J. Inst. Eng. India Ser. B (April 2021) 102(2):329–338
https://doi.org/10.1007/s40031-020-00501-5

ORIGINAL CONTRIBUTION

Text Similarity Measures in News Articles by Vector Space Model Using NLP

Ritika Singh[1] · **Satwinder Singh**[1]

Received: 4 June 2020 / Accepted: 7 October 2020 / Published online: 7 November 2020

Abstract The present global size of online news websites is more than 200 million. According to MarketingProfs, more than 2 million articles are published every day on the web, but Online News websites have also circulated editorial content over the internet that specifies which articles to display on their website's home pages and what articles to highlight, e.g., broad text size for main news articles. Many of the articles posted on a news website are very similar to many other news websites. The selective reporting of top news headlines and also the similarity among news across various news associations is well-identified but not very well calculated. This paper identifies the top news items on the news sites and measures the similarity between two same news items in two languages (Hindi and English) referring to the same event. To accomplish this, a highlighted headline and link extractor has been created to extract top news for both Hindi and English from Google's news feed. First, translate the Hindi news article into English by using Google translator and then compare it with English news articles. Second, we used the cosine similarity, Jaccard similarity, Euclidean distance measure to calculate news similarity score. The frequency of nouns and the next word of nouns from the news articles are also extracted. Our methodology clearly shows that we can efficiently identify top news articles and measure the similarity between news reports.

✉ Ritika Singh
ritikasingh2397@outlook.com

Satwinder Singh
satwinder.singh@cup.edu.in

[1] Department of Computer Science and Technology, Central University of Punjab-Bathinda, Bathinda, India

Keywords Bilingual news article similarity · Cosine similarity · Jaccard similarity · Euclidean distance

Introduction

A huge increase in the number of online newspaper publishing is only because of the digital technology innovations. When in the modern world so much information appears at a tremendous speed, readers need to find out if they are reading true news or false news. False news and information can endanger and confuse not only a person's life, but also an entire society, so it is very important to find out the source of information and compare it with other news. So this study has an interest in extracting online news platforms, specifically to measure the similarity of news articles across various sites. This article provides details about what news is being considered, how it is being presented and, and highlighted on a website [1]. News articles which are published on the website usually appear in similar or rectified form on several different websites. Similar and almost identical news is confusing for users. Similarity slows down the process of discovering new information about a topic, and potentially leads to missing information, if the user mistakenly recognizes two news as similar when in fact one contains new data. It is much more difficult to locate similar news items in websites. This is because of the large amount of miscellaneous content or material on these articles. Although the main news article text can be similar on two different web pages, the extraneous material on the pages may not be the same. Therefore, traditional approaches to equivalent news determination would fail [2]. First, this paper developed a method for scraping top news headline text from web pages, i.e. Google news feed websites which are present in

two different languages (Hindi and English), referring to the same event then use the extracted text to classify news pairs with the same content, avoiding any irrelevant information on the articles. By measuring a similarity score for news pairs based on a method called Cosine similarity and Jaccard similarity and Euclidean Similarity, this research can distinguish similar news articles, as well as different ones. The purpose of this paper is also to discover bilingual news articles in a comparable corpus [3]. In particular, the study is dealing with the representation of news and the measurement of the similarity among new articles. This experiment uses the similarly named entities which they include as representative features of the news. To assess the similarity between articles of the same news, this research proposing a new method focused on a knowledge base framework that aims to provide human information on the value of the category of named entities within the news [4]. In a comparable corpus with news in Hindi and English, we compared our approach to a traditional one which obtains better results. Similarity and also distance measures calculate the similarity of two documents or sentences into a single numerical value and brings out the degree of semantic similarity [5] or distance from one another. Several similarity measures have been used by the researchers, but not much work has been done on the similarity of newspapers. This study aims to compare the semantic similarity between two articles of the same news, present in two different languages (Hindi and English), to optimize human understanding. The basic concept for measuring news similarities is to identify Feature articles vectors, and thereafter measure the difference between those features. Low distance between those features implies a high level of similarity value, while a large distance in between those features implies a low level of similarity value [6]. Euclidean distance, Cosine distance, Jaccard coefficient metrics are some of the distance metrics used in document similarity computation. This study explores two separate methods of generating features from the texts: (1) the Tf-idf vectors, (2) bag of words also implements two methods for calculating textual similarity between news articles: (1) cosine similarity and Jaccard similarity with Tf-idf vectors and (2) Euclidean distance using a bag of words.

Literature Rereview

In the literature, similarity measures have been used for various purposes. In this section, some proposals are reviewed.

