Ensemble-based Short Text Similarity: An Easy Approach for Multilingual Datasets using Transformers and WordNet in Real-world Scenarios

Isabella Gagliardi * and Maria Teresa Artese

Posted Date: 12 July 2023

doi: 10.20944/preprints202307.0690.v1

Keywords: semantic textual similarity; pretrained language models; transformers; WordNet; QueryLab; ensemble methods

Article

Ensemble-Based Short Text Similarity: An Easy Approach for Multilingual Datasets Using Transformers and WordNet in Real-World Scenarios

Isabella Gagliardi * and Maria Teresa Artese

IMATI – CNR; gagliardi , teresa@mi.imati.cnr.it

* Correspondence: isabella@mi.imati.cnr.it or gagliardi@imati.mi.cnr.it; Tel.: +39-02-23699487

Abstract: When integrating data from different sources, there are problems of synonymy, different languages, concepts of different granularity. This paper proposes a simple but effective approach to evaluate the semantic similarity of short texts, especially keywords. The method is capable of matching keywords from different sources and languages by exploiting transformers and WordNet-based methods. Key features of the approach include its unsupervised pipeline, mitigation of the lack of context in keywords, scalability for large archives, support for multiple languages and real-world scenarios adaptation capabilities. The work aims to provide a versatile tool for different cultural heritage archives without requiring complex customization. The objectives of the paper are to explore different approaches to identifying similarities in 1- or n-gram tags, to evaluate and compare different pre-trained language models, and to define integrated methods to overcome limitations. Tests to validate the approach have been conducted using the QueryLab portal, a search engine for cultural heritage archives, to evaluate the proposed pipeline.

Keywords: semantic textual similarity; pretrained language models; transformers; WordNet; QueryLab; ensemble methods

1. Introduction

The wide availability of information on the Web makes it necessary to define and develop tools that allow this information to be integrated, not only as a display of results, but also when searching and making suggestions to the user.

Keywords have always been very useful in representing the content of texts, but at the same time they require skill and ability in identifying those that are general enough to be associated with multiple texts and specific enough to identify homogeneous subsets of data. The use of different keywords, either automatically extracted from the text or associated by hand by experts in the field or catalogers is a limitation that prevents retrieval engines from obtaining good results. It can then happen that texts and keywords are in different languages, either within the same text or the same dataset.

In this paper, a simple and easy-to-use yet effective approach to evaluate the semantic similarity of short texts, represented by keywords, is presented: given two sets of keywords, from different sources and even in different languages, the method is able to perform the best matches between the lists. We present a prototype that exploits pre-trained language models to assess the semantic similarity of context-free keywords, integrating them with dictionary-based methods and string-based similarity. Different models and similarity measures are tested both individually and ensemble, to create an unsupervised, fully automatic approach that can be applied to different lists of terms (1-grams and n), even in different languages.

The effectiveness and ease of use of this approach are attributed to the following key features:

- Unsupervised pipeline: Once hyperparameters and transformation models are optimized, the entire process becomes unsupervised.

- Mitigation of the lack-of-context: the integration of different methods (neural network transformers, dictionary based and syntactic) allows the mitigation of the lack of context of keywords and the definition of final similarity scores.
- Scalability: Using pre-trained language models, the approach can efficiently handle even archives with hundreds of elements.
- Multilingual support: The use of pre-trained multilingual text models enables the approach to efficiently manage archives containing documents in different languages, such as English, French, Italian, and others.
- Real-world scenarios: The experiments performed for this article demonstrate the ability of the method to adapt to real data without having to adapt it to a specific context. The use of ensemble methods makes it possible to overcome any critical problems that may arise due to unfamiliar words or different languages.

The proposed approach incorporates state-of-the-art tools, including transformers and pretrained language models, corpus-based (in this case Word-net) and string-based similarity techniques. By synergistically integrating these tools, a versatile and adaptable tool has been created, that provides a method suitable in a variety of contexts.

The paper aims to define a simple yet effective method that is easily applicable to different cultural heritage archives and provides satisfactory results, without requiring sophisticated methods to adapt to different cases. To achieve this, we set ourselves the following objectives:

1) Provide an overview of different approaches that can be adapted to identify overall similarities of 1- or n-gram tags.
2) Evaluate and compare the performance of different pre-trained language models on short texts/keywords, which by their nature are self-contained.
3) Define methods that can overcome limitations through integrated approaches, after analyzing the results of individual methods.

To evaluate the pipeline, we conducted tests using data from the QueryLab portal, a search engine for archives on tangible and intangible cultural heritage, capable of querying different datasets, both local and via web services simultaneously.

The paper is organized as follows: after a brief analysis of the state of the art, the approach is presented highlighting the technical and innovative features, both individual measures and of the overall approach. Experiments on QueryLab are described, commented on, and discussed, comparing the results of the individual approaches and the overall results. The results obtained by applying the method to the gold standard WordSim353 are also presented. A conclusion and future developments conclude the article.

2. Related works

The study of word similarity is a fundamental task in natural language processing and has gained substantial attention from researchers. Numerous approaches have been proposed to measure the semantic similarity or relatedness between words. In this section, we introduce related works in the field of word similarity.

There is a great number of survey work available in the literature regarding similarity between words and phrases. In the comprehensive survey by Atoum et al. [1], the authors categorize similarity methods into two groups: those for word similarity and those for phrase similarity. Within each group, methods are further classified into three types: corpus-based, knowledge-based, and hybrid methods. Corpus-based approaches utilize statistical information from large text corpora, and they can be further divided into subcategories. Knowledge-based methods, on the other hand, rely on dictionaries or other structured resources to derive semantic knowledge. Hybrid methods combine elements from both corpus-based and knowledge-based approaches to leverage their respective strengths. For sentence-level similarity, the presence of context allows the incorporation of typical information retrieval (IR) features such as tf-idf or the utilization of large corpora like Wikipedia.

Other authors, such as Gomaa et al. [2], Gupta et al [3] or Sunilkumar et al. [4], in their surveys, categorize the similarity methods in analogous manners.

In [5], the authors conduct a systematic review of research on similarity measurement, analyzing the advantages and disadvantages of different methods. They categorize similarity measures into two major groups: those based on distance metrics and those based on text representation.

WordNet, known for its capability to represent semantic relationships, is widely used by researchers to measure semantic similarity. In [6], the authors provide a comprehensive review of various WordNet-based measures. They discuss the strengths and weaknesses of each approach in capturing semantic similarity. The use of WordNet in combination with a corpus is also explored in [7], where a hybrid method is proposed using a novel similarity measure. This approach combines structural information from WordNet with statistical information from a corpus to enhance semantic text similarity.

Furthermore, the evaluation of semantic relatedness between lists of nouns using WordNet is investigated in [8]. The authors conduct experiments to evaluate the ability of different semantic relatedness measures, including latent semantic analysis (LSA), GloVe, FastText, and various WordNet-based measures, to predict differences in word recall between two lists of words.

Word similarity using word embeddings, such as Word2Vec or GloVe, has been a popular approach in the field of natural language processing. These methods generate dense vector representations for words based on their co-occurrence patterns in large text corpora. Word similarity can be measured by calculating cosine similarity or other distance metrics between the respective word representations. Word2Vec and GloVe have shown promising results in capturing semantic relationships and syntactic regularities between words. in [9] the authors study whether similarity between short texts can be assessed using only semantic features. Vector representations of words, computed from unsampled data, are used to represent terms in a semantic space in which the closeness of the vectors can be interpreted as semantic similarity.

In recent years there has been a shift toward the use of transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), for word similarity and related tasks [10]. The preference for BERT- and transformer-based models in word similarity tasks stems from their ability to capture contextualized representations, take advantage of bidirectional language modeling, use large-scale pre-training, offer fine-tuning capabilities, and harness the power of transformer architecture. These advances have shown significant improvements in capturing word semantics and addressing the challenges posed by word polysemy and ambiguity.

In [11] the authors trace the evolution of semantic similarity methods from traditional NLP techniques, such as kernel-based methods, to the most recent research work on transformer-based models, classifying them according to their basic principles as knowledge-based, corpus-based, deep neural network-based and hybrid methods. Another survey using deep learning is presented in [12].

For the choice of which pretrained language model to use, comparisons made in specific fields, such as medical or financial [13,14], can be found in the literature. There are also works comparing these general models [15,16]. Their shortcoming is that these reviews and comparisons age very quickly. On the HuggingFace site [1](from which the models used in this experimentation came) experimentation, for example, models are constantly and continuously added and updated.

