Text Similarity Algorithms.

Sandeep Adobe[*], Rohit Murkute[**], Manisha Bharti [***]

[*] Student at Savitribai Phule Pune University, Department of Technology,
[**]Student at Savitribai Phule Pune University, Department of Technology.
[***] Associate Professor at Department of Technology, Savitribai Phule Pune University,

Abstract

Text similarity measurement compares text with available references to indicate the degree of similarity between those objects. There have been many studies of text similarity and resulting in various approaches and algorithms. This paper investigates two majors text similarity measurement approaches, which include Machine learning & Lexical Based approach. The main target of this survey is to give nearly full image of text similarity measurement techniques and the related fields with brief details.

Introduction: -

Text similarity is one of the applications of linguistics and statistics to natural language processing and it helps in many different applications.

For example, If the user is looking for information about cats, we may want the system to return documents that mention the world kittens and not only the word cat so the document may not have any words in common with the query and still be related because cat and kitten are similar words.

Some of the popular examples are IBM's Watson system famously played on television against the best human contestants in jeopardy, natural language assistance such as Apple's Siri, Alexa and translation systems you're probably familiar with Google Translate there are other applications for example Grammarly there are applications to next generation, for example, the LA Times applies some computer software to generate reports about earthquakes automatically all those techniques use natural language processing

Computers are not inherently designed to understand the human language they're very confused by human language some very specific techniques are needed that would teach computers how to use human language natural language processing is the field that

teaches computers how to understand language and natural language processing is a very multidisciplinary field it draws on research in linguistics which is the study of language, theoretical computer science, mathematics & statistics, artificial intelligence and even fields like psychology and databases and user interfaces and whatnot.

Text similarity has to determine how ‘close’ two pieces of text are both in surface closeness lexical similarity and meaning semantic similarity.

For example, in the phrases "The Lion ate the buffalo" & "The buffalo killed the Lion" can you tell the similarity by just looking at the words?

On the surface, if you considered only the words the two phrases appear to be very similar as 4 of 5 words are a match, as it does not take into account the context.

When taking into consideration the context or the semantics. We need to focus on the phrase/paragraph level (or lexical chain level) where a piece of the sentence is broken into a relevant group of related words before computing similarity. We know that while the words significantly overlap the phrases have a different meanings.

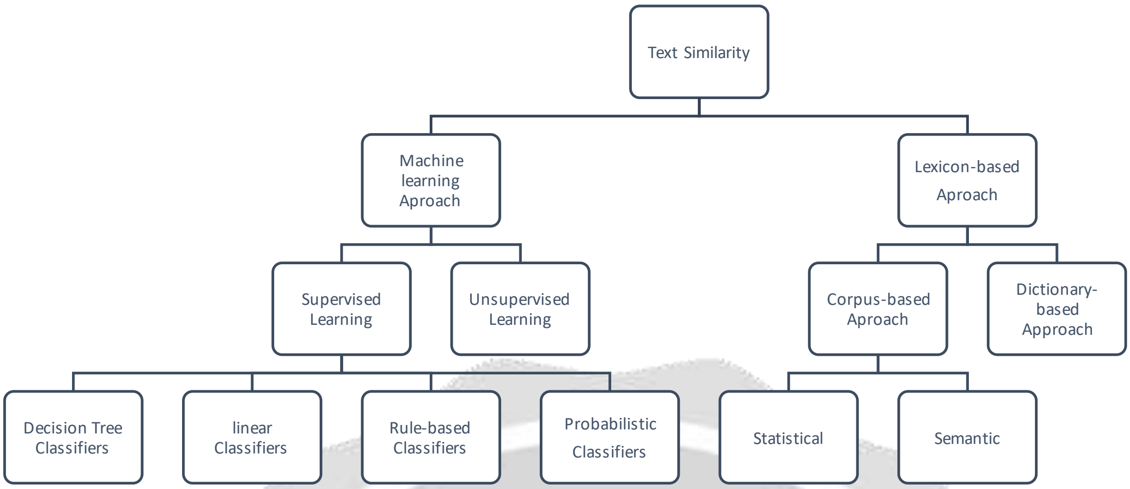

Fig.1 Text Similarity identification techniques.

2. Method

To complete the study of this text similarity, we conducted a performance investigation of text similarity algorithms. This evaluation involves ten algorithms from two categories of text similarity measures we have described.