Atkins et al. [1] describe a technique to assess the top news headline story from a selected set of US-based news websites, and then calculate correlations across them. To do this, they first created a headline and link extractor that parses selected news websites and then searched ten US-based news site home pages for 3 months. They use a parser to extract $k = 1, 3, 10$ for each news site, the maximum number of articles. Second, the author uses the calculation of cosine similarity to quantify the similarity of news. They also provide techniques during this work to assist in analyzing archived news web pages by introducing tools for parsing select HTML news sites for Hero and headline stories using CSS selectors. Author's studies over 3 months have shown that the overall similarity decreased as the number of articles increased. Studies from the author indicate that they would set up synchronous stories for a given day besides relevant national events. This approach can be used to further examine the occasional elections that are being held.

Katarzyna Baraniak and Marcin Sydow work on tools that would support the detection and analysis of the information bias [7]. The author uses methods to automatically identify the articles reporting on the same subject, event, or entity to use them more in comparative analysis or to construct a test or training collection. Within the paper, the author explains representations of the document text and the method of similarity measures for text clustering. Which include tests such as cosine similarity, Euclidean distance, Jaccard coefficient, Pearson coefficient of correlation, and Averaged Kullback–Leibler Divergence. The author also applies a machine learning approach to recognize a similar article and develop a machine learning model that detects similar articles automatically. Identifying fragments of text concerning similar events and identifying bias in them is expected. The author is also working to expand the research study to other languages (e.g., Polish, English).

Maake Benard Magara et al., suggest a system to use 220 artificial intelligent research paper written by 8 artificial intelligence experts [8]. This work uses Recursive Partitioning, Random Forest, and improved machine learning algorithms by having an average accuracy and timing efficiency of 80.73 and 2.354628. Seconds, this algorithm typically performed quite well compared to the Boosted and even the Random Forest algorithms. More sophisticated models can be used in future studies much like the Latent Semantic Analysis (LSA), since documents can be identified as belonging to the same class even if they have no similar words and phrases. Vikas Thada and Dr. Vivek Jaglan authors used the cosine similarity, dice coefficient, Jaccard similarity algorithms [9]. The work is completed on the first 10 pages of the Google search result and will be expanded to 30–35 pages for a reliable efficiency estimate in future study. The cosine similarity eventually concluded was the best fitness compared with others for this dataset. In summary, while the initial

findings are promising, there is still a long way to go to achieve the greatest crawling efficiency possible. A systematic method proposed by Nasab et al. [10] the following points determine the similarities. (1) Article texts are divided into three sections as headings, abstracts and keywords. (2) Abstract, keywords, based on the link to the title of article weighing. (3) The weighted mean is estimated based on the description, abstract, and keyword and use Pearson's correlation method to find the similarity between person and machine scores. They have 87% accuracy in this proposed technique. Use a specialized WordNet it can also concentrate on article similarities. The proposed framework can be used for other texts that require a WordNet of that language, such as texts in Persian and other languages. M. Snover et al., explore a new way of using monolingual target data to enhance the efficiency of a statistical or predictive machine translation for news stories [11]. This method employs comparable text various texts in the target language which explore the same or equivalent stories as mentioned in the source language document. A large monolingual data set for each source document to be translated in the target language, which is searched for documents that may be similar to the source documents. The experimental results of this paper generated through the difference of the language and translation models show vital improvements over the baseline framework.

Qian et al. [12] using a comparable corpus, a bilingual dependency mapping model for bilingual lexicon building from English to Chinese. This model considers both dependent words and their relationships when measuring the similarity between bilingual words and thus offers a more precise and less noisy representation. Author's also illustrated that bilingual dependency mappings can be created and optimized automatically without human input, contributing to a medium-sized set of dependency mappings and that their impacts on Bilingual Lexicon Construction (BLC) can be fully exploited through weight learning using a simple but effective perceptron algorithm, making their approach quickly adaptable to several other language pairs.

Methodology

The major steps of the methodology are given below.

Figure 1 presents the framework of this work. The textual news data are first pre-processed before it is represented into a more structural format. The two representation methods of generating features from the text that are investigated in this study are tf-idf, and Bag of Word. Once represented into these three representation methods, each represented method is compared with three similarity measures as shown in Fig. 1 i.e. Cosine, Euclidean and

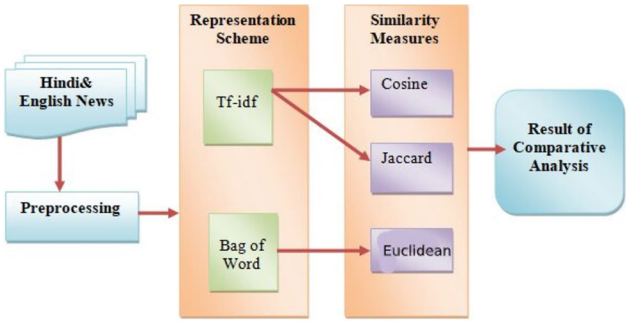

Fig. 1 A framework for comparative analysis

Jaccard similarity measures. The final step in the framework is to compare and analyze the produced results. We further explain each of the steps in detail.