To the best of our knowledge, ours is the first approach to develop a method that relies on three different methods and their integration, entirely unsupervised in the "wild."

3. Materials and Methods

The purpose of this paper is to define a method to identify the most similar or related tags within two lists. As we will see later in the experimental part, the system is able to identify one or more terms that it judges to be similar, regardless of how similar they may appear at first glance. Actually, the system generates a sorted list of all compared terms based on similarity for each individual

[1] Hugging face model for sentence similarity:
https://huggingface.co/models?pipeline_tag=sentence-similarity

 doi:10.20944/preprints202307.0690.v1

method and for ensemble methods. Figure 1 shows the pipeline of the approach, the purpose of which was found in the real case of integrating archives from different sources in QueryLab [17]. The pipeline involves a series of steps or components designed to retrieve and present relevant terms based on their semantic similarity or association with the given word. The main steps of the pipeline for finding related terms:

1) Language identification and Preprocessing: Identify the language and clean and normalize the data by removing punctuation, converting to lowercase, and handling any specific linguistic considerations (e.g., stemming or lemmatization) to ensure consistency in word representations.
2) Single methods similarity:
   a) identify the different similarity/relatedness techniques to be used to evaluate the similarity between terms. Examples can be: word embeddings, transformers using pretrained Bert-like models, dictionary based and syntactic based, and so on.
   b) a similarity matrix is computed for each element in the lists, for each similarity/relatedness method: the similarity scores between the target word and all other words in the dataset using cosine similarity or other similarity measures are computed. The similarity score quantifies the relatedness between two words based on their vector representations or other representation, according to the technique.
   c) Ranking and Selection: Rank the words based on their similarity scores in descending order. You can define a threshold or select the top N words as the most related terms, depending on the desired number of results.
3) Ensemble approach: combine the results of multiple similarity measures to derive final similarity scores.

3.1. Materials

One of the goals of the method is to define a tool that is capable of processing information in natural language in an unsupervised manner. The testing of the method was carried out on QueryLab, a search engine for tangible and intangible cultural assets. Several types of information from QueryLab were used to evaluate the effectiveness of the method, either lists prepared by experts or keywords automatically extracted from QueryLab archives, as described in more details below.

3.2. Language identification and preprocessing

Language identification and preprocessing are important steps in natural language processing (NLP) tasks. Language identification involves determining the language of a given text or document, while preprocessing focuses on preparing the text data for subsequent analysis. Preprocessing is a crucial phase when working on data that involves applying a series of operations and transformations to prepare the data optimally for subsequent analysis. Preprocessing aims to improve data quality, reduce noise, eliminate irrelevant information, and make the data more suitable for machine learning algorithms or other analysis techniques.

3.3. Single methods similarity

Semantic relatedness with transformers: Transformers, as a remarkable development in the field of neural networks, have revolutionized the domain of natural language processing [18]. In contrast to conventional approaches that heavily depend on manually engineered features and statistical models, transformers employ a distinct mechanism known as self-attention [19]. This mechanism enables the model to dynamically allocate attention to various parts of the input, facilitating the capture of long-term dependencies within the language data.

Among the diverse variety of pre-trained language models, BERT (Bidirectional Encoder Representations from Transformers) stands out as one of the most influential and widely adopted models [10]. BERT, being one of the earliest pre-trained models, has significantly impacted the field by providing a powerful representation learning framework. By training on large-scale corpora,

 doi:10.20944/preprints202307.0690.v1

BERT is capable of learning contextualized word embeddings that encapsulate the semantic information of words based on their surrounding context. The availability of BERT has significantly advanced various downstream natural language processing tasks, such as sentiment analysis, named entity recognition, and machine translation. These tasks have experienced notable improvements in terms of accuracy and performance. Similarly, other models like GPT, RoBERTa, and Mini-L6, also based on the transformer architecture, utilize similar techniques as BERT to capture contextualized word representations. Using these techniques, these models enable a wide range of natural language processing applications. Each model offers unique enhancements or modifications tailored to address specific challenges and requirements in NLP tasks. As a result, they contribute to the continuous development and progress of natural language processing, expanding its capabilities and potential in various fields.

These models are built upon a transformer-based neural network architecture. They undergo training using vast amounts of unlabeled text data, enabling them to develop a deep understanding of natural language patterns and structures. BERT and its counterparts developed by Microsoft, Facebook, OpenAI, or HuggingFace are characterized by bidirectional language modeling. This means that they can comprehensively analyze and comprehend the context of a word by considering both the preceding and succeeding words in a sentence. This bidirectional approach allows for a more nuanced understanding of word relationships and linguistic subtleties. In the paper, we conduct experiments where we evaluate various pre-trained models.

The process starts with pretrained models that represent individual words as vectors. These vectors encapsulate the semantic meaning of the respective words. However, when dealing with n-grams or phrases, additional techniques are required to combine the vectors associated with each component and generate a single vector that represents the entire n-gram or phrase. This paper explores three distinct methods employed for this purpose, which are elaborated upon below. In the experimental section, we will present and analyze the results obtained through the utilization of these methods.

One approach is to use the embedding of the [CLS] token, which is added to the beginning of each sentence in a batch during the pre-processing stage. The [CLS] token is designed to represent the entire sentence, and its embedding captures the meaning of the sentence. One way to compute similarity between two sentences (in this case n-grams) using pretrained language models is to take the dot product of their [CLS] token embeddings, which will give a score between -1 and 1, where 1 means the sentences are identical and -1 means they are completely dissimilar. Another way to compute similarity is by using the cosine similarity between the [CLS] token embeddings.

Another approach to compute sentence similarity using BERT-like language models is to take the average of the token embeddings for each sentence. This is known as the mean-pooling approach and from now on, this method will be referred to as [AVG]. To compute sentence similarity using mean-pooling approach, first, we run pretrained models on the input sentence and get the hidden states for each token. Then, we take the average of the hidden states for each sentence. Finally, we can compute the similarity between two sentences by taking the dot product or cosine similarity of their average token embeddings. This approach can be useful in certain use-cases where we want to find the overall similarity between two sentences rather than just comparing the [CLS] token embeddings. However, it should be noted that this approach might not work as well as the [CLS] token embeddings approach in all the cases, as it may not capture the entire meaning of the sentence. In our case, this is not a problem as we are dealing with very short phrases, typically consisting of 3 or 4 words at most.

Another approach to compute sentence similarity using BERT is to take the maximum of the token embeddings for each sentence, also known as the max-over-time pooling approach and referred to as [MAX].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

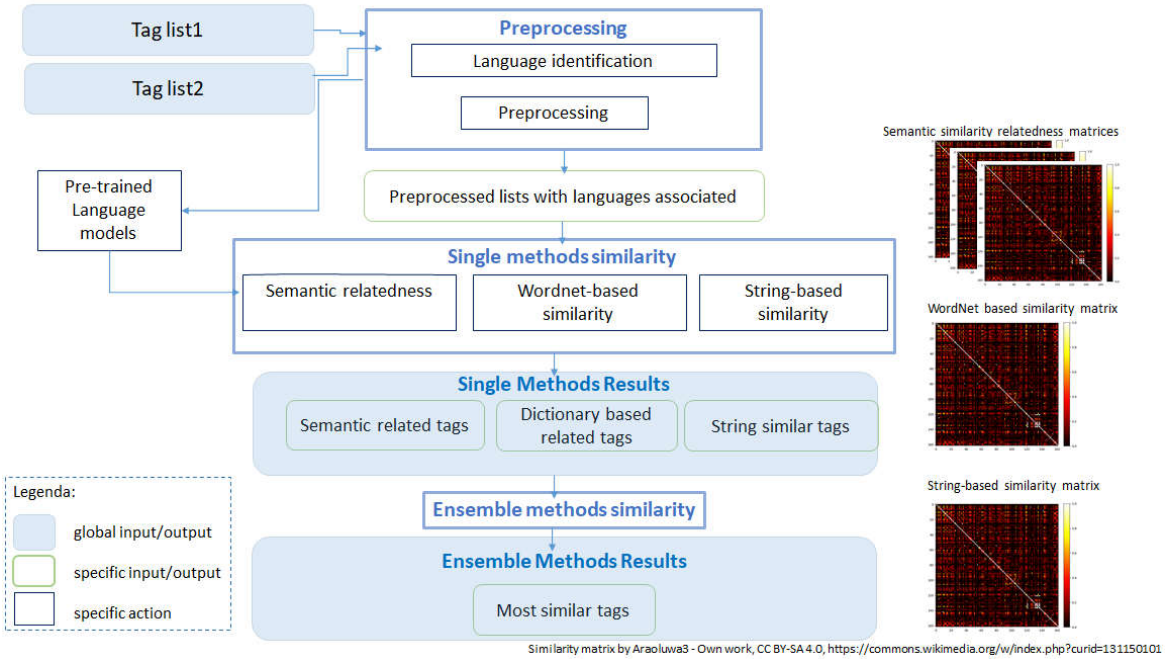

Figure 1. Schema of the proposed method.