3. Algorithms

3.1. Text similarity algorithms

Different approaches have been promoted to measure the similarity between one text with another. The method is divided into two major groups Machine learning approach, and the Lexicon-based approach; as shown in Fig. 1. These approaches will be detailed in the following Subsections.

3.1.1. Machine Learning Approach

The machine learning approach relies on the famous ML algorithms to solve the SA as a regular text classification problem that makes use of syntactic and/or linguistic features. Text Classification Problem Definition: We have a set of training records D = {X1, X2, ..., Xn} where each record is labelled to a class. The classification model is related to the features in

the underlying record to one of the class labels. Then for a given instance of an unknown class, the model is used to predict a class label for it. The hard classification problem is when only one label is assigned to an instance. The soft classification problem is when a probabilistic value of labels is assigned to an instance

1. Supervised learning

The supervised learning methods depend on the existence of labelled training documents. There are many kinds of supervised classifiers in literature. In the next subsections, we present in brief detail some of the most frequently used classifiers in SA.

2. Probabilistic classifiers.

Probabilistic classifiers use mixture models for classification. The mixture model assumes that each class is a component of the mixture. Each mixture component is a generative model that provides the probability of sampling a particular term for that component. These kinds of classifiers are also called generative classifiers. Three of the most famous probabilistic classifiers are discussed in the next subsections

3. Naive Bayes Classifier (NB).

The Naive Bayes classifier is the simplest and most commonly used classifier. The Naïve Bayes classification model computes the posterior prob- the ability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label. set belongs to a particular label.

4. Bayesian Network (BN).

The main assumption of the NB classifier is the independence of the features. The other extreme assumption is to assume that all the features are fully dependent. This leads to the Bayesian Network model which is a directed acyclic graph whose nodes represent random variables, and edges represent conditional dependencies. BN is considered a complete model for the variables and their relationships. Therefore, a complete joint probability distribution (JPD) over all the variables, is specified for a model. In-Text mining, the computation complexity of BN is very expensive; that is why it is not frequently used.

5. Linear classifiers.

Given X ¼ fx1 ...... :xng is the normalized document word frequency, vector A ¼ fa1 ...... ang is a vector of linear coefficients with the same dimensionality as the feature space, and b is a scalar; the output of the linear predictor is defined as p ¼ A:X þ b, which is the output of the linear classifier. The predictor p is a separating hyperplane between different classes. There are many kinds of linear classifiers; among them is Support Vector Machines (SVM) [70,71] which is a form of classifier that attempt to determine good linear separators between different classes.

6. Decision tree classifiers.

A decision tree classifier provides a hierarchical decomposition of the training data space in which a condition on the attribute value is used to divide the data [76]. The condition or predicate is the presence or absence of one or more words. The division of the data space is done recursively until the leaf nodes contain certain minimum numbers of records which are used for classification. There are other kinds of predicates which depend on the similarity of documents to correlate sets of terms which may be used to further partition documents. The different kinds of splits are Single Attribute split which use the presence or absence of particular words or phrases at a particular node in the tree to perform the split. Similarity-based multi-attribute split uses documents or frequent word clusters and the similarity of the documents to these word clusters to perform the split. Discriminant-based multi-attribute split uses discriminants such as the Fisher discriminate for performing the split

3.1.2. Weakly, semi and unsupervised learning

The main purpose of text classification is to classify documents into a certain number of predefined categories. To accomplish that, a large number of labelled training documents are used for supervised learning, as illustrated before. In-text classification, it is sometimes difficult to create these labelled training documents, but it is easy to collect the unlabeled documents. The unsupervised learning methods overcome these difficulties. Many research works were presented in this field including the work presented by Ko and Seo [81]. They proposed a method that divides the documents into sentences and categorized each sentence using keyword lists of each category and sentence similarity measures. The concept of weak and semi-supervision is used in many applications.

The unsupervised approach was used too by Xianghua and Guo [50] to automatically discover the aspects discussed in Chinese social reviews and also the sentiments expressed in different aspects. They used the LDA model to discover multi-aspect global topics of social reviews, then they extracted the local topic and associated sentiment based on a sliding window context over the review text. They worked on social reviews that were extracted from a blog data set (2000-SINA) and a lexicon (300-SINA Hornet). They showed that their approach obtained good topic partitioning results and helped to improve SA accuracy. It helped too to discover multi-aspect fine-grained topics and associated sentiment. Other unsupervised approaches depend on semantic orientation using PMI [82] or lexical association using PMI, semantic spaces, and distributional similarity to measure the similarity between words and polarity prototypes [83].

