IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 9, Issue.4, April 2021 E-ISSN: 2347-2693

Text Similarity on Native Languages Documents

Ramandeep Kaur [1*], Prabhjeet Kaur[2]

[1,2] Sachdeva Engineering College for Girls , Gharuan Punjab India

**Corresponding Author: ramangrewal719@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v9i4.1519 | Available online at: www.ijcseonline.org

Received: 25/Mar/2021, Accepted: 02/Apr/2021, Published: 30/Apr/2021

***Abstract*—** Text similarity of text measuring is a challenging task when text is in local languages and large in amount. Text measuring tools are easily available in the market but for regional languages very few tools are available. To figure out we have introduced a text similarity in native languages. In this paper, we are highlighting the Punjabi language where we find out that cosine similarity measures the accuracy of the Punjabi documents with other Punjabi documents. Text in both documents is divided into n-grams and then the common n-grams are found. The text in the documents is subject to pre-processing, which includes tokenization and punctuation removal, followed by stop words removal and stemming. After the preprocessing step, the similarity score is calculated using the cosine similarity. The purpose of doing this is to one step toward highlighting native languages. The features, performance, advantages, and disadvantages of various similarity measures are discussed. In this paper, we provide an efficient evaluation of all these measures and help the researchers to select the best measure according to their requirement.

***Keywords*—**Semantic similarity, Corpus-based similarity, Knowledge-based similarity, Semantic relatedness

I. INTRODUCTION

IIn the 21st century Text data exponentially grows. Text data is unstructured in nature and to extract information from such data is challenging tasks and to get relevant information from unstructured data, preprocessing or labeling the text data is necessary. Labelling text data is a tedious task as compared to non-text data such as video, images, audio etc. Once text data is well labeled, training that data is the easiest approach. Label Text data is used in supervised learning, to perform text classification whereas for clustering documents we require an unsupervised learning approach of ML. Document clustering is the technique where we categorize the text data into various categories, to do so we use feature engineering on it. Segregating the documents into various categories requires a lot of effort and manual work, if language is not much popular. In India there are 22 major languages, according to Indian constitution[27]. Most languages labelling are still pending but almost all languages are interrelated to each other, for example Dravidian languages- originated in (South India) mainly, Tamil, Malayalam, Telugu and Kannada. Text data clustering is also a popular terminology in terms of NLP. There are various challenges while doing clustering documents,one of the common issues while doing clustering are same sentence different meaning or same word used for different sentences. There are various questions while implementing or dealing with text documents such as: measuring similarities, distance between appropriate documents, calculating metric, how and when clustering document, when we get perfect clustering etc. All these common questions arise while working on document clustering [14]. Various techniques are combined together to increase efficiency of the modules, for example, using text similarity and distance metrics are helpful to make a document clustering module.Using text similarity we can also generate topics of the documents. Topic modeling, clustering, summarization, semantic all require similarity check [7]. Semantic similarity is information gathering from web resources using crawling to integrate all text resources and measure similarity between them , even this approach is also used to check plagiarism. Semantic similarity performs on words/sentence/paragraph/ to calculate distance based on that it analyzes how much similarly two entities , if distance is lesser than similarity is more. Similarity can also be checked through lexical, word sequence and its characteristics relates to lexical whereas its meaning related to semantics, lexical and semantically are parelley in used to measure accuracy. Various types of algorithms are measuring similarity word, sentence or phrase, the string based algorithm is suitable for lexical analysis and corpus and knowledge based suitable for semantic. Semantic similarity measures are being intensively used in various applications of knowledge based and semantic information retrieval systems for identifying an optimal match between user query terms and documents. It is also used in word sense disambiguation for identifying the correct sense of the term in the given context. Semantic similarity and semantic relatedness are two different concepts but related to each other. For example, “mother” and “child” are related terms but are not similar since they have different meanings[5]. The survey paper is structured as follows:

Section two highlights the related work in the field of semantic similarity measures. Section three describes the Corpus based similarity measures and Section four describes the knowledge based similarity measure. Section five highlights various semantic similarity measures in a tabular form. Section six presents the summary of the survey in the form of conclusion.

