.g., dictionary words) or have no formal semantics (e.g., hashtags or slang words), based on their occurrence on Twitter. In this section, we formally define the foundational concepts of our work.

Definition 2.1 (Tweet) A Tweet *t* is defined as a triple, *t = (userId, tweetId, body)*, where *t.userId* is the unique Id associated with each Twitter user, *t.tweetId* is a unique Id of each Tweet *t*, *t.body* is the textual content of *t*.

Based on Definition 2.1, we can classify tweets according to a specific user *u*. We denote a set of Tweets that belong to a specific user as $T_u = \{t | t.userId = u\}$. We define the collection of all Tweets as *T*.

Definition 2.2 (Tweet Token) A Tweet Token *tk* is defined as a quadruple, *tk = (t, tokenId, token, isHashtag)*, where *tk.t* corresponds to the underlying Tweet *t*, *tk.tokenId* is a unique numeric identifier associated with each token, *tk.token* is the stemmed form of the word in *t*.

Based on this definition, the collection of tokens[1] within a given tweet can be represented as $TW_t = \{tk.token | tk.t = t\}$. Furthermore, we denote the Tweet Words set as $TW = \{tk.token | tk.t \in T\}$, which is the collection of all tokens observed across all tweets.

Definition 2.3 (Co-occurrence) Given two tokens w_i and w_j in *TW*, we define their co-occurrence count as $co(w_i, w_j) = |coT(w_i, w_j)|$ where $coT(w_i, w_j) = \{t | w_i \in TW_t \ and \ w_j \in TW_t\}$.

This definition will support our basic assumption that the more two tokens occur in the same tweet, the more related they are.

Definition 2.4 (Conditional Dependency) Given a token w_j and its co-occurrences with other tokens in *TW*, we define conditional dependency, $CD(w_i|w_j)$, as the probability of observing w_i if and when w_j is observed, which is calculated as follows:

$$CD(w_i|w_j) = \frac{co(w_i, w_j)}{\sum_{w_k \in TW} co(w_j, w_k)}$$

The conditional dependency definition ensures that semantic relatedness is dependent not only on the co-occurrence of the two tokens together but also on the co-occurrence of each of the tokens with other tokens in the corpus. In other words, if a token has high co-occurrence with many tokens in the corpus, it is likely that this token is less specific and therefore should receive a lower degree of semantic relatedness.

The basic premise of our work is on the latent relation hypothesis [6] that states that pairs of words that co-occur in similar contexts tend to have similar semantics. Therefore, we hypothesize that the semantics of the words on Twitter can be derived from the context in which they appear, which is typically the tweets where those words are observed. The rationale for choosing individual tweets as the context is that tweets often focus on a very specific subject and therefore each word is only used in one specific sense in a given tweet, even in the case of ambiguous words. For this reason, considering each tweet as the context allows us to focus on specific senses of each word.

Based on this, we build a co-occurrence graph in which the tokens that appear in the same contexts are connected to each other.

Definition 2.5 (Twitter Word Dependency Graph) Given *TW* and $CD(w_i|w_j)$, we define Twitter Word Dependency Graph as a weighted directed graph, denoted as TWDG, where $w_i \in TW$ are the nodes, and $CD(w_i|w_j)$ is the weight of the edge from node w_j to w_i.

Based on TWDG, we model semantic relatedness as being the probability of reaching one token from the other based on a random walk on the graph. In other words, we employ a random walk model where a particle is assumed to float through TWDG starting from a certain token node. The

[1]From here onwards, the terms word and token are used interchangeably.