To compute sentence similarity using max-over-time pooling approach, first, we run the model on the input sentence and get the hidden states for each token, as for [AVG] method. Then, we take the maximum of the hidden states for each sentence by taking the maximum value over the last dimension. Finally, we can compute the similarity between two sentences by taking the dot product or cosine similarity of their maximum token embeddings. This approach can be useful in certain use-cases where we want to find the dominant meaning or feature of the sentence. Also in this case, it should be noted that the max-over-time pooling approach may not capture the entire meaning of the sentence like the mean-pooling approach, and it may also be sensitive to outliers.

WordNet -based similarity: WordNet is a lexical database and semantic network that organizes words and their meanings into a hierarchical structure [20,21]. It provides a comprehensive and structured resource for understanding the relationships between words, synonyms, antonyms, and the hierarchical structure of concepts. In WordNet, words are grouped into synsets (synonym sets), which represent a set of words that are closely related in meaning. Each synset represents a distinct concept or meaning. Synsets are connected through semantic relations, such as hyponyms (subordinate concepts), hypernyms (superordinate concepts), meronyms (part-whole relationships), and holonyms (whole-part relationships).

WordNet's primary purpose is to facilitate the exploration of semantic relationships between words and to measure their semantic similarity. Again, three measures were used to assess the similarity between n-grams:

1. The shortest path length measure computes the length of the shortest path between two synsets in the WordNet graph, representing the minimum number of hypernym links required to connect the synsets. This measure assigns a higher similarity score to word pairs with a shorter path length, indicating a closer semantic relationship. It will be referred to as *path*.
2. Wu-Palmer Similarity: The Wu-Palmer similarity measure utilizes the depth of the LCS (Lowest Common Subsumer - the most specific common ancestor of two synsets in WordNet's hierarchy) and the shortest path length to assess the relatedness between synsets. By considering the depth of the LCS in relation to the depths of the synsets being compared, this measure aims to capture the conceptual similarity based on the position of the common ancestor in the WordNet hierarchy. It will be referred to as *wu*.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

3. Measure based on distance: analogously to shortest path length, this measure is also based on the minimum distance between 2 synsets. This measure, hand-crafted by the authors, takes into consideration that the shorter the distance, the greater the similarity. In this case, the similarity measure is calculated using this equation:

$$\text{min\_dist}=\frac{1}{(0.8+\frac{distance}{4})} \quad (1)$$

When distance >0, in the other case min_dist=1

This measure considers only the distance between synsets, in a depth-independent way, and was obtained by doing several tests and evaluations, so that much weight is given not only to synonyms (with distance 0), but also to hypernyms or hyponyms (distance 1) or siblings (distance 2). This measure will be referred to as *min_dist.*

String comparison algorithms: Jaro, Jaro-Winkler, Levenshtein and other similar similarity measures are string comparison algorithms that focus on quantifying the similarity (or distance) between two strings based on their characters and their order. These measures are useful in various applications, including record linkage, data deduplication, and fuzzy string matching. Here it has been used Jaro-Winkler Similarity, that is an extension of the Jaro similarity measure. It incorporates a prefix scale that rewards strings for having a common prefix. The Jaro-Winkler similarity score ranges from 0 to 1, with 1 indicating a high similarity and a closer alignment of the prefixes.

3.4. Ensemble methods

Since in this paper we have identified, defined, and implemented several measures to calculate similarity between objects, it is necessary to identify appropriate voting mechanisms to determine the best results. The starting point is to consider the following variables:

1. n pretrained language models: as we will see in the experimentation part, the datasets on which we tested our approach can be monolingual (English or Italian, so far), or multilingual (English, Italian or French, in the experiments). It is therefore necessary to identify the models that are best able to represent the specificity of the data under consideration.
2. semantic relatedness: three different methods of calculating the representative vector to be evaluated must be compared.
3. WordNet based similarity: again 3 different ways of calculating the similarity between words.

Ensemble methods [22,23] have the aim of combining multiple individual models or methods to produce a more accurate and robust prediction or decision. It is based on the principle that the collective wisdom of diverse models tends to outperform any individual model in terms of accuracy, generalization, and stability. There are different voting schemes such as majority voting, where the predicted class with the highest number of votes is selected, and weighted voting, where the models' predictions are weighted based on their performance or expertise. We used both methods in different phases, to identify the more accurate results.

3.5. Evaluation of the results

We are interested in evaluating the effectiveness and performance of each method when assessing the similarity between objects. By examining the outcomes, we can gain a deeper understanding of how well these approaches capture and quantify the similarity between different objects, thereby facilitating informed discussions and conclusions.

Borrowing ideas from information retrieval systems [24], a variety of metrics can be used to evaluate semantic similarity measures, including recall, precision, f1 scores, Dice and Jaccard coefficients:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

1) Recall: Recall measures the proportion of relevant items that are correctly identified or retrieved by a model. It focuses on the ability to find all positive instances and is calculated as the ratio of true positives to the sum of true positives and false negatives.
2) Precision: Precision measures the proportion of retrieved items that are actually relevant or correct. It focuses on the accuracy of the retrieved items and is calculated as the ratio of true positives to the sum of true positives and false positives.
3) F1 score: The F1 score combines precision and recall into a single metric. It is the harmonic mean of precision and recall and provides a balanced measure of a model's performance. The F1 score ranges from 0 to 1, with 1 being the best performance. It is calculated as 2 * (precision * recall) / (precision + recall).
4) Dice coefficient: The Dice coefficient is a metric commonly used for measuring the similarity between two sets. In the context of natural language processing, it is often employed for evaluating the similarity between predicted and reference sets, such as in entity extraction or document clustering. The Dice coefficient ranges from 0 to 1, with 1 indicating a perfect match. It is calculated as 2 * (intersection of sets) / (sum of set sizes).
5) Jaccard coefficient: The Jaccard coefficient, also known as the Jaccard similarity index, measures the similarity between two sets. It is commonly used for tasks like clustering, document similarity, or measuring the overlap between predicted and reference sets. The Jaccard coefficient ranges from 0 to 1, with 1 indicating a complete match. It is calculated as the ratio of the intersection of sets to the union of sets.

These metrics provide different perspectives on the performance of semantic similarity measures. Recall focuses on the measure's ability to identify relevant similar pairs, precision emphasizes the accuracy of the identified pairs, and Jaccard similarity provides an overall measure of overlap between identified and reference pairs.

4. The Experimentation

The approach presented in this study comprises a series of sequential steps, accompanied by detailed explanations and specific examples to provide a comprehensive understanding of the implementation process. Table 1 presents the pipeline of the approach.

Table 1. pipeline of the presented approach.

Task 1: Dataset Preparation
- Harvesting process
- Language identification
- Preprocessing (possibly strip stopwords, accents, …)
- Output: items of interest
Task 2: Single Method similarity/relatedness
- # **semantic relatedness**
- Choice of transformers and pre-trained models
- Fine tuning of pre-trained Bert-like models to obtain the vectors
- Computation of similarity matrix using [CLS] tokens, means and max pooling
- Output: for each model, 3 lists of the most related tags, ordered by score value
- # **wordnet-based**
- For each tag1 in list1 and tag2 in list2:
- Synsets identification for the tag1 and tag2, using language and automatic translation if necessary
- Computation of the max path_similarity and wu_p_similarity
- Computation of the min distance and hand crafted similarity

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

- Output: 3 lists of the most related tags, ordered by score value
- #**string-based similarity**
- Computation of jaro-wrinkler similarity among string
- Output: list of the most related tags, ordered by score value

Task 3: Ensemble similarity

- #**ensemble for each pretrained model**
- Weighted Voting mechanism to create, for each language model, a unique list of most related tags
- Output: list of the most related tags, ordered by score value
- #**global ensemble**
- Majority Voting mechanism
- Output: lists of the most related tags, ordered by number of votes

Task 4: Evaluation

- Ground truth creation
- Computation of recall, precision, and Jaccard similarity

4.1. Datasets

QueryLab Platform

The methods presented here has been designed with the idea of their use in the "wild", integrating them in QueryLab, a prototype dedicated to the integration, navigation, searching and preservation of tangible and intangible heritage archives on the web, with the aid of themed paths, keywords, semantic query expansion and word cloud [17].