II. RELATED WORK

A. *Information Retrieval (IR)*

Information retrieval (IR) is the technique to get exact information from corpus or the documents that are related to the search entity. For example, On google when we write in search bar google automatically find out exact match related to search the technique, this approach is Information retrieval that make relevance document, content related to search or in other word based on rank, score and similarity, we easily get similar output on google.[11]

B. *Feature Engineering*

Feature engineering is a domain or way to make machine learning approach easier. Generally preprocessing, Bag of Words, TF-IDF, and word vectorization models are usually used to filter or extract meaningful information from raw data, so feature engineering is a mixture of all to make it easier for ML/DL [10][8].

C. *Text Normalization*

Text Normalization is a common approach in NLP, and used in various text related valuable techniques such summarization ,clustering etc. The most common steps in text normalization are tokenization where sentences are tokenized into words, removing stop words like *a, an, in* etc, then lemmatization and stemming to remove in words and use of dictionary to make sentence correct[3][4]

D. *Similarity Measures*

Similarity measures are popular terms in text similarity measures or text clustering. similarity measures use degree of closeness in term of word, sentence, documents. Similarity measures help to measure the close relation between two sentences, paragraphs & documents, based on similarity measures the performance of text based models is quite better. Various algorithms that are used in text measuring are generally derived from similarity measures. The main principle behind this measure is distance between two documents, how far and near two documents from each other are clearly understandable, in the paper[13], text similarity of english documents easily calculated.

Consider a separation measure d and two substances (state they are archives in our unique situation) x and y. The separation among x and y, which is utilized to decide the level of likeness between them, can be spoken to as d(x, y), yet the measure d can be called as a separation metric of similarity[1][2] if and just in the event that it fulfills the accompanying four conditions:

(a) The separation estimated between any two elements, state x and y, should consistently be non-negative, that is, dxy,()³0. (b) The separation between two elements should consistently be zero if and just on the off chance that they are both indistinguishable, that is, dxy iff xy,()³=0 .(c) This separation measure ought to consistently be symmetric, which implies that the good ways from x to y is consistently equivalent to the good ways from y to x. Numerically this is spoken to as dxy dyx,,()=(). (d) This separation measure ought to fulfill the triangle imbalance property, which can be numerically spoken to dxz dxy dyz

A. Text Similarity

Text similarity is implemented on word,sentence,paragraph & documents. To measure similarity there are two approaches one is lexical another is semantic both are popular and useful, lexical observing the content of text with respect to syntax, structure based on that it measures similarity [15][16]. whereas semantics use meaning, grammar of the content,how close the meaning of one word with respect to another [17].

The most mainstream zone is lexical likeness, in light of the fact that the strategies are more clear, simple to actualize, and you can likewise cover a few pieces of semantic closeness utilizing straightforward models like the Bag of Words. Normally separation measurements will be utilized to quantify likeness scores between text elements, and the accompanying two wide zones of text closeness areTerm closeness,Document comparability[26].

B. Analyzing Term Similarity

Analyzing term similarity is very popular while using search engines. Several applications are available such as when we search on Google, Google automatically corrects or gives suggestions based on similar search by most of the time, similarly in google doc autocomplete or autosuggestion is commonly used Term Similarity approach. Different words representation and its metrics are clearly explained using the Bag of Characteristics, Vectorization in paper[18].

C. *Hamming Distance*

This metric is like the Hamming separation theoretically used in information and communication, where we take away the contrast between each pair of characters at each position of the two strings[19]. Numerically it tends to be indicated as where u and v are the two terms of length n.

$$hd(u,v)=\sum_{i=1}^{n}(u_i \neq v_i)$$

D. Manhattan Distance

Adroitly it is like a Hamming, where instead of calculating mismatches, we deduct the contrast between each pair of characters at each position of the two strings.