probability of finding the particle at a certain node such as w_i after t iterations is equivalent to the sum of all paths through which the particle could have reached w_i starting from any other node at the time t-1; this is formalized as:

$$w_i^{(t)} = \sum_{w_j \in TW} w_j^{(t-1)} CD(w_i|w_j)$$

Now, given token w_j, the objective is to find a stationary distribution for it by releasing the particle into TWDG and iteratively applying the random walk process. The stationary distribution for w_j can be represented as the distribution of the probability of the particle being found in each of the nodes of the graph after the application of the random walk process. In order to compute the stationary distribution, we first define an initial distribution $v(w_j)^{(0)}$ that places all of the probability mass on a single token node. Then, at each iteration of the walk, the distribution is updated with parameter β as follows:

$$v(w_j)^{(t)} = \beta v(w_j)^{(0)} + (1-\beta) M v(w_j)^{(t-1)}$$

where M is the transition matrix corresponding to the TWDG graph denoting the conditional dependency $CD(w_i|w_j)$ moving from node w_j to w_i. Hughes and Ramage [7] have proposed that a random walk process is rather insensitive to the value of the β parameter and have suggested that it can be set to 0.1. They have also empirically evaluated that $v(w_j)^{(t)}$ converges to its unique stationary distribution $v(w_j)^{(\infty)}$ after a number of iterations proportional to $\beta - 1$. For us, the convergence criteria was set to $|v(w_j)^{(t)} - v(w_j)^{(t-1)}| < 10^{-6}$ for which our experiments showed to converge in around 20 iterations.

Given the stationary distribution of each token derived from the random walk on TWDG, we measure the similarity of two tokens by calculating the similarity between their stationary distributions. As suggested in the literature [7], we use cosine similarity to measure the similarity of two tokens according to their distributions.

Definition 2.6 (TSSR) Semantic similarity of two tokens w_i and w_j in TSSR is defined based on the cosine similarity of their respective stationary distributions, $v(w_i)^{(\infty)}$ and $v(w_j)^{(\infty)}$ as follows:

$$SR(w_i, w_j) = \frac{v(w_i)^{(\infty)} \cdot v(w_j)^{(\infty)}}{|v(w_i)^{(\infty)}||v(w_j)^{(\infty)}|}$$

IV. Evaluation

We have benefited from the tweets dataset released by Cheng et al. [8] as the information source for building TWDG and for computing TSSR. After parsing the tweets and performing preprocessing such as removing stopwords and stemming the words in the dataset, we obtained 8,770,157 tweets with an average length of 8.4 words published by 106,349 users. These tweets were collected from 10 Nov 2006 until 17 March 2010. There were 4,148,886 unique words in total, which served as the vertices of the TWDG.

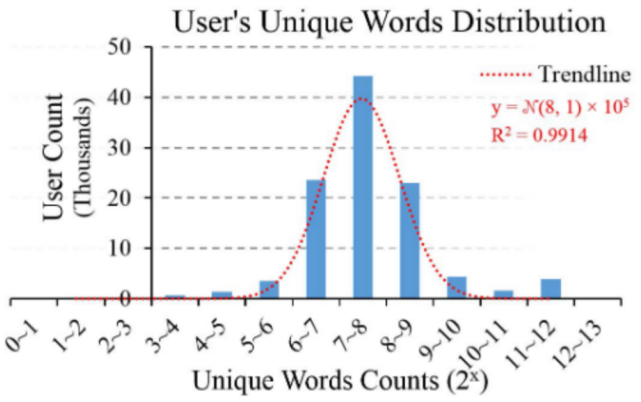

Figure 1. User's Unique Words Distribution in the Twitter dataset

There are typically two approaches for evaluating the performance of semantic relatedness techniques. The first approach relies on a gold standard dataset of word pair similarities collected from a group of human subjects. The performance of a relatedness technique is measured through its degree of correlation with the subjective assessment of human subjects. The second approach evaluates two or more semantic relatedness techniques by measuring their impact on an application specific problem, e.g. their impact on improving product search. In this paper, we evaluate TSSR using both approaches as well as a third strategy that consists of subjective assessment of the ability of TSSR to to describe frequent Twitter hashtags that do not have direct English language semantics. To sum up, we evaluate our work from three perspectives:

1. First, we employ the gold standard-based evaluation approach to compare the performance of our technique to the state of the art semantic relatedness techniques. We benchmark our work against five other techniques from the literature on three different datasets.
2. Second, we implement and compare our work against the best performing semantic relatedness technique (ESA) on the application-specific problem of tweet search in order to observe how our technique performs in contrast to ESA.
3. Finally, given the fact that none of the existing semantic relatedness techniques is able to calculate the semantic relatedness of non-dictionary words, we perform an experiment involving human subjects to determine the suitability of our technique for semantically relating such words in practice.

