International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 07
June 2019

Short Text Similarity Measures for Social Sentiment Comments

Zar Zar Hnin[1], Ei Ei Mon[2], Cho Cho Khaing[2]
[1]Faculty of Computer Science, University of Computer Studies (Mandalay), UCS-MDY, Myanmar
[2]Faculty of Computer Science, University of Computer Studies (Loikaw), Kayah State, Myanmar

zarzarhnin@gmail.com, eieimon80@gmail.com, chokhaing28@gmai.com

Abstract:

The feeling of social networks is the attitude and feelings people have about their brand on social networks. Adding context to all commissions, comments, and actions need to analyze. It is important that the brands listen carefully to what is said about their online business. And more importantly to know whether the conversation is positive or negative. In this paper, we find the similar groups of comments which talk about the same context depending on particular topic. By doing so, we can help online business works to group the customers who shares common interest and same feeling of their products. This paper introduces how to find the similar text in semantic ways in both word-level and phrase-level measures by filling the gap of syntactic measures in text similarity. For the datasets, Twitter dataset is used for system implementation because their comments are short and compatible with our proposed system. According to the experimental results, the results get promising results in terms of higher accuracy rate but lower error rate by switching two datasets available from online.

Keywords: social network, text similarity, semantic analysis, word-level, phrase-level

1. Introduction

Sentiment analysis is also useful when monitoring keywords. In addition to seeing what the general public has to say, you can find influential people and opinion leaders relevant to your industry [1-5].

Measuring belief in society-often referred to as a social-learning agenda, an integral part of any social media monitoring plan. It helps us understand what a person feels behind a social media post. Knowing the feeling behind a post can give us an important context on how we can continue and respond [6, 7, 12].

Sentiment Analysis also known as Opinion Mining is a field within Natural Language Processing (NLP) that builds systems that try to identify and extract opinions within text. Usually, besides identifying the opinion, these systems extract attributes of the expression e.g, positive, negative, etc.

Instead of querying final sentiment decisions, this paper boost the analysis procedures by proposing semantically similarity measures between comments groups of different users so that they can further be used for further analysis by corresponding business area such as online shopping, public policy making, etc.

Similarity measures plays a key role in classification of texts in several fields, including signal processing, natural language processing, statistics and information retrieval and also sentiment analysis. This measure is needed to retrieve the documents relevant to user query.

An order structure is also required to solve some violations in a large text copyright by extracting knowledge extracted from structured data such as Wikipedia and for automatic text correction where an improper spelling word is replaced by a dictionary a word of high degree of similarity. Measurement of text equivalence is more widely used in discovering plagiarism. Since plagiarism often occurs in certain parts of the text, the text should be split into smaller fragments before measuring similarity [12,13,16,18].

In this paper, we propose a new text similarity measures by means of phrase-wise and word-wise similarity to reveal the similar groups of writing styles with different word usage and styles. We even explore embedded words using the help of WordNet so that we can explore more similarity by semantically rather than syntactic matching in contemporary matching processes.

The rest of the paper is organized as follows. Related works are described in Section 2, theory background for text similarity techniques and sentiment analysis are explored in section 3. The proposed method is presented in Section 4 and paper is concluded in Section 5.

2. Related Works

Previous efforts focus on calculating the same semantics between documents, concepts or phrases. In recent natural language processing applications, they demonstrate the stronger need to find effective methods to measure semantic similarities between variable length texts, and general methods are suggested for these people [8-11].

The work of Zhao et al (2013a) indicates that the relevance of the users' ratings is good for the same function, but only using search rankings from a set of similar users is not so all right. The idea behind the computation of semantic similarities between text analysis focuses on that generates a range of users with revision texts similar to both articles and a range of users.

It also provides descriptions of similar elements in their reviews, which is the main task of collaborative filtering and content-based techniques.

Checking that semantic analogy texts provides the correct path to understand, compare and study the concepts that are subject to each term in the test texts for both of them. The words in two different test texts may not necessarily mean the same concepts. The correct concept of word mapping and word comprehension is required according to the study [2] with the use of evaluation text refers to rating predictions.

Some of the existing works of Leung et al., [3] and Zhang et al., [4] use texts of change by conducting an assessment of the user's opinion that is reflected in their texts of change, to improve personalized recommendations. However, we are debating for effective exploitation approaching, we should use the same semantics between user revision text instead of emotionally because the factors behind the preferences of the users are well reflected in semantic similarities as contrary to the study of feelings of change texts.

Only a few methods combine the rules of character level and token in Cohen et al., [5,6] These methods are called important steps. The principle of a soft proposal is to apply a level indicator on all pairs of tokes between the strings and only consider tokens that ensure some criteria (e.g, threshold) as input to a measure at the token level. According to Jiménez et al. [7] soft cosine, both the character level and the token level appear to match the name. Soft cosine has cosine to match tokens and bigrams to match the character level.