4.1. Lexicon-based approach

Opinion words are employed in many sentiment classification tasks. Positive opinion words are used to express some desired states, while negative opinion words are used to express some undesired states. There are also opinion phrases and idioms which together are called the opinion lexicon. There are three main approaches to compiling or collecting the opinion word list. The manual approach is very time consuming and it is not used alone. It is usually combined with the other two automated approaches as a final check to avoid the mistakes that resulted from automated methods. The two automated approaches are presented in the following subsections.

4.1.1. Dictionary-based approach

A small set of opinion words is collected manually with known orientations. Then, this set is grown by searching in the well-known corpora WordNet or thesaurus for their synonyms and antonyms. The newly found words are added to the seed list then the next iteration starts. The iterative process stops when no new words are found. After the process is completed, a manual inspection can be carried out to remove or correct errors. The dictionary-based approach has a major disadvantage which is the inability to find opinion words with domain and context-specific orientations. Qiu and He [12] used a dictionary-based approach to identify sentiment sentences in contextual advertising. They proposed an advertising strategy to improve ad relevance and user experience. They used syntactic parsing and a sentiment dictionary and proposed a rule-based approach to tackle topic word extraction and consumers' attitude identification in advertising keyword extraction. They worked on web forums from automotvieforums.com. Their results demonstrated the effectiveness of the proposed approach in advertising keyword extraction and ad selection.

4.2.1. Corpus-based approach

The Corpus-based approach helps to solve the problem of finding opinion words with context-specific orientations. Its methods depend on syntactic patterns or patterns that occur 1102 W. Medhat et al. together along with a seed list of opinion words to find other opinion words in a large corpus. One of these methods was represented by Hatzivassiloglou and McKeown . They started with a list of seed opinion adjectives and used them along with a set of linguistic constraints to identify additional adjective opinion words and their orientations. The constraints are for connectives like AND, OR, BUT, EITHER-OR. . .. . .; the conjunction AND for example says that conjoined adjectives usually have the same orientation. This idea is called sentiment consistency, which is not always consistent practically. There are also adversative expressions such as but, however which are indicated as opinion changes. To determine if two conjoined adjectives are of the same or different orientations, learning is applied to a large corpus. Then, the links between adjectives form a graph and clustering is performed on the graph to produce two sets of words: positive and negative. The Conditional Random Fields (CRFs) method was used as a sequence learning technique for extracting opinion expressions. It was used too by Jiaoa and Zhou [23] to discriminate sentiment polarity by a multi-string pattern matching algorithm. Their algorithm was applied to Chinese online reviews. They established many emotional dictionaries. They worked on car, hotel and computer online reviews. Their results showed that their method has achieved high performance. Xu and Liao [25] have used a two-level CRF model with unfixed interdependencies to extract the comparative relations. This was done by utilizing the complicated dependencies between relations, entities and words, and the unfixed interdependencies among relations. Their purpose was to make a graphical model to extract and visualize comparative relations between products from customer reviews. They displayed the results as comparative relation maps for decision support in enterprise risk management. They worked on mobile customer reviews from amazon.com, epinions.com, blogs, SNS and emails. Their results showed that their method can extract comparative relations more accurately than other methods, and their comparative relation map is potentially a very effective tool to support enterprise risk management and decision making. A taxonomy-based approach for extracting feature-level opinions and mapping them into feature taxonomy was proposed by Cruz and Troyano. This taxonomy is a semantic representation of the opinionated parts and attributes of an object. Their main target was a domain-oriented OM. They defined a set of domain-specific resources which capture valuable knowledge about how people express opinions on a given domain. They used resources which were automatically induced from a set of annotated documents. They worked on three different domains (headphones, hotels and car reviews) from epinions.com. They compared their approach to other domain-independent techniques. Their results proved the importance of the domain to building accurate opinion extraction systems, as they led to an improvement of accuracy, concerning the domain-independent approaches. Using the corpus-based approach alone is not as effective as the dictionary-based approach because it is hard to prepare a huge corpus to cover all English words, but this approach has a major advantage that can help to find the domain and context-specific opinion words and their orientations using a domain corpus. The corpus-based approach is performed using a statistical approach or semantic approach as illustrated in the following subsections:

4.2.1.1. Statistical approach.

Finding co-occurrence patterns or seed opinion words can be done using statistical techniques. This could be done by deriving posterior polarities using the co-occurrence of adjectives in a corpus, as proposed by Fahrni and Klenner. It is possible to use the entire set of indexed documents on the web as the corpus for the dictionary construction. This overcomes the problem of the unavailability of some words if the used corpus is not large enough. The polarity of a word can be identified by studying the occurrence frequency of the word in a large annotated corpus of texts. If the word occurs more frequently among positive texts, then its polarity is positive. If it occurs more frequently among negative texts, then its polarity is negative. If it has equal frequencies, then it is a neutral word. Similar opinion words frequently appear together in a corpus. This is the main observation that the state of the art methods is based on. Therefore, if two words appear together frequently within the same context, they are likely to have the same polarity. Therefore, the polarity of an unknown word can be determined by calculating the relative frequency of co-occurrence with another word. This could be done using PMI. Statistical methods are used in many applications related to SA. One of them is detecting the review's manipulation by conducting a statistical test of randomness called the Runs test. Hu and Bose [31] expected that the writing style of the reviews would be random due to the various backgrounds of the customers if the reviews were written actually by customers. They worked on Book reviews from amazon.com and discovered that around 10.3% of the products are subject to online review manipulation. Latent Semantic Analysis (LSA) is a statistical approach which is used to analyze the relationships between a set of documents and the terms mentioned in these documents to produce a set of meaningful patterns related to the documents and terms. Cao and Duan [18] have used LSA to find the semantic characteristics from review texts to examine the impact of the various features. The objective of their work is to understand why some reviews receive many helpfulness votes, while others receive few or no votes at all. Therefore, instead of predicting a helpful level for reviews that have no votes, they investigated the factors that determine the number of helpfulness votes which a particular review receives (including both "yes" and "no" votes). They worked on software programs users' feedback from CNET Download.com. They showed that the semantic characteristics are more influential than other characteristics in affecting how many helpfulness vote reviews receive. The semantic orientation of a word is a statistical approach used along with the PMI method. There is also an implementation of semantic space called Hyperspace Analogue to Language (HAL) which was proposed by Lund and Burgess. Semantic space is the space in which words are represented by points; the position of each point along with each axis is somehow related to the meaning of the word. Xu and Peng [6] have developed an approach based on HAL called Sentiment Hyperspace Analogue to Language (S-HAL). In their model, the semantic orientation information of words is characterized by a specific vector space, and then a classifier was trained to identify the semantic orientation of terms (words or phrases). The hypothesis was verified by the method of semantic orientation inference from PMI (SO-PMI). Their approach produced a set of Sentiment analysis algorithms and applications: A survey of 1103 weighted features based on surrounding words. They worked on news pages and used a Chinese corpus. Their results showed that they outperformed the SO-PMI and showed advantages in modelling semantic orientation characteristics when compared with the original HAL model.

4.2.1.3. Semantic approach.

The Semantic approach gives sentiment values directly and relies on different principles for computing the similarity between words. This principle gives similar sentiment values to semantically close words. WordNet for example provides different kinds of semantic relationships between words used to calculate sentiment polarities. WordNet could be used too for obtaining a list of sentiment words by iteratively expanding the initial set with synonyms and antonyms and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word. The Semantic approach is used in many applications to build a lexicon model for the description of verbs, nouns and adjectives to be used in SA as in the work presented by Maks and Vossen [7]. Their model described the detailed subjectivity relations among the actors in a sentence expressing separate attitudes for each actor. These subjectivity relations are labelled with information concerning both the identity of the attitude holder and the orientation (positive vs. negative) of the attitude. Their model included a categorization into semantic categories relevant to SA. It provided means for the identification of the attitude holder, the polarity of the attitude and also the description of the emotions and sentiments of the different actors involved in the text. They used Dutch WordNet in their work. Their results showed that the speaker's subjectivity and sometimes the actor's subjectivity can be reliably identified. The semantics of electronic WOM (eWOM) content is used to examine eWOM content analysis as proposed by Pai and Chu. They extracted both positive and negative appraisals and helped consumers in their decision making. Their method can be utilized as a tool to assist companies in better understanding product or service appraisals, and accordingly, translate these opinions into business intelligence to be used as the basis for product/service improvements. They worked on Taiwanese Fast food reviews. Their results showed that their approach is effective in providing eWOM appraisals related to services and products. Semantic methods can be mixed with the statistical methods to perform SA tasks as in the work presented by Zhang and Xu [38] who used both methods to find product weaknesses from online reviews. Their weakness finder extracted the features and grouped explicit features by using the morpheme-based method to identify feature words from the reviews. They used the Hownet-based similarity measure to find the frequent and infrequent explicit features which describe the same aspect. They identified the implicit features with the collocation statistics-based selection method PMI. They have grouped product feature words into corresponding aspects by applying semantic methods. They have utilized the sentence-based SA method to determine the polarity of each aspect in sentences taking into consideration the impact of adverbs on the degree. They could find the weaknesses of the product, as it was probably the most unsatisfied aspect in customers' reviews or the aspect which is more unsatisfied when compared with their competitor's product reviews. Their results expressed the good performance of the weakness finder.