In the following, we describe the details of our Twitter dataset and report on the three evaluation tasks.

A. Overview of the Twitter Dataset

As mentioned earlier, we used the Twitter dataset provided by Cheng et al. [8] that contains over 8.5M tweets and over 4M unique words. As shown in Figure 1, the majority of

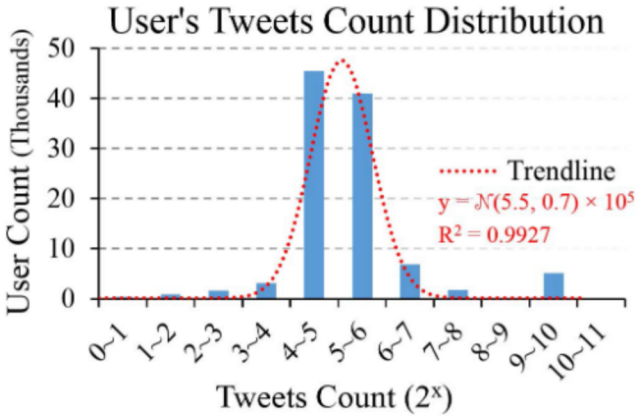

Figure 2. User's Tweets Count Distribution in the Twitter dataset

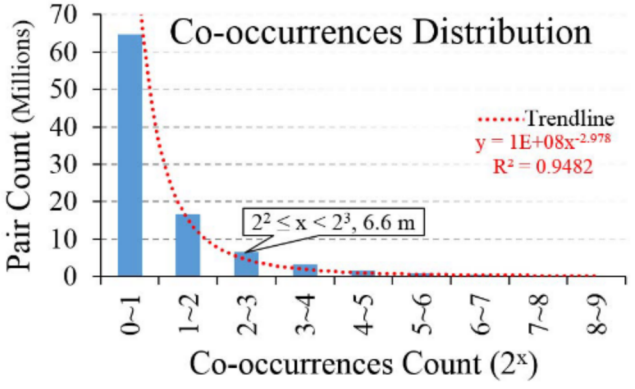

Figure 3. Co-occurrences Distribution in the Twitter dataset

users used between 128 to 256 unique words across all the tweets in their timeline. There is a small number of users with a very small vocabulary, i.e., less than 64 unique words in total, or very large vocabulary, i.e., more than 512 unique words. This shows that the Twitter users that were covered in our dataset had a very focused and limited vocabulary that they frequently used. While we do not generalize this observation, we believe this might be a trend on Twitter since our dataset included over 8.5M tweets. Furthermore, as shown in Figure 2, most of the users in our dataset posted between 16 to 32 tweets in the 3.5 year period. In terms of the co-occurrence of words in the Twitter data, a significant number of words had only been observed together once, which does not allow the derivation of any meaningful semantic relatedness between such words. Figure 3 shows the co-occurrence of words in the Twitter dataset. The co-occurrences are calculated by counting the number of times two stemmed words are seen together in the same tweet.

B. Gold Standard-based Evaluation

Traditionally, semantic relatedness techniques have been evaluated based on the correlation of their results with a gold standard dataset collected from human judges. These datasets include a collection of word pairs along with the assessment of human experts with regards to the similarity of the words in each word pair. For instance, WordSimilarity-353 collection contains 353 English word pairs [9], RG-65 consists of 65 word pairs [10] and MC-30 is a collection of 30 word pairs [11], which have been widely used in the literature. Many researchers [5] have used these datasets to show that their method is able to reasonably reproduce the word pair similarity rankings (not the actual relatedness value but the ranking of the word pair in the word pair dataset) by calculating spearmans rank correlation (ρ).