However, when applied as independent measures, neither the cosine nor the bigram are the best option [7]. Hopefully we can find a better combination of other characters and token level measures through effective evaluation.

Regarding the chain-based similarity, Islam et al. [8] proposed a standardized and modified version of the string matching algorithm of the Longest Common Undercurrent (LCS) to measure the similarity of the text. It works together with a corpus-based measure, its methods achieved a very competitive result.

3. Background Theory

3.1. Sentiment Analysis

The AI system should understand similar identifiers from users and provide a consistent response. The emphasis on semantic objectives is to create a system that identifies language and word patterns to generate responses similar to how human conversations work. For example, if the user asks "What happened to the color I ordered?" Or "What's wrong with my shirt's design?". In this case, this paper organizes the similar user group for online analyzers so that they can make further decision in a timely manner without needing individual comparison on every online comments of the social newt wok users.

With the growth of the use of social networks, online learning, communities and groups have become an attractive research domain. In this context, the integration of users as minds is one of the emerging problems. In fact, it gives a good idea about the formation and evolution of groups, explains various social events and leads to many applications, such as product recommendations, public opinion discussions, etc.

3.2. Similarity Measures

3.2.1 Syntactic measures: Works with words and their characters without any language or opinion is difficult to extract the meaning of the content. Therefore, they are more general than semantic measures. In general, the synthetics measure makes a distance, which indicates how two data elements differ. The greater the distance between the two elements, the more they are. Distance and similarity can be used differently, such as inverse functions. In document analysis, all distance measurements are converted into equivalence measurements, which returns a score in the range [0, 1] where 0 means unusual and 1 means exact match.

3.2.2 Semantic measures: Word embedding has become extensively in Natural Language Processing.

International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 07
June 2019

They allow us to easily calculate the same semantics between the two words, or to find words that are very similar to a target word. However, we are often more interested in similarities between two sentences or short texts.

Many NLP applications are needed to calculate equivalence in meaning between two short texts. Search engines, for example, need to model the relevance of a document in a query, beyond the merging of words between the two.

3.2.3 Methods of Similarity Measures: there are some popular ways of computing similarity in both syntactic and semantic analysis as follows.

- Jaccard Similarity
- Different embedding, K-means and Cosine Similarity
- Word2Vec and Cosine Similarity
- Different embedding and Siamese Manhattan LSTM
- Different embedding and Variation Auto Encoder, etc.

Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers.

4. Proposed Semantic-based Similarity Measures

In this paper, we use pre-trained sentence encoders as combination of Smooth Inverse Frequency and referenced model Google Sentence Encoder.

The method for estimating the semantic analogy between a pair of sentences is to average the words of the inlays of all the words in two sentences and calculate the cosine between the resulting inlays. Obviously, this simple baseline leaves a considerable space for diversity.

Taking the average of word inlays in a sentence tends to give too much weight to words that are quite irrelevant, semantically speaking. Smooth Inverse Frequency (SIF) tries to solve this problem in two ways:

- Weighting the key terms: using term-frequency and inverse-document frequency (tf-idf)
- Common component removal: SIF computes the principal component of the resulting embedding for a set of sentences. It then subtracts from these sentence embedding their projections on their first principal component. This should remove variation related to frequency and syntax that is less relevant semantically.

SIF removes unimportant words, alternatively known as stopping words such as but, just, etc., and keeps the information that can expose the most to the semantics of the sentence.

4.1. Proposed Architecture of Similarity Matching

By referencing the similarity matching models of InferSent [19] and the Google Sentence Encoder, we built a pre-trained encoder for phrase-level and word-level semantic matching of social network comments and opinion-contained short texts.

In this encoders, we use soft-max combination phases with three layers to organized the partial results obtained from phrase and word level similarity matching results.

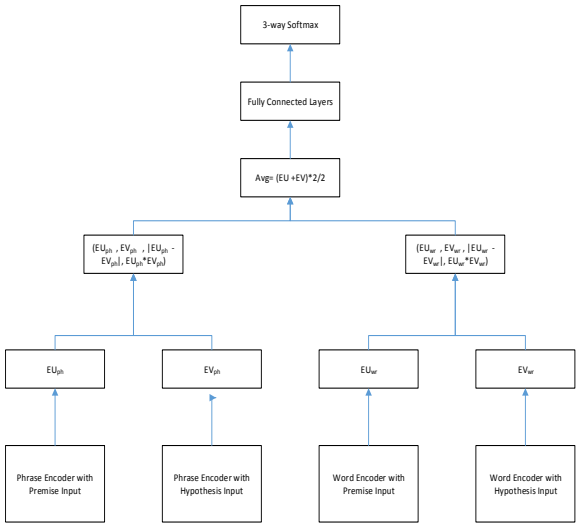


Figure 1. Pre-trained Encoder Model for Phrase and Word Level Semantic Similarity Matching

To demonstrate how proposed system works on similarity matching upon the short text that can probably found on social networks.