Formall[20],Numerically it very well may be meant as,where u and v are the two terms of length n.

$$md(u,v)=\|u-v\|_1=\sum_{i=1}^{n}|u_i-v_i|$$

E. Euclidean Distance

The "ED" is additionally called Euclidean norm[22], L2 standard, or L2 separation and is characterized as the most brief straight-line separation between two focuses. Mathematically this can be signified where the two focuses u and v are vectorized text terms in our situation, each having length n.

$$ed(u,v)=\|u-v\|_2=\sqrt{\sum_{i=1}^{n}(u_i-v_i)^2}$$

F. Cosine Distance and Similarity

The Cosine distance and similarity is the way to find out the actual similarity between two documents or vectors [24]. Cosine similarity use cosine function to measure similarity of non-zero positive vectors,if the value is closer to 1 meansvector are similar to each other whereas if value is closer to -1, it mean vectors are converse to each other

$$cs(u,v)=\cos(\theta)=\frac{u\cdot v}{\|u\|\,\|v\|}=\frac{\sum_{i=1}^{n}u_i v_i}{\sqrt{\sum_{i=1}^{n}u_i^2}\sqrt{\sum_{i=1}^{n}v_i^2}}$$

III. PROPOSED APPROACH

Document similarity analysis is an important part of Information Retrieval. Two documents can be treated as similar if they are semantically close and are related to similar concepts. Similarity measures can also be used to detect duplicates. As the number of information resources and document quantity is exploding, we need efficient tools to help the users to analyse the documents for how similar they are and/or to detect plagiarism in the documents, if any. Punjabi language is the world's 10th most spoken language as per *Nationalencyklopedin 2010* [7], but not much resources are available for text processing in Punjabi. A lot of work and research has been done on English and we get many tools and inbuilt libraries for carrying out different operations and manipulation of text in English but to the best of our knowledge, not much work has been done on Document similarity on Punjabi documents till date. In this paper, the proposed method finds the similarity between two given documents using the Cosine similarity approach, and detects similarity based on the *n-grams* approach. A user friendly GUI has also been developed using python's Tkinter for the same *various steps has been taken to implement cosine and gui interface such as Tokenization,* Language detection, Stemming, Lemmatization, POS tagging, Identify named entities (NER recognition), Chunking etc.

Cosine likeness is a proportion of similitude between two non-zero vectors of an internal item space that measures the cosine of the point between them.

The normalised tokens are fed to the module that calculates the similarity between the documents and returns the similarity index as an output.

The final step is to output the similarity index/measure obtained as a result of cosine similarity calculation algorithm on the GUI.

In this paper we are using Cosine Similarity to measure the accuracy between two documents of Punjabi for that we are using python language In the next section we have experimentally proof.

IV. RESULTS AND DISCUSSION

Our model is designed to check text similarity between two Punjabi documents written in Gurmukhi script. For checking the similarity of documents, cosine similarity has been used. Other techniques such as frequency distribution, *tf-idf* and jaccard distance were also analyzed to find the similarity between the documents but the cosine similarity technique was found to be giving the best results. Jaccard similarity gave 100% similarity whereas Cosine similarity gave 99.89% similarity in case of documents that have exactly the same content except one extra line in one of the documents. So, cosine similarity is found to be better. For plagiarism detection, ngrams have been used. Plagiarism index has been find out by changing the value of n-grams. By experimenting on the various values of n-grams, it is found out that using n-gram value greater than 5 degrades the performance of the system in terms of similarity. When we work on unigram (1-grams) or bigram (2-grams) we see that it considers the text as a bag of words. This shows that they don't contribute to the similarity of two documents. For a low value of n, a large number of n-grams will be generated , so the processing time will be more. So by comparing the processing time and accuracy, the value of n in n-gram is taken to be 5.