The dataset used in this paper is known as 'Google News', and is publicly available [13]. Google News: Google is offering a special experience to Google News which combines all its news items into one. It provides a constant, personalized flow of newspapers from thousands of publishers and magazines grouped around. Google News is a combination of global events, local news and news stories that you've been reading. Then you can turn to Headlines to show top news from all over the world. Additional sections here allow you to delve into various topics such as sports, business and technology. And its greatest value is that this service delivered the news in 35 languages so using Google news this experiment extracts the news articles in both Hindi and English languages.

Headline and Link Extractor

A basic python library for searching and downloading live news articles from Google News feeds is GoogleNews or gnewsclient [14]. Using this, one can pick up the top headlines running on Google's news websites or check for top headlines on a particular subject (or keyword). So this experiment can use this, to extract links from both Hindi and English news that related to the same event.

Article Scraping

'Newspaper' is a Python module used to extract newspaper articles and to parse them. Newspapers are using specialized Web scrapping algorithms to extract all the valuable text from a website. This works extremely well on websites of the online newspapers. This experiment has extracted links from both Hindi and English news, so now also extract their text using the Newspaper module.

Translator

Through using this package, Google offers a language translation package for Python; words are taken from the Hindi news articles and translated into different languages (English language). Either Hindi corpus can be translated into English or English corpus can be translated into Hindi. Here we have translated Hindi corpus into English. The translation is performed at a level of the sentences. This translation also generates a map of words in various languages, from English. This research used bilingual dictionaries ranging from Hindi to English.

Pre-processing and Data Cleaning

Pre-processing steps such as the elimination of stop-words, lemmatization, and parsing letters, punctuation marks, and numbers have been completed. The words were lemmatized by WordNetLemmatizer and NLTK library took the English stop-words [15].

Vector Space Model

A mathematical model is also called the term vector model, which describes text documents as identifier variables, such as terms or tokens. Of course, the term depends on the comparisons, but usually, only words, keywords or sentences are compared.

Feature Vectors

In the Artificial Intelligence feature vector is an n-dimensional vector of computational features that describe some entity. That is a really important method of calculating semantic similarity among texts. Methods were used during this experiment to measure the function vectors is TF-IDF (Term Frequency-Inverse Document Frequency) is a simple algorithm for transforming a text into a meaningful representation of numbers. Tf-idf weight is a measure of fact which evaluates the importance of a specific word in a text. In mathematics,

$$tfi\,df\ \text{weight} = \sum_{i \in d} tf_{i,d} * \log\left(\frac{N}{df_i}\right) \tag{1}$$

where in document d, $tf_{i,d}$ is the number of occurrences of the ith term, df_i is the number of documents which contain ith term; N is the total number of documents. The sklearn-vectorized function was used to construct a tf-idf function. This whole model was constructed by using the documents, and a group of such tf-idf vectors was generated consisting of the tf idf weight of and term in the documents. Such tf-idf vectors have now been used as feature vectors to measure the similarity between articles in news-results.

Similarities Measures

Similarity function is a real-valued function that calculates the similarity between two items. The calculation of similarity is achieved by mapping distances to similarities within the vector space. This experiment provides two tests of similarity: cosine similarity, similarity with Jaccard, and Euclidean distance.

(1) *Cosine Similarity* It is a cosine angle in an n-dimensional space, between two n-dimensional vectors. This is the dot product of the two vectors, divided by-product of the two vectors' lengths (or magnitudes) [16]. The similarity of the cosine is measured by using the following formula:

$$\text{similarity}(A,B) = \frac{A.B}{||A|| \times ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

As shown in Fig. 2, suppose there is two point's p1 and p2, as the distance within these points increases the similarity between these points decreases and vice versa.

$$1 - \text{Cosine Similarity} = \text{Cosine Distance}$$

The result of the angle will show the result. If the angle is 0 between the document vectors then the cosine function is 1 and both documents are the same. If the angel is any other value then the cosine function will be less than 1. Does the angle reach -1 then the documents are completely different? Thus this way by calculating the cosine angle between the vectors of $P1$ and $P2$ decides if the vectors are pointing in the same direction or not.