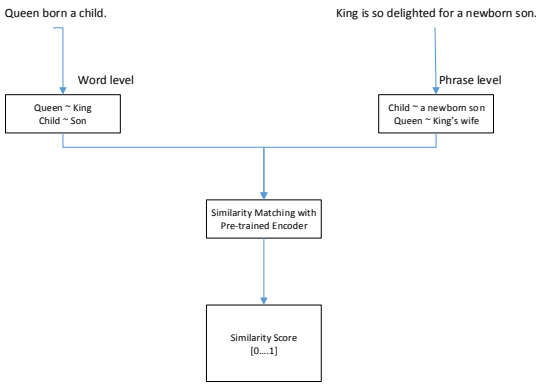


Figure 2. Working Flow of Proposed Similarity Matching Process

In matchmaking process, word-level measures can be categorized into the following three classes:

- Exact match,
- Word Transformed Match, and
- Longest common substring (LCS)
- Ignorance of word sequence (IWS)

In this paper, we assume all types of similarity as similar so, we take the matching result as one, whereas phrase level, with the help of WordNet, we find the sparsity of a word and find their semantic meanings, in this case, the similarity values is regarded depending on the distance they are found.

5. Datasets and Implementation

5.1. Datasets and Setting

To test the efficiency of proposed models and matching process, we use two alternative datasets so that the proposed system is shown to be independent of datasets. For ever experimental works, we set different variations of words and sentence structure so that the overall average result is summarized and shown in the figures.

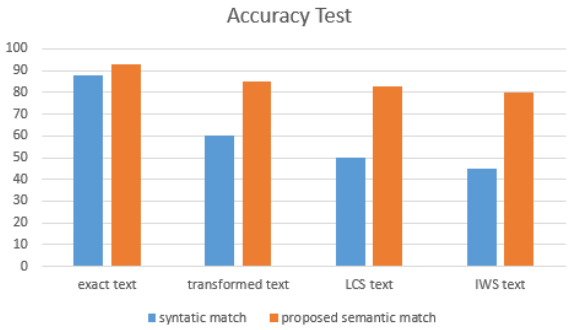
5.2. Experimental Works

The experiments are performed in two setting aspects as shown in following subsections to measure the accuracy and error rate executed by the proposed system. The accuracy rate is used to determine if a value is accurate compare it to the accepted value. In an experiment observing a parameter with an accepted value of VA and an observed value VO, there are two basic formulas for percent accuracy:

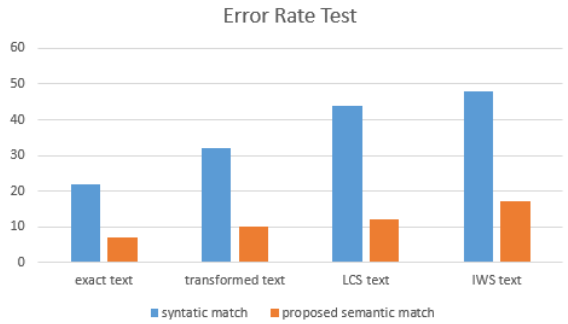
Accuracy (%) = (VA - VO)/VA X 100

Percent error is the difference between a measured and known value, divided by the known value, multiplied by 100%.

5.2.1 Impact of text variations: This experiment is to measure how the proposed system accurately or mistakenly match the text which are disguised in different synonyms and semantic relations. According to the results illustrated in Figure 3 (a) and (b), our proposed semantic match achieves significant better results against traditional syntactic matching.



(a) accuray test on text variations

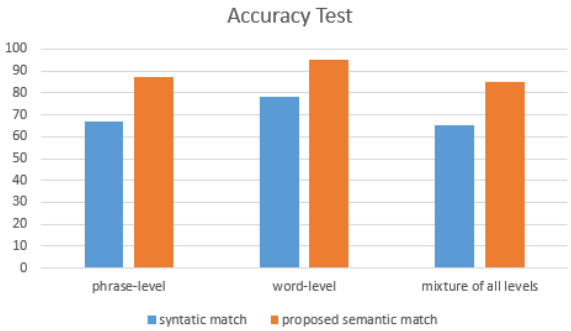


(b) error rate test on text variations

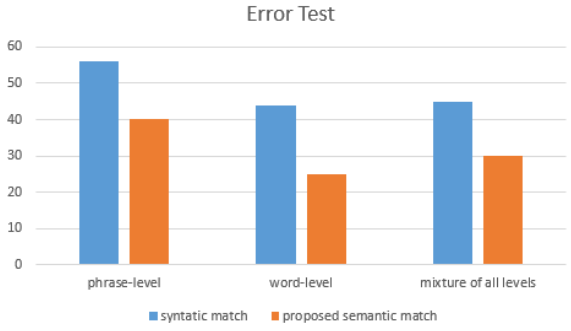
Figure 2. Measurements depending on text variations

5.2.2 Impact of different level matching:

To experiment the performance of phrase-level and word-level matching, we perform this experiment by switching different levels on the test and takes average accuracy and error rate to show the figures as follows.