4.3. Lexicon-based and natural language processing techniques

Natural Language Processing (NLP) techniques are sometimes used with the lexicon-based approach to find the syntactical structure and help in finding the semantic relations. Moreo and Romero [37] used NLP techniques in preprocessing stage before they used their proposed lexicon-based SA algorithm. Their proposed system consists of an automatic focus detection module and a sentiment analysis module capable of assessing user opinions of topics in news items which use a taxonomy lexicon that is specifically designed for news analysis. Their results were promising in scenarios where colloquial language predominates. The approach for SA presented by Caro and Grella [35] was based on a deep NLP analysis of the sentences, using dependency parsing as a pre-processing step. Their SA algorithm relied on the concept of Sentiment Propagation, which assumed that each linguistic element like a noun, a verb, etc. can have an intrinsic value of sentiment that is propagated through the syntactic structure of the parsed sentence. They presented a set of syntactic-based rules that aimed to cover a significant part of the sentiment salience expressed by a text. They proposed a data visualization system in which they needed to filter out some data objects or contextualize the data so that only the information relevant to a user query is shown to the user. To accomplish that, they presented a context-based method to visualize opinions by measuring the distance, in the textual appraisals, between the query and the polarity of the words contained in the texts themselves. They extended their algorithm by computing the context-based polarity scores. Their approach was approved high efficient after applying it to a manual corpus of 100 restaurant reviews. Min and Park [39] have used NLP from a different perspective. They used NLP techniques to identify tense and time expressions along with mining techniques and a ranking algorithm. Their proposed metric has two parameters that capture time expressions related to the use of products and product entities over different purchasing periods. They identified important linguistic clues for the parameters through an experiment with crawled review data, with the aid of NLP techniques. They worked on product reviews from amazon.com. Their results showed that their metric was helpful and free from undesirable biases.

Similarity Result

No	Approach/Algorithm	Pair 1	Pair 2	Pair 3
1	Jaro-Winkler	0.8333	0	0.7714
2	N-gram	0.375	1.0	0.5
3	Cosine similarity	0.4999	0	0
4	Jaccard	0.5	0	0.2
5	LSA	0.1485	0.5080	0.1164
6	Wu Palmer	0.5000	0.9091	0.8696
7	Lin	0.1647	0.7355	0.0000
8	Path	0.1429	0.3333	0.2500
9	Monge Elkan	0.7500	0.4000	0.3714
10	SoftTFIDF	0.8333	0	0

Table 1. Lexical and Semantic

4. Conclusion

This Paper has summarized surveys of measurements of text similarity categorized into two major groups: Machine Learning Approach & Lexicon-Based Approach. The most common and familiar algorithms in each category have also been reviewed. This survey paper presented an overview of the recent updates in SA algorithms and applications. After analyzing these articles, it is clear that the enhancements of SC and FS algorithms are still an open field for research. Naı¨ve Bayes and Support Vector Machines are the most frequently used ML algorithms for solving SC problems. They are considered a reference model which many proposed algorithms are compared to. The interest in languages other than English in this field is growing as there is still a lack of resources and research concerning these languages. The most common lexicon source used is WordNet which exists in languages other than English. Building resources, used in SA tasks, are still needed for many natural languages. Information from micro-blogs, blogs and forums as well as a news source, is widely used in SA recently. This media information plays a great role in expressing people's feelings, or opinions about a certain topic or product. Using social network sites and micro-blogging sites as a source of data still needs deeper analysis. There are some benchmark data sets, especially in reviews like IMDB which are used for algorithm evaluation. In many applications, it is important to consider the context of the text and the user preferences. That is why we need to make more research on context-based SA. Using TL techniques, we can use related data to the domain in question as training data. Using NLP tools to reinforce the SA process has attracted researchers recently and still needs some enhancement