(2) *Jaccard Similarity* Jaccard similarity calculates similarities among sets. It's defined as the intersection size

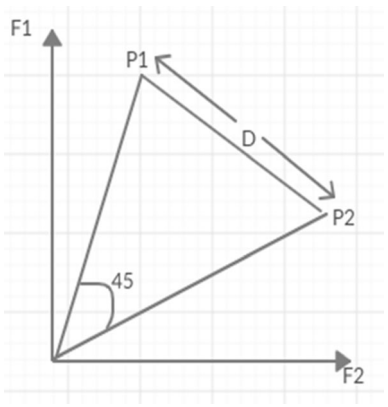

Fig. 2 Cosine similarity

divided by the union size of two sets. Jaccard similitude is determined using the formula [16] below.

$$J(A,B) = \frac{A \bigcap B}{A \bigcup B} = \frac{A \bigcap B}{`|A| + |B| - |A \bigcap B|} \quad (3)$$

where ∩ represents intersection and ∪ represents the union. In this paper, A and B are bags of words that contain news articles.

- Jaccard(*A*,*A*) = 1
- Jaccard(*A*,*B*) = 0 if $A \bigcap B = 0$
- *A* and *B* don't have to be the same size
- Always assign a number between 0 and 1.

Jaccard distance which instead of similarity measures dissimilarity between can be found by subtracting Jaccard similarity coefficient from 1:

$$JD(A,B) = 1 - JD(A,B) \quad (4)$$

$$\text{or} \quad JD(A,B) = \frac{|A \cup B| - |A \cup B|}{|A \cup B|} \quad (5)$$

(3) *Euclidean Distance* Another similarity measure in the vector space model is Euclidean distance or *L*2 distance, or Euclidean norm. This similarity measure differentiates similarity measurements from the other vector space model by not judging from the angle like the rest but rather the direct distance between the vector inputs.

As shown in Fig. 3, if there are two points like (*X*1, *Y*1) and (*X*2, *Y*2) and let us consider any dimension point so if one wants to find out the distance between (*X*1, *Y*1) and (*X*2, *Y*2) then basically use this particular parameter like Euclidean distance to check that if this particular points are nearer to each other than it will consider that this two-point are similar with each other. Euclidean distance is calculated based on the Pythagoras theorem. Let *D* represent the measure of distances between (*X*1, *Y*1) and (*X*2, *Y*2). Hence the distance from *A* to *C* can be expressed as:

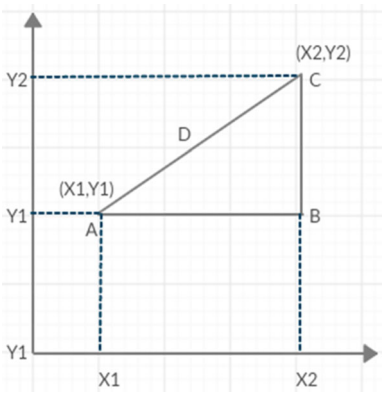

Fig. 3 Euclidean distance

$$AC^2 = AB^2 + BC^2 \quad (6)$$

$$AC = \sqrt{AB^2 + BC^2} \quad (7)$$

$$AC = \sqrt{(x2 - x1)^2 + (y2 - y1)^2} \quad (8)$$

$$|X \to Y| = \sqrt{\sum_{i=1}^{m} (X_i - Y_i)^2} \quad (9)$$

Table 1 shows a comparative analysis of the methods based on their relative pros and cons. The table also describes the application areas where the selected techniques can be used.

Similarity Score

Similarity score means that two data sets are how similar to one another. The data collection will include two separate texts as in this case. The similarity between the two texts is evaluated according to the scoring system. Euclidean distance does not find the similarity between the texts, but finds the metric, the distance between both texts [18]; there are different ways to calculate similarity:

$$\text{Similarity}(A,B) = \frac{1}{1 + \text{Distance}(A,B)} \quad (10)$$

Noun Phrase Extraction

Noun Phrase Extraction is a technique of text analysis, consisting of the automated extraction of nouns in a text. It helps to summarize the contents of a text and identify the key topics being discussed. This paper concludes that the extraction of the frequency of noun phrases and the frequency of the next word of the noun from news articles can considerably improve similarity measures. TextBlob is a Python module that is used to extract a noun [19].

Proposed Method

This paper introduces two methods for calculating the similarity between two articles of the same news, which are present in two different languages (Hindi and English), based on methods for calculating the feature vector and similarity measures.