There are four possible cases in this model:

1. *When exactly same documents are uploaded*
2. *When documents about same topic are uploaded*
3. *When documents about different topic are uploaded*
   When documents of language other than Punjabi are uploaded

The below figure is when documents about same topic are uploaded

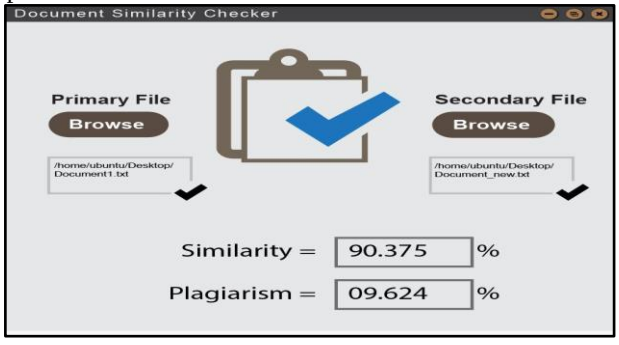

Fig. 1 Similarity & Plagiarism Tool

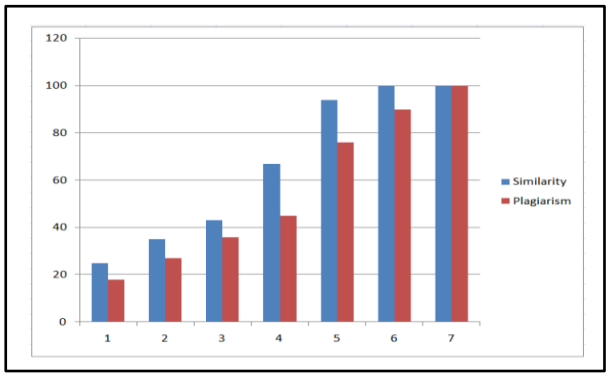

Fig. 2 Graphical Representation of Plagiarism & similarity

With this figure we have concluded that similarity is alway less than or equal to plagiarism and our model is better to find Similarity & plagiarism between Punjabi documents.

V. CONCLUSION AND FUTURE SCOPE

Text Similarity for native language is a complex task and has many issues associated with it. There are a number of tools available for English, but not much work has been done in the field of finding similarity between Punjabi documents. For similarity calculation, the cosine similarity approach has been used and for plagiarism detection, n-grams approach has been used. A simple and user-friendly interface has been created for easy access. Document similarity and Plagiarism detection for Punjabi documents can be useful in various fields such as human resource applications for finding similar employees using automatic CV mining, clustering of documents and auto-categorization, Patent research by matching potential patent applications against a corpus of existing patent grant and so on.

REFERENCES

[1] Gentner, Dedre; Markman, Arthur B. (1997). "Structure mapping in analogy and similarity" (PDF). American Psychologist. 52 (1): 45–56. CiteSeerX 10.1.1.87.5696. doi:10.1037/0003-066X.52.1.45. Archived from the original on 2016-03-24.

[2] Greg Aloupis, Thomas Fevens, Stefan Langerman, Tomomi Matsui, Antonio Mesa, Yurai Nunez, and David Rappaport, and Godfried T. Toussaint, "Algorithms for computing geometric measures of melodic similarity," Computer Music Journal, Vol. 30, No. 3, Fall 2006, pp. 67–76

[3] Gentner, Dedre; Markman, Arthur B. (1997). "Structure mapping in analogy and similarity" (PDF). American Psychologist. 52 (1): 45–56. CiteSeerX 10.1.1.87.5696. doi:10.1037/0003-066X.52.1.45. Archived from the original on 2016-03-24.

[4] Balkova, Valentina; Sukhonogov, Andrey; Yablonsky, Sergey (2003). "Russian WordNet From UML-notation to Inter net/Intranet Database Implementation" (PDF). GWC 2004 Proceedings: 31–38. Retrieved 12 March 2017.

[5] Novotný, Vít (2018). Implementation Notes for the Soft Cosine Measure. The 27th ACM International Conference on Information and Knowledge Management. Torun, Italy: Association for Computing Machinery. pp. 1639–1642. arXiv:1808.09407. doi:10.1145/3269206.3269317. ISBN 978-1-4503-6014-2.