References

[1] A. Yunianta, O. M. Barukab, N. Yusof, N. Dengen, H. Haviluddin, and M. S. Othman,

“Semantic data mapping technology to solve semantic data problem on heterogeneity aspect,” Int. J. Adv. Intell. Informatics, vol. 3, no. 3, pp. 161–172, Dec. 2017, doi: https://doi.org/10.26555/ijain.v3i3.131.

[2] W. H. Gomaa and A. A. Fahmy, “A survey of text similarity approaches,” Int. J. Comput. Appl., vol. 68, no. 13, 2013, doi: https://doi.org/10.5120/11638-7118.

[3] E. Y. Hidayat, F. Firdausillah, K. Hastuti, I. N. Dewi, and A. Azhari, “Automatic Text Summarization Using Latent Drichlet Allocation (LDA) for Document Clustering,” Int. J. Adv. Intell. Informatics, vol. 1, no. 3, p. 132, Dec. 2015, doi: https://doi.org/10.26555/ijain.v1i3.43.

[4] R. W. Barron and L. Henderson, “The effects of lexical and semantic information on same-different visual comparison of words,” Mem. Cognit., vol. 5, no. 5, pp. 566–579, Sep. 1977, doi: https://doi.org/10.3758/BF03197402.

[5] J. Wang, G. Li, and J. Fe, “Fast-join: An efficient method for fuzzy token matching based string similarity join,” in 2011 IEEE 27th International Conference on Data Engineering, 2011, pp. 458–469, doi: https://doi.org/10.1109/ICDE.2011.5767865.

[6] R. W. Hamming, “Error Detecting and Error Correcting Codes,” Bell Syst. Tech. J., vol. 29, no. 2, pp. 147– 160, Apr. 1950, doi: https://doi.org/10.1002/j.1538-7305.1950.tb00463.x.

[7] V. I. Levenshtein, “Binary codes capable of correcting spurious insertions and deletions of ones,” Probl. Inf. Transm., vol. 1, no. 1, pp. 8–17, 1965.

[8] F. J. Damerau, “A technique for computer detection and correction of spelling errors,” Commun. ACM, vol. 7, no. 3, pp. 171–176, Mar. 1964, doi: https://doi.org/10.1145/363958.363994.

[9] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” J. Mol. Biol., vol. 48, no. 3, pp. 443–453, Mar. 1970, doi: https://doi.org/10.1016/0022-2836(70)90057-4.

[10] R. A. Wagner and M. J. Fischer, “The String-to-String Correction Problem,” J. ACM, vol. 21, no. 1, pp. 168–173, Jan. 1974, doi: https://doi.org/10.1145/321796.321811.

[11] T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences,” J. Mol. Biol., vol. 147, no. 1, pp. 195–197, Mar. 1981, doi: https://doi.org/10.1016/0022-2836(81)90087-5.

[12] M. A. Jaro, “Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida,” J. Am. Stat. Assoc., vol. 84, no. 406, pp. 414–420, Jun. 1989, doi: https://doi.org/10.1080/01621459.1989.10478785.

[13] W. E. Winkler, “String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of Record Linkage.,” p. 8, 1990, available at: http://files.eric.ed.gov/fulltext/ED325505.pdf.

[14] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, “Syntactic clustering of the Web,” Comput. Networks ISDN Syst., vol. 29, no. 8–13, pp. 1157–1166, Sep. 1997, doi: https://doi.org/10.1016/S0169- 7552(97)00031-7.

[15] G. Kondrak, “N-gram similarity and distance,” in International symposium on string processing and information retrieval, 2005, pp. 115–126, doi: https://doi.org/10.1007/11575832_13.

[16] A. M. Mahdi and S. Tiun, “Utilizing wordnet for instance-based schema matching,” in Proceedings of the International Conference on Advances in Computer Science and Electronics Engineering (CSEE 2014), pp. 59– 63, available at : http://www.academia.edu/download/34671264/ahmed_CSEE_2014.pdf.