QueryLab foresees different ways of querying inventories by collecting and managing data in a transparent way for the user. Archives are integrated in QueryLab, in the most automatic way possible, both via web services and local ones. The datasets are constantly growing and currently are those presented in Table 2.

Due to the diverse nature of the archives comprising QueryLab, the results obtained for the same query can exhibit significant variations. This discrepancy arises from keyword lists that may differ in form or language while still maintaining semantic similarity. Consequently, there is a need to devise a method for harmonizing keywords, serving a twofold purpose:

- During the search phase, expanding the query to include all tags that surpass a predetermined similarity threshold. This expansion allows for a broader search scope, encompassing tags that are semantically similar to the original query. By including such tags, we aim to enhance the search results by considering related keywords.
- During the fruition stage, suggesting elements that contain similar keywords to enhance the user experience. By identifying and recommending elements that share similar keywords, we provide users with relevant and related content. This approach enables users to explore and access information beyond their initial query, promoting a comprehensive and enriched browsing experience.

Table 2. Archives in the QueryLab portal.

Name (source)	Country	Data	Language	Description
Intangible Search (local)	Italy	ICH	Italian, English, French, German	Arts and Entertainment, Oral Traditions, Rituals, Naturalistic knowledge, Technical knowledge
ACCU Databank (local)	Pacific Area	ICH	English	Performing Arts

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

Sahapedia (local)	India	ICH	English (used here) and other local idioms	Knowledge Traditions, Visual and Material Arts, Performing Arts, Literature and Languages, Practices and Rituals, Histories…
Lombardy Digital Archive (local)	Italy	ICH, CH	Italian	Oral history, Literature, Performing Arts, Popular traditions, Work tools, Religion, Musical instruments, Family…
Museum Contemporary Photography (local)	Italy	CH	Italian	Architecture, Cities, Emigration, Family, Objects, Work, Landscapes, Sport…
Europeana [6] (web)	European countries	ICH, CH	multilingual	1914-18 World War, Archaeology, Art, Fashion, Industrial Heritage, Manuscripts, Maps, Migration, Music, Natural History, Photography…
Victoria&Albert Museum (web)	UK	ICH, CH	English	Architecture, Embroidery, Costume, Paintings, Photography, Frames, Wallpaper, Jewelry, Illustration, Fashion, Manuscripts, Wedding dress, Books, Opera, Performance…
Cooper Hewitt (web)	USA	CH	English	Decorative Arts, Designs, Drawings, Posters, Models, Objects, Fashion, Textiles, Wallcoverings, Prints…
Réunion des Musée Nat – RMN (web)	France	ICH, CH	French with tags in English	Photography, Paintings, Sculptures, Literature, Dances, Modern Art, Decorative Arts, Music, Drawings, Architecture…
Auckland Museum (web)	New Zealand	ICH, CH	English	Natural Sciences, Human History, Archaeology, War history, Arts and Design, Manuscripts, Maps, Books, Newspapers…
Digital Public Library of America – DPLA (web)	USA	ICH, CH	English	American civil war, Aviation, Baseball, Civil rights movement, Immigration, Photography, Food, Women in science…
Harvard Art Museums – HAM (web)	USA	ICH, CH	English	art history, conservation, and conservation science; learning and supporting activities
CookIT	Italy	ICH	Italian	Traditional Italian Recipes

Gold standard datasets

In addition to the data collected, the evaluation of the quality of the approach is based on two gold standards for word similarity: WordSim353 [25] and SimLex999[26][2].

WordSim353 is a widely used benchmark dataset for evaluating word similarity and relatedness. It consists of 353-word pairs, and each pair is assigned a similarity score by human annotators based

[2] WordSim353 and SimLex999 datasets:
https://github.com/kliegr/word_similarity_relatedness_datasets

 doi:10.20944/preprints202307.0690.v1

on their perceived similarity or relatedness. The dataset covers a range of semantic relationships, including synonyms, antonyms, hypernyms, and more. The WordSim353 dataset serves as a standard evaluation resource for measuring the performance of word embedding models and other techniques in capturing semantic similarities between words. Researchers often use this dataset to assess the quality of their models and compare their results with other approaches.

SimLex-999 is another widely used benchmark dataset for evaluating word similarity and relatedness. It consists of 666 word pairs, and each pair is assigned a similarity score by human annotators based on their perceived similarity. The dataset covers a diverse range of semantic relationships and includes words with different levels of similarity and relatedness.

4.2. Dataset preparation

Language identification and preprocessing

In many of the processing and similarity evaluation tasks, language plays a crucial role. While certain datasets used in our experiments have known languages, such as benchmark datasets or Italian recipes, the data collected through automated methods often lacks fine-grained language information. To address this, we perform preprocessing to extract detailed language information from the data [27].

It is important to note that the preprocessing step is only applied to data collected in QueryLab archives. Since we are already working with keywords, the main activity performed is grouping tags within each dataset based on specific criteria. Specifically, we group tags that have a distance of 0 in WordNet and a Jaro similarity score greater than 0.95.

4.3. Single Methods similarity

The assessment of similarity in this experiment placed emphasis on several key aspects:

1. Determining the best similarity score: Various similarity assessment methods were utilized to calculate similarity scores between tags. The evaluation aimed to identify the similarity score that best captured the semantic similarity between the given tag and other tags in the datasets.
2. Identifying the single tag with the highest similarity: Based on the calculated similarity scores, the tags that exhibited the highest similarity to the given tag were identified, one for each similarity measure.
3. Identifying the three most similar features: In addition to finding the single tag with the highest similarity, the evaluation process also focused on identifying the three most similar features. These features are the elements (1-gram or n-grams) that demonstrated high similarity to the given tag.

By considering these three aspects, the similarity process aimed to provide a comprehensive assessment of the semantic similarity between tags.

Semantic relatedness

For evaluating semantic similarity, we first convert tags, which can be 1-gram or n-grams, into vectors using pretrained language models specifically designed for "sentence similarity" tasks. The availability of pretrained models is extensive and continuously expanding.

One of the aims of the paper is to compare how different pretrained models perform in the evaluation of semantic similarity: we tested 2 or 3 different language models in the various languages. We select these models based on information gathered from literature sources or platforms such as Hugging Face, as well as models developed by major tech companies like Microsoft, Facebook, or Google.

We utilize three distinct sets of pretrained models:

Models designed for the English language: we tested three pretrained models, i.e.:

- 'sentence-transformers/paraphrase-MiniLM-L6-v2'
- 'flax-sentence-embeddings/all_datasets_v3_mpnet-base'
- 'tgsc/sentence-transformers_paraphrase-multilingual-mpnet-base-v2'

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

Models tailored for the Italian language.

- `'nickprock/sentence-bert-base-italian-xxl-uncased'`
- `'tgsc/sentence-transformers_paraphrase-multilingual-mpnet-base-v2'`

Multilingual models capable of handling multiple languages.

- `'LLukas22/paraphrase-multilingual-mpnet-base-v2-embedding-all'`
- `'tgsc/sentence-transformers_paraphrase-multilingual-mpnet-base-v2'`

The initial step in evaluating semantic similarity involves generating similarity matrices. The similarity matrix requires a single vector for each element being compared. When utilizing a BERT-like model, we employ specific tokens, such as [CLS], which represents the entire sentence, or the average of tokens [AVG], or the [MAX] as mentioned previously. The input text undergoes preprocessing, involving tokenization and encoding into numerical vectors. These vectors are then fed into the transformation model, which produces contextualized embeddings for each token in the input text.

To construct the similarity matrix, the [CLS], [AVG] or [MAX] tokens are compared pairwise using a distance metric, typically cosine similarity. The resulting scores reflect the similarity relationships between the [CLS], [AVG] or [MAX] tokens within the input text, allowing for the creation of the similarity matrix**.**

WordNet based similarity

The approach for WordNet-based similarity is based on the principle that words with greater similarity exhibit smaller distances within the WordNet structure. However, the search for tag synsets in WordNet is language-dependent. We began with two key considerations:

1. Not all WordNet synsets have been translated into all languages. This implies that some synsets may not be available in certain languages, which can affect the matching process.
2. By employing machine translation on tags, we can increase the likelihood of finding a corresponding synset. This strategy helps overcome language barriers and enhances the chances of locating relevant synsets.