Cosine Similarity and Jaccard Similarity with TF-IDF Vectors

The pre-processed news articles were turned into vectors of tf-idf by using a vectorized model of tf-idf. The vectors obtained were a sparse-matrix containing tf-idf weights for news article word having the dimensions of [number of

Table 1 Comparison of the pros and cons of different measures and their application area

SI.	Similarity measures	Pros	Cons	Application area
1	Cosine similarity	Both continuous and categorical variables may be used	Doesn't work effectively with nominal data [17]	Text mining, document similarity
2	Jaccard coefficient	Both continuous and categorical variables may be used	Doesn't work effectively with nominal data	Document classification
3	Euclidean distance	Easy to Compute and work well with a dataset with compact or isolated clusters [17]	Does not work with image data efficiently	Application involving interval data, DNA analysis, K-mean algorithm

news articles * number of features (distinct words)] [16]. That tf-idf weight from the matrix was now used as a feature for every text, and similarity among news articles is calculated using cosine similarity and Jaccard similarity. Sklearn's built-in cosine and Jaccard similarity module was used to measure the similarity.

Table 2 Sample pair of completely similar news

Pair	News Headlines
1	भारतवंशी राजनेता प्रीतम सिंह सिंगापुर की संसद में विपक्ष के नेता नियुक्त, हासिल की यह उपलब्धि Indian-Origin Politician Appointed Singapore's First Leader Of Opposition In Parliament
2	नई राष्ट्रीय शिक्षा नीति में कोई भाषा किसि पर थोपी नहीं गई है :के कस्तूरीरंगन No Language Is Being Imposed In The New National Education Policy: K Kasturirangan
3	UK में बिल्ली हुई Covid-19 पॉजिटिव, डॉक्टरों ने कहा, "कोरोना फैलाने के सबूत नहीं मिले" Covid-19 positive cat in UK, doctors say no evidence of corona spread found
4	भारत में होने वाला FIFA U-17 महिला वर्ल्ड कप टला FIFA Under-17 Women's World Cup, scheduled November postponed
5	8 अप्रैल को दिखेगा सुपर पिंक मून, भारत में ऐसे देखें लाइव Super Pink Moon: When to witness the brightest full moon of 2020
6	अमेरिकी राष्ट्रपति ने भारत, चीन और रूस पर लगाए गंभीर आरोप, कहा -नहीं रखते वायु गुणवत्ता का ध्यान Donald Trump also shines on India along with China, says- 'Does not take care of air quality

Bag of Words Euclidean Distance

The pre-processed documents have been described as a vector with the frequency of each word and compare how similar they are by comparing their bag of vector words. This experiment uses the bag-of-words model because the computer processes vectors much faster than a vast file of

Table 3 Sample pair of different news stories about the same topic

Pair	News Headlines
7	गवर्नर से चौथी बार मिले CM गहलोत, 31 जुलाई को विधानसभा सत्र बुलाने के लिए नहीं माने राज्यपाल Gehlot Cabinet has sent a revised proposal to Governor on Assembly session, says report
8	डीसीजीआई ने ऑक्सफोर्ड कोविड-19 वैक्सीन के ट्रायल प्रोटोकॉल को संशोधित करने को कहा Russia may launch world's first coronavirus vaccine by 10 August: Report
9	चेन्नई के अस्पताल में 3 साल के रूसी बच्चे का हुआ बर्लिन हॉर्ट ट्रांसप्लांट Berlin Heart Completes Post Approval Surveillance; Report Details Improved Outcomes for Pediatric Heart Failure Patients
10	अमेरिका ने दिखाई चीन को ताकत, शंघाई के बेहद करीब पहुंचे US के फाइटर जेट Chinese jets intercept US aircraft over East China Sea, US says
11	चीन ने लॉन्च किए तीन नए सैटेलाइट, उठाएंगे धरती के अनछुए राज से पर्दा China receives data from newly-launched mapping satellite