Our first assessment method consisted of benchmarking our work against the three aforementioned gold standard datasets and comparing the results with the state of the art

Table III
SPEARMAN'S RANK CORRELATION AND MAE RESULTS

Method	Spearman's Rank Correlation (ρ)			Mean Absolute Error (MAE)		
	WSW -353	RG -65	MC -30	WSW -353	RG -65	MC -30
ESA[3]	0.75	0.82	0.73	4.2	1.3	1.9
WikiRelate[12]	0.49	0.52	0.45	-	-	-
Hughes and Ramage[7]	0.47	0.76	0.84	-	-	-
WordNet-Res[13][14]	0.30	0.55	0.72	4.3	1.5	1.5
WordNet-Path[15][14]	0.19	0.50	0.48	3.6	1.7	2.1
TSSR	0.61	0.56	0.73	2.1	1.2	0.9

semantic relatedness techniques [5] (see Section 4 for an overview of these techniques). The first three columns of Table III show the results of Spearmans rank correlation on the three datasets. As this table shows, on two of the datasets, TSSR does not perform as well as ESA. This is an expected result, consistent with the main hypothesis of our work: the semantics of words when used on Twitter differ from their more formal widely-used definition that has been employed in these datasets. However, even though the semantic relatedness derived by TSSR deviates from the one exposed by ESA, it is not overly remote from formal judgement rankings, as the computed rank correlations on the WSW-353 dataset demonstrate. Figure 4[2] clearly depicts the difference between the semantic relatedness score distributions produced by TSSR compared to ESA. As an example the figure shows that word pairs such as *game* and *victory* are not considered to be too highly semantically related in ESA or the WSW-353 dataset but are considered to be highly related by TSSR due to their frequent co-occurrence on Twitter. We will show in our next two experiments that such differences are a desirable effect of capturing the semantics of words based on Twitter context.

Apart from Spearmans rank correlation, in Table III we also report on the Mean Absolute Error (MAE) of the estimated semantic relatedness values produced by each

[2]Due to space limitation, not all word pairs are listed on the x-axis.

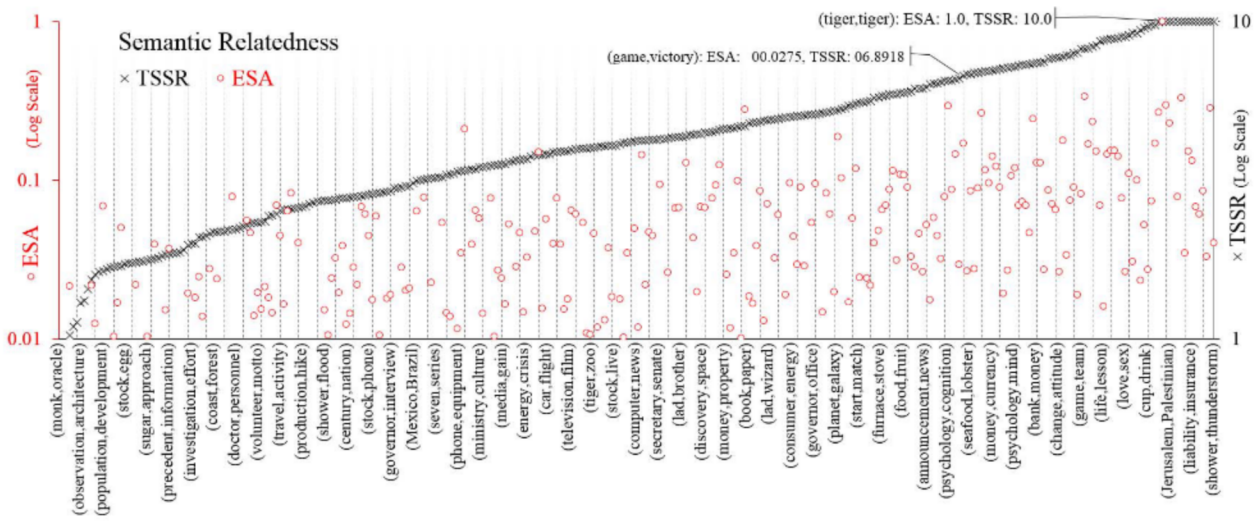

Figure 4. Comparison between ESA and TSSR semantic relatedness scores on the WSW-353 dataset

method and the actual human judgments:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|m_i - g_i|$$

where n is the number of word pairs, m_i is the score produced by a semantic relatedness method for word pair i and g_i is the gold standard score for the same pair. In order to calculate MAE, the semantic relatedness values produced by different methods were scaled to [0,10], which is the scale used in gold standard datasets. As shown in Table III, TSSR produces the smallest mean absolute error across all of the three gold standard datasets. This means that the value proposed by TSSR in the range of [0, 10] for each pair of words is closer to the actual value attached by human subjects compared to other methods. However, statistically speaking, as observed in Table III, the lowest MAE does not result in the highest rank correlation. In other words, a method can have a low MAE but produce a ranking that is not the same as the gold standard. It should be noted that the implementation of methods [12] and [7] were not publicly available; therefore, we were not able to generate MAE values for these two methods.