[17] L. Gravano et al., “Approximate string joins in a database (almost) for free,” in VLDB, 2001, vol. 1, pp. 491– 500, available at : http://www.vldb.org/conf/2001/P491.pdf.

[18] M. Yu, G. Li, D. Deng, and J. Feng, “String similarity search and join: a survey,” Front. Comput. Sci., vol. 10, no. 3, pp. 399–417, Jun. 2016, doi: https://doi.org/10.1007/s11704-015-5900-5.

[19] M. Y. Bilenko, “Learnable similarity functions and their application to record linkage and clustering,” 2006.

[20] P. Jaccard, “Étude comparative de la distribution florale dans une portion des Alpes et des Jura,” Bull Soc Vaudoise Sci Nat, vol. 37, pp. 547–579, 1901.

[21] L. R. Dice, “Measures of the Amount of Ecologic Association Between Species,” Ecology, vol. 26, no. 3, pp. 297–302, Jul. 1945, doi: https://doi.org/10.2307/1932409.

[22] A. Bhattacharya, “On a measure of divergence of two multinomial populations,” Sankhya. v7, pp. 401–406.

[23] E. F. Krause, Taxicab geometry: An adventure in non-Euclidean geometry. Courier Corporation, 1975.

[24] J. H. Friedman, “On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality,” Data Min. Knowl. Discov., vol. 1, no. 1, pp. 55–77, 1997, doi: https://doi.org/10.1023/A:1009778005914.

[25] A. Kulkarni, C. More, M. Kulkarni, and V. Bhandekar, “Text Analytic Tools for Semantic Similarity,” Imp.

J. Interdiscip. Res., vol. 2, no. 5, 2016, available at: http://imperialjournals.com/index.php/IJIR/article/view/688.

[26] K. Lund, “Semantic and associative priming in high-dimensional semantic space,” in Proc. of the 17th Annual conferences of the Cognitive Science Society, 1995, 1995, pp. 660–665.

[27] T. K. Landauer and S. T. Dumais, “A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.,” Psychol. Rev., vol. 104, no. 2, pp. 211–240, 1997, doi: https://doi.org/10.1037/0033-295X.104.2.211.

[28] E. Gabrilovich and S. Markovitch, “Computing semantic relatedness using wikipedia-based explicit semantic analysis.,” in IJcAI, 2007, vol. 7, pp. 1606–1611, available at: http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-259.pdf.

[29] R. L. Cilibrasi and P. M. B. Vitanyi, “The Google Similarity Distance,” IEEE Trans. Knowl. Data Eng., vol. 19, no. 3, pp. 370–383, Mar. 2007, doi: https://doi.org/10.1109/TKDE.2007.48.

[30] P. Kolb, “Disco: A multilingual database of distributionally similar words,” Proc. KONVENS-2008, Berlin, vol. 156, 2008, available at: http://www.ling.uni-potsdam.de/~kolb/KONVENS2008-Kolb.pdf.

[31] R. Mihalcea, C. Corley, C. Strapparava, and others, “Corpus-based and knowledge-based measures of text semantic similarity,” in AAAI, 2006, vol. 6, pp. 775–780, available at: http://www.aaai.org/Papers/AAAI/2006/AAAI06-123.pdf.

[32] A. Budanitsky and G. Hirst, “Evaluating WordNet-based Measures of Lexical Semantic Relatedness,”

Comput. Linguist., vol. 32, no. 1, pp. 13–47, Mar. 2006, doi: https://doi.org/10.1162/coli.2006.32.1.13.

[33] T. Slimani, “Description and Evaluation of Semantic Similarity Measures Approaches,” Int. J. Comput. Appl., vol. 80, no. 10, pp. 25–33, Oct. 2013, doi: https://doi.org/10.5120/13897-1851.

[34] J. J. Lastra-Díaz, A. García-Serrano, M. Batet, M. Fernández, and F. Chirigati, “HESML: A scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset,” Inf. Syst., vol. 66, pp. 97–118, Jun. 2017, doi: https://doi.org/10.1016/j.is.2017.02.002.

[35] L. Meng, R. Huang, and J. Gu, “A review of semantic similarity measures in wordnet,” Int. J. Hybrid Inf. Technol., vol. 6, no. 1, pp. 1–12, 2013, available at: https://pdfs.semanticscholar.org/da95/ceaf335971205f83c8d55f2292463fada4ef.pdf.