Table 4 Sample pair of completely dissimilar news

Pair	News Headline
12	दक्षिण चीन सागर तनाव पर अमेरिका की चेतावनी, "यदि फ्री नेशंस कुछ नहीं करते हैं तो... "Cops "Kneeling On Necks Of Black Americans": Barack Obama Attacks Trump
13	UP सरकार ने जारी की अनलॉक-3 की गाइडलाइन, पूरे अगस्त में रहेगा वीकेंड लॉकडाउन Google Parent Alphabet's Profit Drops 30% As COVID-19 Hits Ad Market
14	हल्दी पाउडर में मिलावट के आरोपी को खुद को निर्दोष साबित करने में लग गए 38 साल NEP A Major Step To Enhance Access To Quality Education: Vice President
15	अरविंद केजरीवाल ने कोरोना योद्धा डॉ .जावेद के परिजनों को दिया एक करोड़ की सहायता राशि 4 Arrested For Cheating On False Assurance Of Interest-Free Loans In Delhi: Police
16	भारत और भूटान में घुसपैठ के जरिए चीन देखना चाहता है कि दुनिया उसका विरोध करेगी या नहीं: US4Arrested For Cheating On False Assurance Of Interest-Free Loans In Delhi: Police

text for a lot of data [20]. So this paper load all news articles in a list called corpus then calculate the feature vectors from the documents and finally compute the Euclidean distance and then to check how similar they are. Greater the distance, less similar they are. This paper uses a module or library called sklearn which is a machine learning library.

Result and Analysis

Proposed algorithms are implemented using Python 3.7.3(64-bit). For the experiment, around 1000 news stories were randomly picked from the dataset. The algorithm runs on that dataset, and it measures and compares the various similarity score. Every news article's similarity has been calculated against itself and every other article.

Comparative Analysis

To analyze the performance of the representation method on different similarity measures, the experiment was performed on pairs of news headline obtained from Google News [14]. The chosen news articles are listed in Tables 2, 3 and 4. The news articles were given to a human expert to judge the similarity and dissimilarity. As a result, the human expert has determined 6 pairs (pair 1–6) are completely similar news and 5 pairs (pair 7–11) are different news about the same topic and the other 5 pairs (pair 12–16) are completely dissimilar news. The expert judgment is used as a benchmark to evaluate the automatic similarity calculation on these news articles. The cosine similarity, Jaccard coefficient, and Euclidean distance are applied. The result of all three measures is shown in Tables 5, 6 and 7.

To provide a better understanding of the three compared measures, the results are shown on a bar graph as depicted in Fig. 4.

Figure 5 shows the similarity measures bar graph for different news stories about the same topic.

Figure 6 shows the similarity measures bar graph for completely dissimilar news.

The performance measures used in the experiment are accuracy, precision, recall and F-measures. These measures are calculated by determining the number of news articles correctly identified as similar or dissimilar compared to the decisions by human experts [21]. In other words, using the human decisions as a benchmark the number of true positive (TP) which is equivalent to actual similar news correctly identified as similar, true negative (TN) which is equivalent to actual dissimilar news correctly identified as dissimilar, false positive (FP) which is equivalent to actual similar news incorrectly identified as dissimilar, and false-negative (FN) which is equivalent to actual dissimilar news incorrectly identified as similar are determined. Then, the accuracy is calculated as (TP + TN)/all data, precision is TP/(TP + FP), recall is TP/(TP + FN) and the F-measures as the harmonic mean of precision and recall, which is equal to 2TP/(2TP + FP + FN) [21]. The results are presented in the next section.

Results and Discussion

Figure 7 presents the graph of similarity measurements of the sample pair of news articles using Euclidean, Jaccard and cosine similarity measures for each representation schemes i.e. tf-idf, and a bag of word representation. As can be learned from Fig. 7, Cosine performs similar to benchmark for news with the same meaning (pair 1–6) and different news about the same topic (pair 7–11) and however for completely dissimilar news (pair 12–16) Jaccard's and Euclidean score are similar to the human benchmark.

To prove our point further, we calculated the correlation scores for each similarity measures against the human benchmark as shown in Table 8.

Table 5 Similarity measures of completely same news

S.no.	Cosine similarity	Jaccard similarity	Euclidean similarity
1	0.8931	0.38075	0.04679
2	0.856	0.2134	0.02764
3	0.8476	0.3589	0.04861
4	0.7434	0.3289	0.04610
5	0.8034	0.2086	0.08609
6	0.7899	0.2756	0.02440

Table 6 Similarity measures for different news stories about the same topic

S.no.	Cosine similarity	Jaccard similarity	Euclidean similarity
7	0.7063	0.1168	0.02311
8	0.5301	0.047	0.01377
9	0.5459	0.1516	0.04511
10	0.6316	0.1196	0.03771
11	0.7182	0.132	0.02990

Table 7 Similarity measures of completely dissimilar news

S.no.	Cosine similarity	Jaccard similarity	Euclidean similarity
12	0.3447	0.066	0.0434
13	0.4032	0.0705	0.00804
14	0.4843	0.0996	0.0334
15	0.5490	0.1003	0.0298
16	0.3466	0.08503	0.02949