C. Tweet Search

The second evaluation strategy that we adopted was an application-based method. Given the fact that one of the most important application areas of semantic relatedness techniques is to improve search, we compared TSSR to ESA, which showed the best performance in the first experiment, when applied to the domain of tweet search. In order to integrate semantic relatedness into tweet search, we extended the baseline vector-based comparison of query terms with tweet space terms. Hence, the similarity of a tweet to a query is calculated as the sum of semantic relatedness between query terms and tweet terms as follows:

$$S_{tweet}(T,Q) = \sum_{i=1..k}\sum_{j=1..n} SR(q_i, w_j)$$

where n is the number of words in a tweet (T) and w_j is the j^{th} word in the tweet; k is the number of words in the query (Q) and q_i is the i^{th} word in the search query. For a given tweet T and a query Q, $S_{tweet}(T,Q)$ calculates the semantic relatedness between T and Q. In the search process, tweets are ranked based on their degree of semantic relatedness to the input query. We used TSSR and ESA for $SR(q_i, w_j)$ and performed our evaluation as follows.

For a given single-term query, we find 100 tweets that have the exact query term in their content; we refer to these as target tweets. We then identify 900 tweets that neither contain the exact query term nor have topical similarity with the query term (determined by human expert); we refer to these as irrelevant tweets. We then anonymize the target tweets by removing the exact query term from the tweet content. Therefore, the overall dataset of 1,000 tweets contains 10% of relevant tweets and 90% irrelevant tweets, none of which have the exact query term in them. The objective is to study whether and to what extent the search method based on $S_{tweet}(T,Q)$ is able to find the target tweets based on the computed semantic relatedness values and without the presence of the exact query term.

For the purpose of experimentation, we selected the 100 most frequent words in the overall dataset used in this study, as the 100 queries to be used for search. We performed the above procedure for each of the 100 queries and employed the standard TREC evaluation tool to compute the performance measures. We report three metrics in Table IV, namely i) Mean Average Precision (MAP), which is the mean of the average precision scores of each query; ii) Reciprocal Rank that shows the multiplicative inverse of the rank of the first correct answer; and iii) Precision at 100 (P@100), which

shows the ratio of correct tweets in the top 100 results.

Table IV
SEMANTIC SEARCH OVER TWEETS

Method	MAP	Reciprocal Rank	P@100
ESA	0.31	0.79	0.33
TSSR	0.39	0.95	0.38

The results reported in Table IV show that TSSR is more effective in finding a higher number of tweets from the target tweet set. Given the search for relevant tweets in this evaluation strategy is only dependent on the performance of the semantic relatedness technique, we believe that the results are an indication that the semantic relatedness derived by TSSR for word pairs on Twitter is more accurate and representative of the semantics of words as they are used by Twitter users. Therefore, as observed in Figure 4, a shift can be observed between the semantics of a word on Twitter and it's common semantics, which if captured as done in TSSR, can lead to a higher performance when performing tweet search and possibly other Twitter related applications.

D. Describing Hashtags

The third evaluation strategy was to determine whether our semantic relatedness technique is able to identify the correct semantics of words that do not necessarily have explicit English language semantics such as hashtags. We performed this evaluation with 35 human participants, all with a good understanding of the Twitter dynamics. Each participant was given a hashtag along with a set of 25 descriptive words that described that hashtag. The descriptive words for a hashtag were derived by using TSSR to find the top 25 words that had the highest semantic relatedness with that hashtag. Table V shows five sample hashtags and their descriptive words that were included in the experiments. Each participant was then asked to provide their perspective on the following three statements regarding the relationship between the hashtag and its 25 descriptive words:

1. *The 25 words for the given hashtag are highly descriptive.*
2. *There are no irrelevant descriptors within the 25 words set.*
3. *There are no important missing descriptors from the 25 words set.*

The participants were asked to provide their assessment using a five level Likert-type scale. The purpose of the first statement was to determine whether the correct semantics of the hashtag was identified by TSSR. The second statement focused on an informal assessment of precision, while the third statement evaluated perceived recall. We selected the top 50 most frequent hashtags in our dataset and each hashtag was independently assessed by seven participants.