For tags comprising multiple words, if the lemma (base form) of a word is not found directly, we split and identify the lemmas and synsets of individual words. This allows for a more granular analysis and matching process.

Similar to the previous approach, a similarity matrix is generated using the three measures described above. This involves computing similarity scores using cosine similarity between the similarity scores from the WordNet-based analysis. These scores serve as the basis for constructing the similarity matrix, facilitating further analysis and evaluation.

String-based similarity

String-based similarity techniques are valuable for detecting minor script changes, small errors, or variations such as singulars and plurals, which can hinder effective searches across archives or data from different sources. By utilizing string-based similarity, we can address these issues and enhance the search process.

4.4. Ensemble Methods similarity/relatedness

Various similarity assessment methods using different characteristics were employed in this experiment, making it necessary to take an overall approach that provides a comprehensive and balanced assessment of similarity, taking into account multiple measures and their respective strengths in the overall assessment process. The ensemble evaluation approach followed the path of individual evaluation and the best three results. This involved consideration of the best similarity score as a weighted voting mechanism, both overall for all methods and based on semantic or WordNet-based similarity.

For each pre-trained model, this ensemble approach considered the best similarity score by combining the results of the various similarity evaluation techniques. Using this weighted voting

 doi:10.20944/preprints202307.0690.v1

mechanism, the ensemble evaluation aimed to capture the overall similarity between the tags, while also taking into account the specific evaluation techniques employed.

Moreover, an ensemble solution was performed on the results at a higher granularity by pooling the results from all the language models considered. In this case, a majority voting approach was employed to capture the collective decision of multiple language models.

We defined ensemble methods by considering a combination of methods, language models and frequency of the most similar terms:

1. Single Model Ensemble: The combined results for a single model are referred to as enx.modelname, considering either the most similar term (n=1) or the first three most similar terms (n=3). The 'x' in the notation can be empty (indicating no specific method), 's' for semantic methods, or 'w' for WordNet-based methods. The 'modelname' refers to the individual pretrained models used for the specific experiment.
2. All Models Ensemble: The combined results (enx.all) considering all models, with the majority voting mechanism applied to all terms. The 'n' value can be 1 or 3, while 'x' can be empty, 's' for semantic methods, or 'w' for WordNet-based methods.
3. Most Frequent Terms Ensemble: The combined results (enx.max) considering all models, with the majority voting mechanism applied only to the most frequent terms. The 'n' value can be 1 or 3, and 'x' can be empty, 's' for semantic methods, or 'w' for WordNet-based methods.

4.5. Evaluation

User ground truth

To evaluate the quality of the approach, it is crucial to have a "ground truth" dataset for comparing the results and calculating metrics such as recall and precision. For this purpose, a procedure was established involving annotators, including both experts in the Cultural Heritage field and ordinary individuals.

In the annotation process, the annotators were tasked with identifying the word or words most similar to a given target word. This task allowed for the identification of terms that closely matched the target word's meaning. However, in cases where finding suitable terms proved challenging due to a lack of similar options, the annotators were provided with the following choices:

- Eliminate the word: If no suitable term could be found, the annotators had the option to exclude the word from the annotation process due to the unavailability of appropriate alternatives.
- Make an extremely bland association: In cases where the annotators faced significant difficulty in finding similar terms, they had the option to make a less precise or less contextually relevant association. This option aimed to capture any resemblance or connection, even if it was not an ideal match.

In the instructions for annotators, the first choice is preferred. To assist the annotators in this task, the system provided suggestions that exhibited very high Jaro similarity values or had a WordNet distance of less than 2. These suggestions were intended to guide the annotators and help them identify potentially relevant terms.

In the annotation process, the annotators were given the task of identifying not only the most similar tag but also the three most related or similar tags in some way to the given target tag. This approach aimed to capture a broader range of related terms and expand the understanding of semantic relationships. By requesting the annotators to identify the three most similar or related tags, the evaluation process went beyond a single best match and encompassed a small set of related tags that shared some semantic similarities or connections with the target tag.

5. Results and discussion

5.1. Datasets used

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

The data collection process was based on a series of queries to the QueryLab archives, using the REST API protocol at the source and storing the following data for the first n results. Each archive organizes the search results differently and therefore requires ad hoc procedures.

List similarity experiments were performed on the following data:

1. keywords extracted dynamically in QueryLab. An interesting thing is that QueryLab's archives are multilingual and almost all of them are able to respond with English terms. when tested, however, different languages coexist in e.g. RNM or DPLA. Here are the results of the query:
   1) *mariage* on Europeana and Rnm
   2) *wedding* on Victoria &Albert Museum and DPLA.

An initial analysis of the data, shown visually in Figures 2 and 3, shows that there are indeed clearly identifiable synonyms and that the language of the words is different. These two datasets were chosen because they possess certain characteristics that make them interesting. The *Mariage* results consist of lists of terms in both English and French; whereas the *Wedding* results contain only English tags, using synonyms. Therefore, the chosen pre-trained models are multilingual models for *Mariage* to handle terms in both English and French, whereas English-specific pre-trained models were used for *Wedding*.

2. Intangible heritage-related tag lists extracted from QueryLab. These lists are hand-built by experienced ethnographers. The tags we compare are those defined for the Archives of Ethnographies and Social History AESS and its transnational version for the inventory of the intangible cultural heritage of some regions in Northern Italy, some cantons in Switzerland and some traditions in Germany, France and Austria [28] and those of UNESCO [29], imported in QueryLab. The language is the same (English), but there are extremely specialized terms, making it difficult to assess similarity. This dataset is called *ICH_TAGS*.
3. tag lists referring to cooking and ingredients, in Italian, again taken from QueryLab. The interest is to handle data, belonging to a specific domain, in a language other than English, both for semantic and dictionary-based similarity. The pretrained models used are those for Italian. A comparison was made with a multilingual model. In the following called *Cook_IT*
4. WordSim363, gold standard

Figure 2. word cloud for Mariage on Europeana and Rnm.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

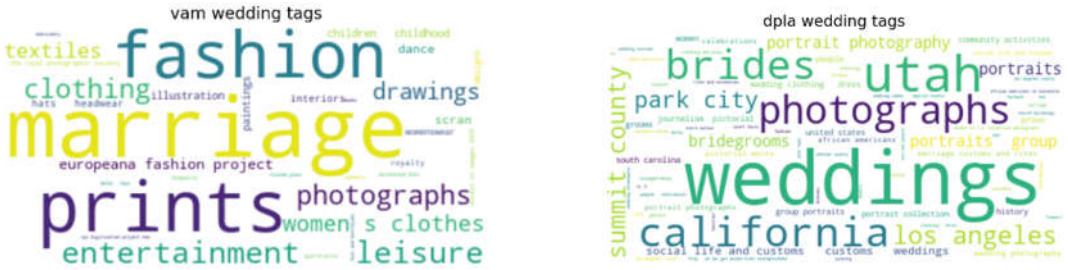

Figure 3. word cloud for Wedding on Victoria & Albert Museum and DPLA.

5.2. Results Evaluation

The performance of the approach on different datasets has been evaluated, and the results were summarized in graphs displaying the recall, precision, f1, Dice and Jaccard scores.

To enhance clarity, the figure labels have been included in Table 3. The individual semantic similarity methods are labeled in Table 3a, while the ensemble results are presented in Table 3b. In Table 3a, each model is depicted as a graph, whereas in Table 3b, each model is presented as a set of results in terms of recall, precision, and F1 scores.