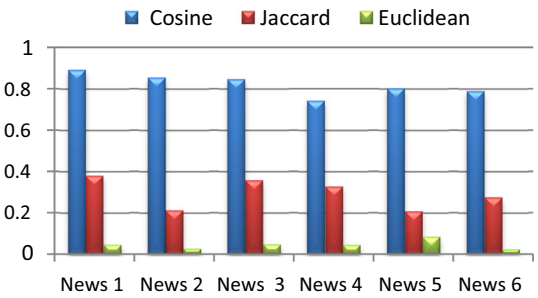

Fig. 4 Comparison of similarity coefficients for articles of same news

From the correlation score in Table 8, it can be perceived that the Cosine and Jaccard similarity is more correlated to the benchmark scores. We further analyze the produced result by calculating the Confusion Matrix [3]

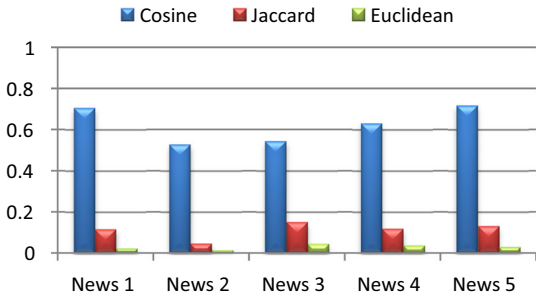

Fig. 5 Comparison of similarity coefficients for different news articles about the same topic

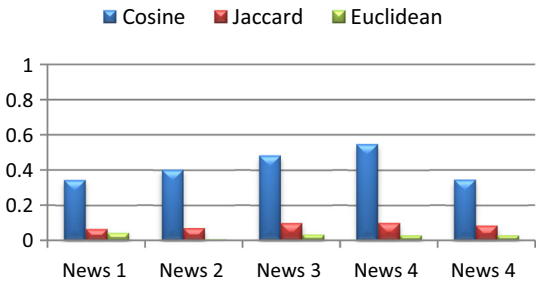

Fig. 6 Comparison of similarity coefficients for completely dissimilar news

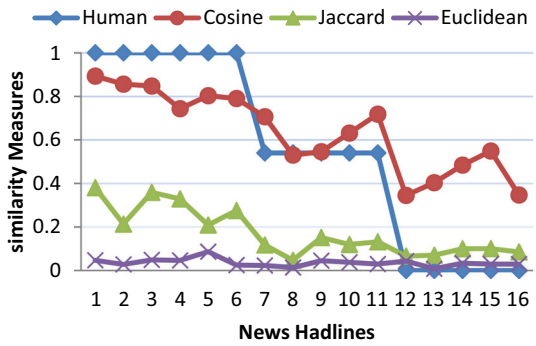

Fig. 7 Similarity score graph

(Tables 9, 10, 11) to find out their accuracy, precision, recall and F- measures as explained in the previous section.

Table 12 gives a clear picture of the performance of each similarity measure. Analyzing the results we see that the Precision value of Jaccard measures is 1.0 or 100% but less than 50% in Euclidean Distance. However, Euclidean gives a high value of Recall as compared to Precision. Cosine measure gives a good accuracy level and $F1$ score, but the difference between Recall value and Precision is high. But, among these three methods cosine similarity using tf-idf showed greater accuracy, recall and F-measure scores of 81.25%, 100% and 76.92%, respectively.

Table 8 Correlation of the similarity scores to the benchmark

Method	Correlation
Cosine and benchmark	0.919847
Jaccard and benchmark	0.816131
Euclidean and benchmark	0.422671

Table 9 Confusion matrix for cosine similarity

16 News	Predicted: No	Predicted: Yes
Total:	8	8
Actual: No	TN = 8	FP = 3
Actual: Yes	FN = 0	TP = 5
Threshold value: 0.788		
Total refined news: 8		

Table 10 Confusion matrix for Jaccard similarity

16 News	Predicted: No	Predicted: Yes
Total:	12	4
Actual: No	TN = 8	FP = 0
Actual: Yes	FN = 4	TP = 4
Threshold value: 0.245		
Total refined news: 4		

Table 11 Confusion matrix for Euclidean similarity

News	Predicted: No	Predicted: Yes
Total:	8	8
Actual: No	TN = 8	FP = 7
Actual: Yes	FN = 0	TP = 1
Threshold value: 0.0529		
Total refined News: 8		

Table 12 Accuracy level of each similarity measures

Similarity measures	Performance measures			
	Accuracy	Precision	Recall	F-Measure
Cosine	**0.8125**	62.5	**1.0**	**0.7692**
Jaccard	0.750	**1.0**	0.50	0.666
Euclidean	0.5625	0.125	**1.0**	0.222