Table V
SAMPLE HASHTAGS AND THEIR DESCRIPTIVE WORDS SET

Hashtag	The 25 Descriptive Word Set	Meaning
HW	finish, done, home, class, school, help, hour, math, read, study, tired, book, assign, write, paper, idea, problem, pic, homework, stupid, page, monday, spanish, teacher	Homework
MJS	jackson, show, michael, song, movie, miss, music, die, listen, world, dance, memory, death, fan, perform, hear, beat, remember, white, gone, sing, left, album	Michael Jackson
TCOT	tlot, p2, obama, gop, sgp, teaparty, health, news, care, bill, show, healthcare, hcr, vote, video, palin, ocra, help, job, conservation, read, talk, president, plan	Top Conservatives on Twitter
STEM	research, study, current, grow, health, world, education, institution, school, derive, approve, challenge, science, stress, success, conflict, learn, brain, technology, develop, vote, university, student, support	Science, Technology, Engineering, and Math
BOHO	style, chick, bag, leather, brown, hair, shop, tan, tote, fashion, saddle, wooden, trendy, show, wear, sale, rock, pretty, design, dress, store, cut, jacket, shoe	Informal and Unconventional Fashion

Table VI
INTRA-CLASS CORRELATION (ICC) OF THE PARTICIPANTS

ICC single measure	ICC average measure
0.757	0.956

First, in order to examine whether the opinions provided by human participants were consistent and that valid conclusions can be drawn from the data, we performed inter-rater reliability analysis. In particular, we applied intra-class correlation (ICC), which is a descriptive statistic that is used to measure the agreement within a group of individuals. In Table VI we report both ICC single measure and ICC average measure. The former defines the extent to which the opinion of a single participant is similar with the other participants, whereas the latter shows how reliable it is to use the average opinions of participants. As shown in Table VI, the ICC single measurement value (0.757) shows a reasonable agreement among the participants, while the ICC average result (0.956) shows the reliability of the study.

Given that we have demonstrated that the participants were highly consistent in their responses to the three questions (high ICC values); therefore, it is reliable to use the median of the values received for each of the questions to represent the subjective opinion of the participants. The median of the answers for all three questions was 4, which corresponds to *agree* (5 for strongly agree and 0 for strongly disagree), which is an indication that the participants collectively agreed with the three statements regarding descriptiveness, precision and recall of the 25 descriptive words set extracted by TSSR for each of the hashtags. This shows the fact that TSSR has been able to identify the semantics of the hashtags, i.e., words that do not have explicit English language semantics, with a reasonably

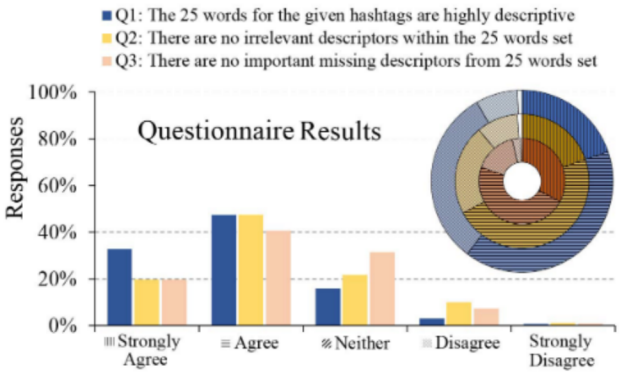

Figure 5. Results of the Hashtags study

high quality. The distribution of the answers received from the participants are shown in Figure 5.