(a) accuracy test on different levels



(b) Error rate test on different levels

Figure 3. Measurements depending on different phrase and word levels

As shown in figures, the results of our proposed system outperform the traditional syntactic approaches in both accuracy and error results under different parametric setting.

6. Conclusions

In this paper, we search users of the group to share the same interests by studying their textual comments. The main purpose is to retrieve the user group by analyzing their context's semantically similarity so that the publishers' interest centers and, group interests can be revealed. This paper presents a text similarity in semantic ways using pre-defined encoders in both word and phrase level matching and gains far better results compared with other approaches. As future work, we have plan to extend this approach with better promising techniques so that better results could be revealed.

7. References

[1] Zhao, X., Niu, Z., & Chen, W. (2013b). Opinion-based collaborative filtering to solve popularity bias in recommender systems. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 8056 LNCS, pp. 426–433). https://doi.org/10.1007/978-3-642-40173-2_35

[2] Leacock, C., & Chodorow, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. WordNet: An Electronic Lexical Database., (JANUARY 1998), 265–283. https://doi.org/citeulike-article-id:1259480

[3] Leung, C. W., Chan, S. C., & Chung, F(2006), Integrating Collaborative Filtering and Sentiment ANalysis: A Rating Inference Approach. ECAI 2006 Workshop on Recommender Systesm, 62-66, https://doi.org/10.1.1.69.1870

[4] Zhang, W., Ding, G., Chen, L., Li, C., & Zhang, C. (2013). Generating virtual ratings from Chinese reviews to augment online recommendations. ACM Transactions on Intelligent Systems and Technology, 4(1), 1–17. https://doi.org/10.1145/2414425.2414434

[5] Cohen, W., Ravikumar, P., & Fienberg, S. (2003a). A comparison of string metrics for matching names and records. In Kdd workshop on data cleaning and object consolidation, 3, (pp. 73-78).

[6] Cohen, W., Ravikumar, P., & Fienberg, S. (2003b). A comparison of string distance metrics for name-matching tasks. In II Web (pp. 73-78).

[7] Jimenez, S., Gonzalez, F., & Gelbukh, A. (2010). Text comparison using soft cardinality. In International Symposium on String Processing and Information Retrieval (pp. 297-302). Springer Berlin Heidelberg.

[8] Islam Aminul, & Inkpen Diana. (2008). Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data (TKDD), 2(2), 10.

[9] N Gali, R Mariescu-Istodor, D Hostettler , Framework for syntactic string similarity measures, Expert Systems with Applications, 2019.

[10] Liu Yang, Sun Chengjie, Lin Lei, & Wang Xiaolong yiGou: A Semantic Text Similarity Computing System Based on SVM. Paper presented at the Proceedings of the 9th International Workshop on Semantic Evaluation, pages 80-84, Denver, Colorado, USA.

[11] Manning Christopher D, Raghavan Prabhakar, & Schütze Hinrich. (2008). Introduction to information retrieval (Vol. 1): Cambridge university press Cambridge.

[12] Meng Lingling, Huang Runqing, & Gu Junzhong. (2013). A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology, 6(1), 1-12.

[13] Sultan Md Arafat, Bethard Steven, & Sumner Tamara. (2014b). DLS@CU: Sentence Similarity from Word Alignment. Paper presented at the Proceedings of the 8th International Workshop on Semantic Evaluation, pages 241-246, Dublin, Ireland

[14] Wu Zhibiao, & Palmer Martha. (1994). Verbs semantics and lexical selection. Paper presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics.

[15] Mikolov Tomas, Chen Kai, Corrado Greg, & Dean Jeffrey. (2013). Efficient estimation of word representations in vector space. Paper presented at the ICLR, 2013.

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

[17] Fattah Mohamed Abdel, & Ren Fuji. (2008). Automatic text summarization. World Academy of Science, Engineering and Technology, 37, 2008.

[18] Gabrilovich Evgeniy, & Markovitch Shaul. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. Paper presented at the IJCAI.

[19] https://github.com/facebookresearch/InferSent

[20] https://www.kaggle.com/c/twitter-sentiment-analysis2/data

[21] https://blog.cambridgespark.com/50-free-machine-learning-datasets-sentiment-analysis-b9388f79c124