[6] Langer, Stefan; Gipp, Bela (2017). "TF-IDuF: A Novel Term-Weighting Scheme for User Modeling based on Users' Personal Document Collections" (PDF). IConference.

[7] Rogers, David J.; Tanimoto, Taffee T. (1960). "A Computer Program for Classifying Plants". Science. 132 (3434): 1115–1118. doi:10.1126/science.132.3434.1115.

[8] A Survey of Encoding Techniques for Reducing Data-Movement Energy", JSA, 2018

[9] Winkler, W. E. (2006). "Overview of Record Linkage and Current Research Directions" (PDF). Research Report Series, RRS.

[10]Andoni, Alexandr; Krauthgamer, Robert; Onak, Krzysztof (2010). Polylogarithmic approximation for edit distance and the asymmetric query complexity. IEEE Symp. Foundations of Computer Science (FOCS). arXiv:1005.4033. Bibcode:2010arXiv1005.4033A. CiteSeerX 10.1.1.208.2079.

[11]Backurs, Arturs; Indyk, Piotr (2015). Edit Distance Cannot Be Computed in Strongly Subquadratic Time (unless SETH is false). Forty-Seventh Annual ACM on Symposium on Theory of Computing (STOC). arXiv:1412.0348. Bibcode:2014arXiv1412.0348B.

[12]Chapman, S. (2006). SimMetrics: a java & c# .net library of similarity metrics, http://sourceforge.net/projects/simmetrics/.

[13]Hall , P. A. V. & Dowling, G. R. (1980) Approximate string matching, Comput. Surveys, 12:381-402.

[14]Peterson, J. L. (1980). Computer programs for detecting and correcting spelling errors, Comm. Assoc. Comput. Mach., 23:676-687.

[15]Jaro, M. A. (1989). Advances in record linkage methodology as applied to the 1985 census of Tampa Florida,Journal of the American Statistical Society,vol. 84, 406, pp 414-420.

[16]Jaro, M. A. (1995). Probabilistic linkage of large public health data file, Statistics in Medicine 14 (5-7), 491-8

[17]Winkler W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage,Proceedings of the Section on Survey Research Methods, American Statistical Association, 354–359

[18]Needleman, B. S. & Wunsch, D. C.(1970).A general method applicable to the search for similarities in the amino acid sequence of two proteins",Journal of Molecular Biology48(3): 443–53

[19]Smith, F. T. & Waterman, S. M. (1981).Identification of Common Molecular Subsequences, Journal of Molecular Biology147: 195–197

[20]Alberto, B. , Paolo, R., Eneko A. & Gorka L. (2010). Plagiarism Detection across Distant Language Pairs, In Proceedings of the 23rd International Conference on Computational Linguistics, pages 37–45

[21]Eugene F. K. (1987).Taxicab Geometry , Dover.ISBN0-486-25202-7

[22]Dice, L. (1945). Measures of the amount of ecologic association between species. Ecology, 26(3

[23]Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 547-579

[24]Lund, K., Burgess, C. & Atchley, R. A. (1995). Semantic and associative priming in a high-dimensional semantic space.Cognitive Science Proceedings (LEA), 660-665

[25]Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence.Behavior Research Methods, Instruments & Computers, 28(2),203-208

[26]Landauer, T.K. & Dumais, S.T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge", Psychological Review, 104

[27]https://www.mhrd.gov.in/sites/upload_files/mhrd/files/upload_document/languagebr.pdf

AUTHORS PROFILE

Er. Ramandeep kaur is presently working as an Assistant Professor in the Department of Computer Science and Engineering of Adesh College of Engineering, India. She received the degree of Bachelor of Technology(B.Tech.) in Computer Science and Engineering from the PTU. She is presently pursuing her M.Tech in Computer Science & Technology at Sachdeva Engineering College for Girls , Gharuan Punjab India. Her research interests include NLP, Network Security, SDN, Big Data and Computer Applications.