V. Conclusion

In this paper, we have proposed a novel approach for computing semantic relatedness between two words on Twitter by looking at word co-occurrences on this social network. We have conducted three different types of experiments to assess how well our approach is able to identify the semantics of words within the context of Twitter and measure the semantic relatedness of two words. We have shown that semantics of some words may shift when used on Twitter. Therefore, state of the art semantic relatedness techniques that focus on encyclopedic knowledge sources are not able to accurately identify the semantics of words in the Twitter context and therefore, would not be ideal for application on this platform. Our proposed approach is able to not only identify the semantics of dictionary words on Twitter but also to capture their semantic shift. In addition, it is able to semantically describe new words on Twitter that do not have formal dictionary semantics such as Internet slang.

There are two avenues of future work that we would like to explore. First, we are interested in studying whether using a broader context for words on Twitter would impact the quality of the semantic relatedness measure. In the current form, TSSR considers words in the same tweet to have the same context. We would like to expand the context to cover words from the tweets of one user in the same day, or words in a tweet and all of its responses. Second, we are also interested in automatically deriving different senses of a word based on its context on Twitter. In the current form, each word, regardless of how many senses it may have, is represented as a single node in our graph. As future work we will try to determine the multiple senses of a single word so that semantic relatedness can be measured more accurately.

References

[1] Budanitsky, A., & Hirst, G., Evaluating wordnet-based measures of lexical semantic relatedness, *Computational Linguistics, 32(1)*, 13-47, 2006.

[2] Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., & Motta, E., Semantically enhanced Information Retrieval: an ontology-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web, 9(4)*, 434-452, 2011

[3] Gabrilovich, E., & Markovitch, S. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. *IJCAI*, 1606-1611, January, 2007.

[4] Cilibrasi, R. L., & Vitanyi, P., The google similarity distance. *IEEE TKDE, 19(3)*, 370-383, 2007.

[5] Zesch, T., Study of semantic relatedness of words using collaboratively constructed semantic resources. *TU Darmstadt*, 2010.

[6] Turney, P. D., The latent relation mapping engine: Algorithm and experiments. *JAIR*, 615-655, 2008.

[7] Hughes, T., & Ramage, D., Lexical Semantic Relatedness with Random Graph Walks. *EMNLP-CoNLL*, pages 581-589, June, 2007.

[8] Cheng, Z., Caverlee, J., & Lee, K., You are where you tweet: a content-based approach to geo-locating twitter users. *CIKM*, 759-768, 2010.

[9] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E., Placing search in context: The concept revisited. *WWW*, 406-414, 2001.

[10] Rubenstein, H., & Goodenough, J. B., Contextual correlates of synonymy. *Communications of the ACM, 8(10)*, 627-633, 1965.

[11] Miller, G. A., & Charles, W. G., Contextual correlates of semantic similarity. *Language and cognitive processes, 6(1)*, 1-28, 1991.

[12] Strube, M., & Ponzetto, S. P., WikiRelate! Computing semantic relatedness using Wikipedia. *AAAI* 1419-1424, 2006.

[13] Resnik, P., Using information content to evaluate semantic similarity in a taxonomy. *arXiv cmp-lg/9511007*, 1995.

[14] Pedersen, T., Patwardhan, S., & Michelizzi, J., WordNet:: Similarity: measuring the relatedness of concepts. *HLT-NAACL*, 38-41, 2004.

[15] Hirst, G., & St-Onge, D., Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305-332, 1998.

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

[17] Hamp, B., & Feldweg, H., Germanet-a lexical-semantic net for german. *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 9-15, July, 1997.

[18] Henrich, V., & Hinrichs, E. W., GernEdiT-The GermaNet Editing Tool. *ACL*, 19-24, 2010.

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

[20] Patwardhan, S., Banerjee, S., & Pedersen, T., Using measures of semantic relatedness for word sense disambiguation. *In Computational linguistics and intelligent text processing*, pages 241-257, 2003.

[21] Sahami, M., & Heilman, T. D., A web-based kernel function for measuring the similarity of short text snippets. *WWW*, 377-386, 2006.

[22] Radinsky, K., Agichtein, E., Gabrilovich, E., & Markovitch, S.,A word at a time: computing word relatedness using temporal semantic analysis. *WWW*, 337-346, 2011.