Table 3. a) single semantic method result labels; b) ensemble methods result labels

s1s.[CLS]
s1s.[AVG]
s1s.[MAX]
s1w.path
s1w.wu
s1w.min_dist
s.Jaro
s3s.[CLS]
s3s.[AVG]
s3s.[MAXS]
s3w.path
s3w.wu
s3w.min_dist
e1.model
e3.model
e1s.model
e3s.model

e1.model
e3.model
e3s.model
e3w.model
e1.all
e3.all
e3s.all
e3w.all
e1.max
e3.max
e3s.max
e3w.max

Figure 4 and 5 reports evaluation results on *Mariage* from Europeana and Rnm. It should be noted that Europeana is a multilingual archive, while Rnm primarily features French content with some metadata available in English. For our approach, we identified two pretrained models that were specifically trained on multilingual datasets, employing automatic language detection.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

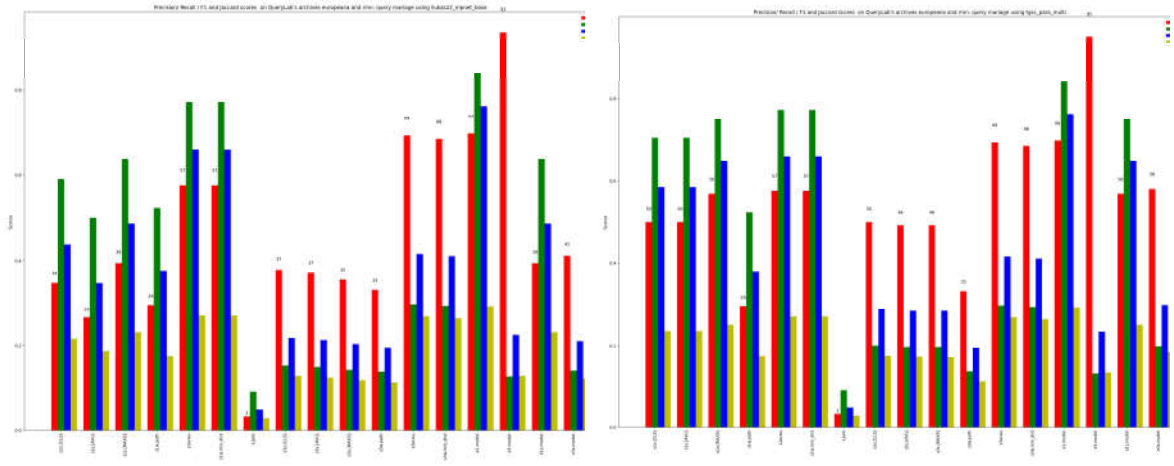

Figure 4. single methods results for Mariage. Labels are those of Table 3a.

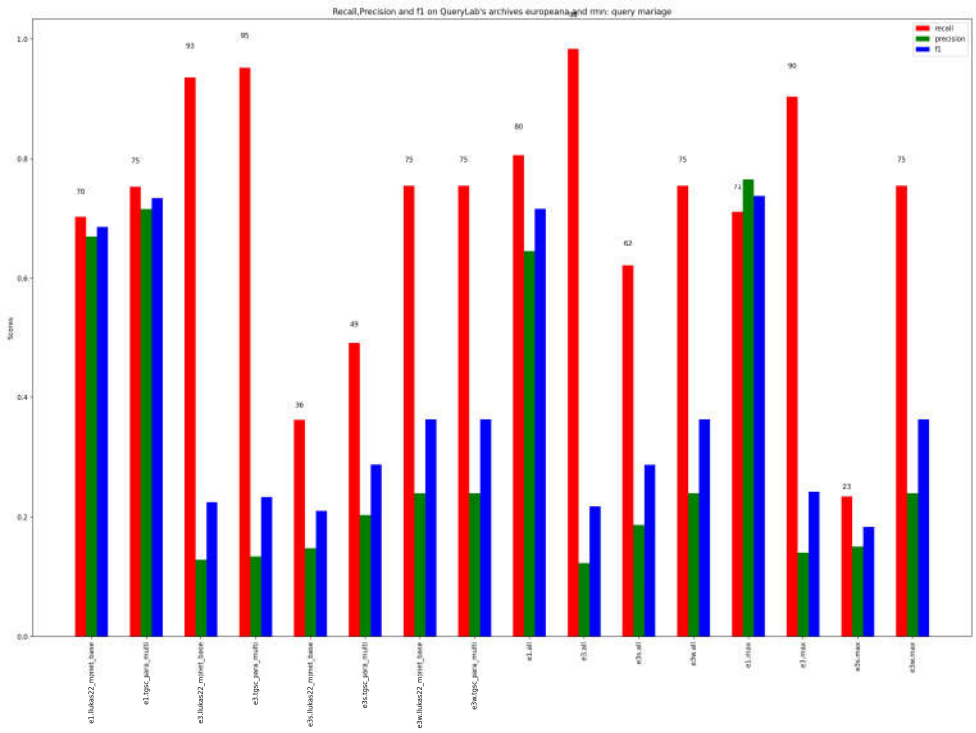

Figure 5. ensemble methods results for Mariage. Labels are those of Table 3b.

In Figure 4, the depicted outcomes correspond to the single similarity methods as previously defined. Specifically, for the semantic similarity methods, the results pertain to the two language models used for multilingual texts. The figure presents the results obtained by combining the individual methods using a weighted voting strategy applied to the methods within the current model.

Our primary focus is on achieving high recall, indicated by the red color in the graph, considering the objective of the approach. When examining single semantic models, where only the first result (s1s) is taken into account, the recall is comparatively lower than that of WordNet-based methods (s1w). This observation holds true even when considering the top three results (s3s and s3w). As expected, string-based similarity yields very low results, and it is utilized in ensemble methods to reinforce similarity when it surpasses a specified threshold, empirically set at 0.9.

The last 4 columns of figure 4, obtained integrating the results using ensemble methods on single language model, the recall for considering all methods applied to the first three results achieves a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

promising 93 (e3.model), surpassing both the recall obtained by solely considering the term evaluated as most similar through majority voting (recall 69 with e1.model) and employing only semantic methods. The latter approaches yield a recall of 39 and 41 for a single term (e1s.model) and the first three terms (e3s.model), respectively.

Figure 5 presents the ensemble results for different scenarios, including all methods or limiting only to semantic methods or WordNet-based methods. The highest recall is obtained by using a result composition method that considers all pre-trained models and takes the best three results (e3.all). This result significantly improves both the scores obtained by semantic and/or WordNet-based methods alone and by taking only the best result. We do not observe much variation in the use of different pre-trained language models.

Figures 6 and 7 show the results obtained on the wedding query in the DPLA and Victoria&Albert Museum archives. In this case the tags were identified in English and the pretrained language models for that language were used. The results obtained replicate those for *Mariage*. Again, the ensemble methods both for single model and overall on all models outperform the single methods. The solution of taking the first three results yields a recall of 98 percent with the e3.model, in both models (figure 6). This value reaches 100% recall in e3.all.

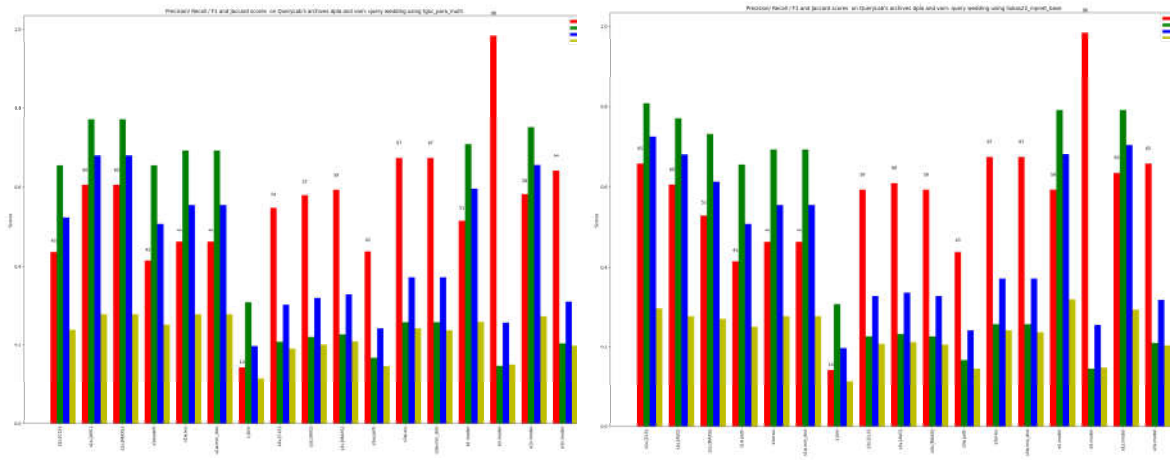

Figure 6. single methods result for *Wedding*. Labels are those of Table 3a.

The distinctive feature of ICH tags is their utilization of highly specific terms. The potential challenge posed by classical word embeddings, which could result in out-of-vocabulary (OOV) terms, is effectively addressed through the tokenization process of transformers. As a result, the similarity between terms is accurately computed, even in cases where the model is unfamiliar with individual words, thanks to the incorporation of tokenizers and fine-tuning.