The Highest value is shown in bold

Conclusion

This ongoing research conducted a comparison of three different methods to estimate the semantic similarity among two news articles on (nearly) the same topic/event to measure the similarity between them in two different languages (Hindi and English). The experiment was tested using the GoogleNews data sets. The three methodologies are the similarity of Cosine with tf-idf vectors, similarity of Jaccard with tf-idf vectors, Bag of words Euclidean distance. All three of these methods showed promising results, but among these three methods, cosine similarity using tf-idf showed greater accuracy, recall and *F*-measure scores of 81.25%, 100% and 76.92%, respectively. The accuracy of the other two methods may be improved with the Doc2Vec model [6], which takes text corpus as input and generates document vectors as output. This experiment is also looking to expand the work to other languages.

References

1. G. Atkins, M. Weigle, and M. Nelson, Measuring news similarity across ten U.S. news sites, arXiv preprint arXiv, pp. 1–11, 2018
2. J. Gibson, B. Wellner, and S. Lubar, Identification of duplicate news stories in web pages, in *The MITRE Corporation 202 Burlington Rd. Bedford MA 01730 USA*, 202 Burlington Rd. Bedford MA 01730 USA, 2008
3. M. Singh, D.A. Kumar, D.V. Goyal, Review of techniques for extraction of bilingual lexicon from comparable corpora. Int. J. Eng. Technol. **7**(2), 16–20 (2018)
4. S. Montalvo, R. Martínez, A. Casilla, *Bilingual News Clustering Using Named Entities and Fuzzy Similarity* (Springer, Heidelberg, 2007), pp. 108–114
5. S. Mohd Saad and s. S. Kamarudin, Comparative analysis of similarity measures for sentence level semantic measurement of text. in *IEEE International Conference on Control System, Computing and Engineering,* pp. 90–94, 2013
6. P. Sitikhu, A Comparison of Semantic Similarity Methods for Maximum Human Interpretability, 2019. arXiv:1910.09129v2 [cs.IR]
7. K. Baraniak and M. Sydow, News articles similarity for automatic media bias detection in Polish news portals, in *Proceedings of the Federated Conference on Computer Science and Information Systems*, 2018
8. M. B. Magara and T. Zuva, A comparative analysis of text similarity measures and algorithms in research paper recommender systems, in *Conference on Information Communications Technology and Society (ICTAS)*, 2018
9. V. Thada, D.V. Jaglan, Comparison of Jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. Int. J. Innov. Eng. Technol. (IJIET) **2**(4), 202–204 (2013)
10. M.I. Nasab, A new approach for finding semantic similar scientific articles. J. Adv. Comput. Sci. Technol. (JACST) **4**, 563-59 (2015)
11. M. Snover, B. Dorr, and R. Schwartz, Language and translation model adaptation using comparable corpora, in *Proceedings of*

the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, 2008
12. Q. Long Hua and W. Hong Ling, Bilingual lexicon construction from comparable corpora via dependency mapping, in *Proceedings of COLING*, 2012
13. Y. Y Dani Deahl@danideahl, Google News is getting an overhaul and customized news feeds, THE VERGE, 8 May 2018. [Online]. Available: https://www.theverge.com/2018/5/8/17329074/google-news-update-new-features-newsstand-io-2018
14. H. Hu, "GoogleNews.PyPI," PyPI.org, Mar 13, 2020. [Online]. Available: https://pypi.org/project/GoogleNews/
15. J. Brownlee, How to clean text for machine learning with python, Machine Learning Mastery, October 18, 2017. [Online]. Available: https://machinelearningmastery.com/clean-text-machine-learning-python/
16. S. Polamuri, Five most popular similarity measures implementation-in-python/ https://dataaspirant.com/2015/04/11/five-most-popular-similarity-measures-implementation-in-python/
17. M. Goswami, A. Babu, B. Purkayastha, A comparative analysis of similarity measures to. Int. J. Manag. Technol. Eng. **8**(XI), 786–797 (2018)
18. A. Ali, Textual similarity, ISSN 2011-19, 2011
19. TextBlob: simplified text processing, [Online]. Available: https://textblob.readthedocs.io/en/dev/
20. Bag of words Euclidean distance, [Online]. Available: https://pythonprogramminglanguage.com/bag-of-words-Euclidean-distance/
21. S.S. Kamaruddi, Graph-based representation for sentence similarity measure: a comparative analysis. Int. J. Eng. Technol. **7**(2.4), 32–35 (2018)