[36] R. Rada, H. Mili, E. Bicknell, and M. Blettner, “Development and application of a metric on semantic nets,” IEEE Trans. Syst. Man. Cybern., vol. 19, no. 1, pp. 17–30, 1989, doi: https://doi.org/10.1109/21.24528.

[37] J. J. Lastra-D’\iaz and A. Garc’\ia-Serrano, “A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet,” 2016, available at: http://e-spacio.uned.es/fez/eserv/bibliuned:DptoLSI-ETSI-Informes-Jlastra-refinement/Refinement_Espace_LastraGarcia.pdf.

[38] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. M. Petrakis, and E. E. Milios, “Semantic similarity methods in wordNet and their application to information retrieval on the web,” in Proceedings of the seventh ACM international workshop on Web information and data management - WIDM ’05, 2005, p. 10, doi: https://doi.org/10.1145/1097047.1097051.

[39] A. Tversky, “Features of similarity.,” Psychol. Rev., vol. 84, no. 4, pp. 327–352, 1977, doi: https://doi.org/10.1037/0033-295X.84.4.327.

[40] T. B. Huedo-Medina, J. Sánchez-Meca, F. Mar’in-Mart’inez, and J. Botella, “Assessing heterogeneity in meta-analysis: Q statistic or I2 index?,” Psychol. Methods, vol. 11, no. 2, p. 193, 2006.

[41] A. E. Monge, C. Elkan, and others, “The Field Matching Problem: Algorithms and Applications.,” in KDD, 1996, pp. 267–270, available at : http://www.aaai.org/Papers/KDD/1996/KDD96-044.pdf.

[42] W. Cohen, P. Ravikumar, and S. Fienberg, “A comparison of string metrics for matching names and records,” in Kdd workshop on data cleaning and object consolidation, 2003, vol. 3, pp. 73–78, available at: https://www.cs.cmu.edu/afs/cs/Web/People/wcohen/postscript/kdd-2003-match-ws.pdf.

[43] C. Lin, D. Liu, W. Pang, and Z. Wang, “Sherlock: A Semi-automatic Framework for Quiz Generation Using a Hybrid Semantic Similarity Measure,” Cognit. Comput., vol. 7, no. 6, pp. 667–679, Dec. 2015, doi: https://doi.org/10.1007/s12559-015-9347-7.

[44] M. Al-Hassan, H. Lu, and J. Lu, “A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system,” Decis. Support Syst., vol. 72, pp. 97–109, Apr. 2015, doi: https://doi.org/10.1016/j.dss.2015.02.001.

[45] I. Atoum and A. Otoom, “Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus,” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 9, pp. 124–130, 2016, doi: 10.14569/IJACSA.2016.070917, available at: http://thesai.org/Publications/ViewPaper?Volume=7&Issue=9&Code=ijacsa&SerialNo=17.

[46] Classification of Network Traffic using Ensemble Methods to Secure Internet, International Conference on Smart Innovations in Design, Environment,

Management, Planning and Computing (ICSIDEMPC 2020)-IEEE.

[47] Deep Convolutional Neural Network Based Intrusion Detection System, International Journal of Multidisciplinary Educational Research (IJMER).

[48] Zhang, P.; Boulos, K.M.N. Chapter 50-blockchain solutions for healthcare. In Precision Medicine for Investigators, Practitioners and Providers; Faintuch, J.,

Faintuch, S., Eds.; Academic Press: Cambridge, MA, USA, 2020; pp. 519–524.

[49] Patrick, B. Meet boston's digital twin. Esri. Newsroom Blog. 2018.

[50] Schwartz, S.M.; Wildenhaus, K.; Bucher, A.; Byrd, B. Digital twins and the emerging science of self: Implications for digital health experience design and

"small" data. Front. Comput. Sci. 2020,

[51] Masison, J.; Beezley, J.; Mei, Y.; Ribeiro, H.; Knapp, A.C.; Vieira, S.L.; Adhikari, B.; Scindia, Y.; Grauer, M.; Helba, B.; et al. A modular computational

framework for medical digital twins. Proc. Natl. Acad. Sci. USA 2021, 118, e2024287118.

[52] World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance; World Health Organization: Geneva, Switzerland,

2021;

[53] Rao, D.J.; Mane, S. Digital twin approach to clinical DSS.