As shown in figures 8 and 9, the previous results are also confirmed here: the best single model result is from e3.model, and the best ensemble methods result is e3.all.

Figures 10 and 11 pertain to the Italian dataset comprising terms associated with recipes and ingredients used in Italian cuisine. These terms are commonplace within the specific domain but not typically found in a general-purpose lexicon. However, by employing a tokenizer, this challenge is effectively addressed, enabling the measurement of similarity between the two sets with promising outcomes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

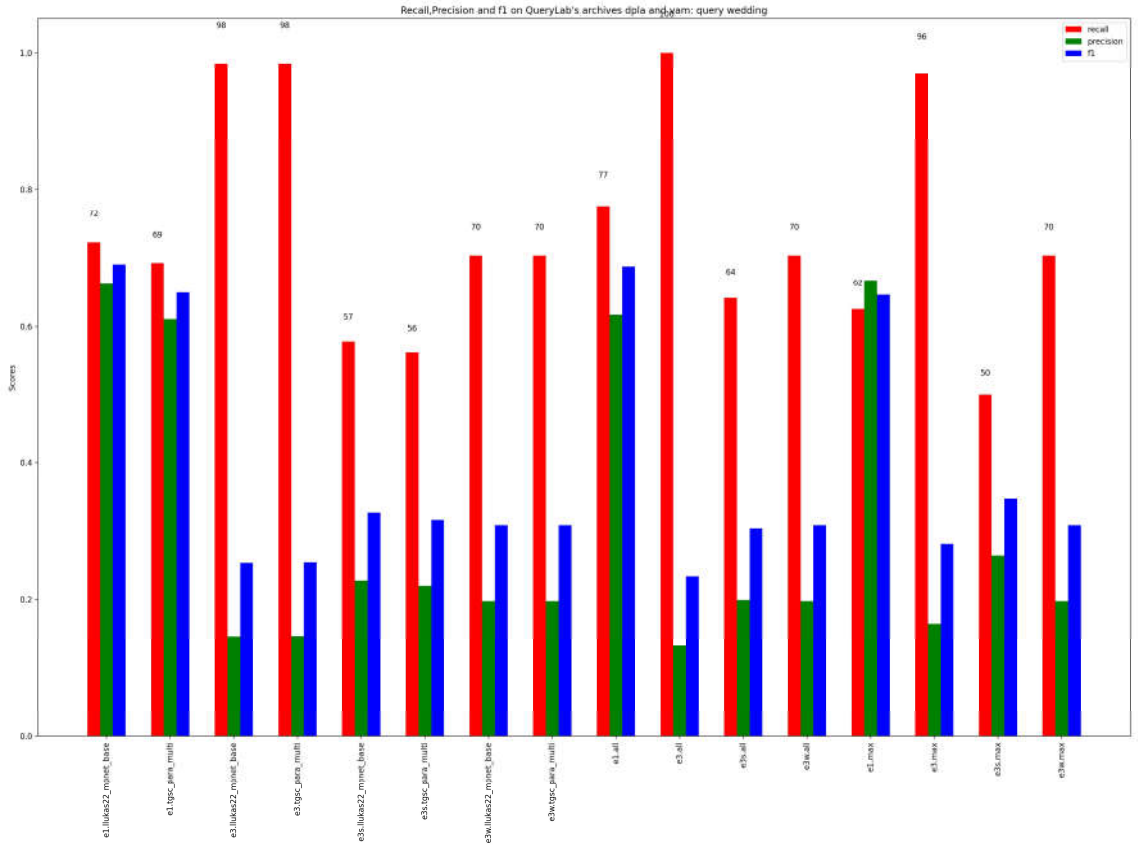

Figure 7. ensemble methods result for *Wedding*. Labels are those of Table 3b.

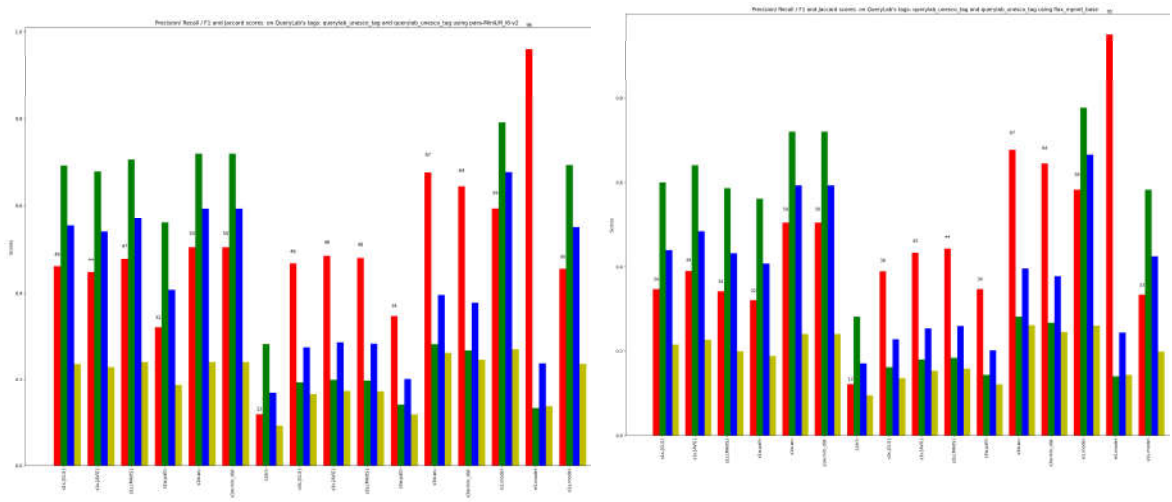

Figure 8. single methods result for ICH tags. Labels are those of Table 3a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

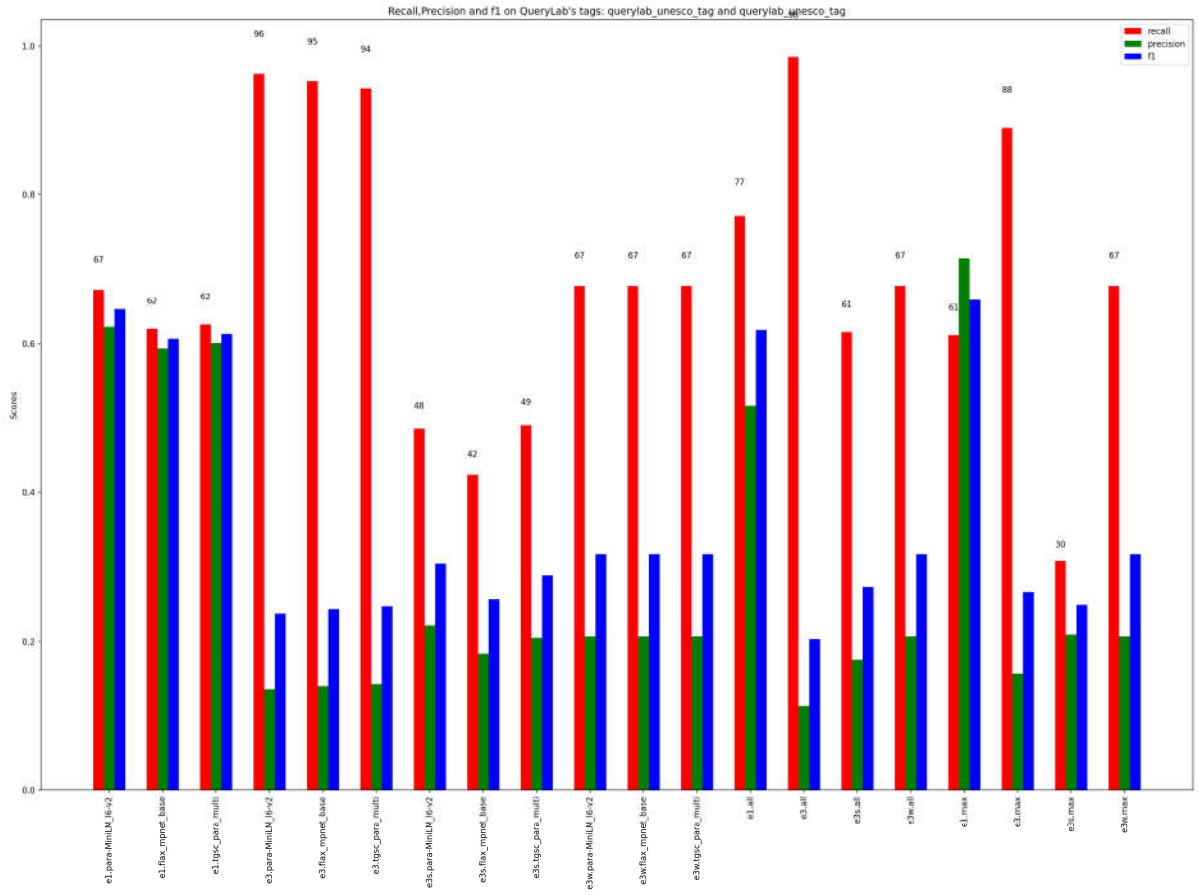

Figure 9. ensemble methods result for ICH tags. Labels are those of Table 3b.

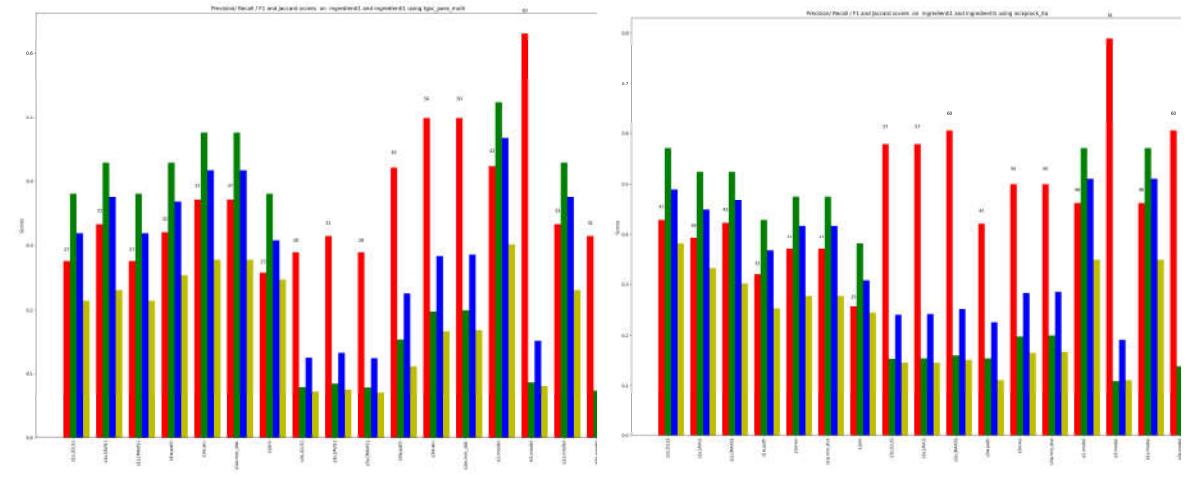

Figure 10. single methods result for cook_IT. Labels are those of Table 3a.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

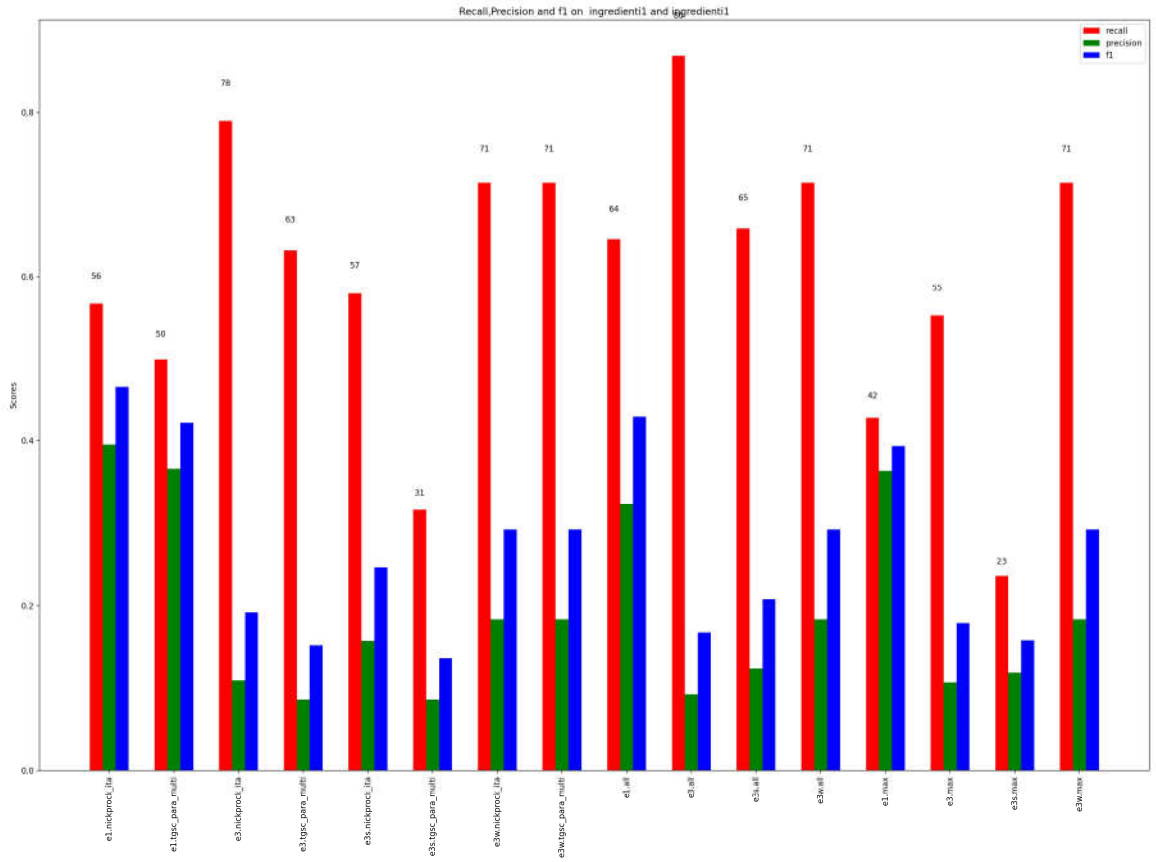

Figure 11. ensemble methods result for cook_IT. Labels are those of Table 3b.

Figure 12. and 13 are for the gold standard dataset. The recall in this case is very high for most of the single methods. Once again, it is observed that ensemble methods focusing on the first three results demonstrate superior performance. This finding applies to both individual models (Figure 3, e3.model) and ensemble methods that combine multiple methods and language models (Figure 4, e3all). In both cases, considering the first three results leads to improved outperforms compared to considering only the most similar term or using other combination mechanisms.

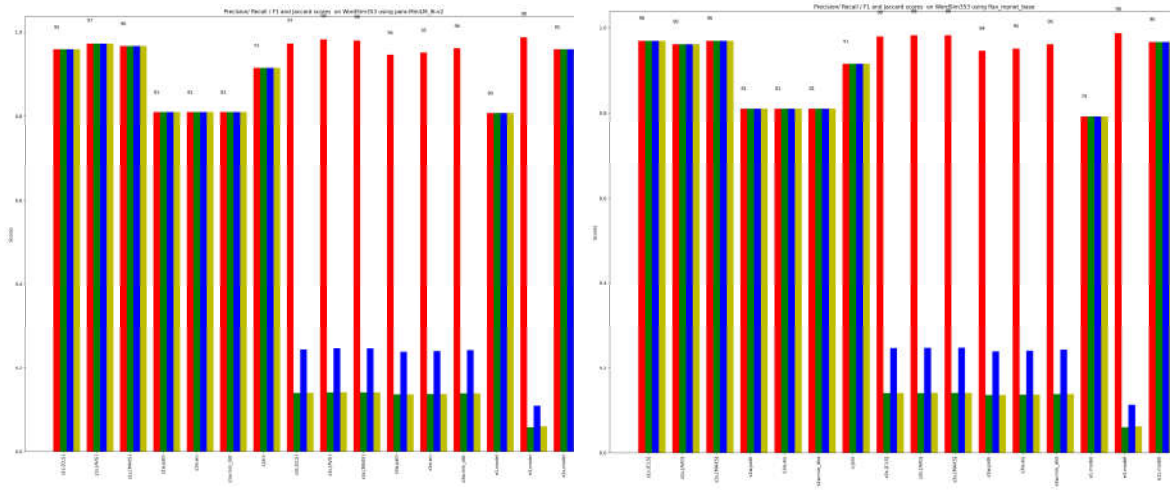

Figure 12. single methods result for WordSim353. Labels are those of Table 3a.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0690.v1

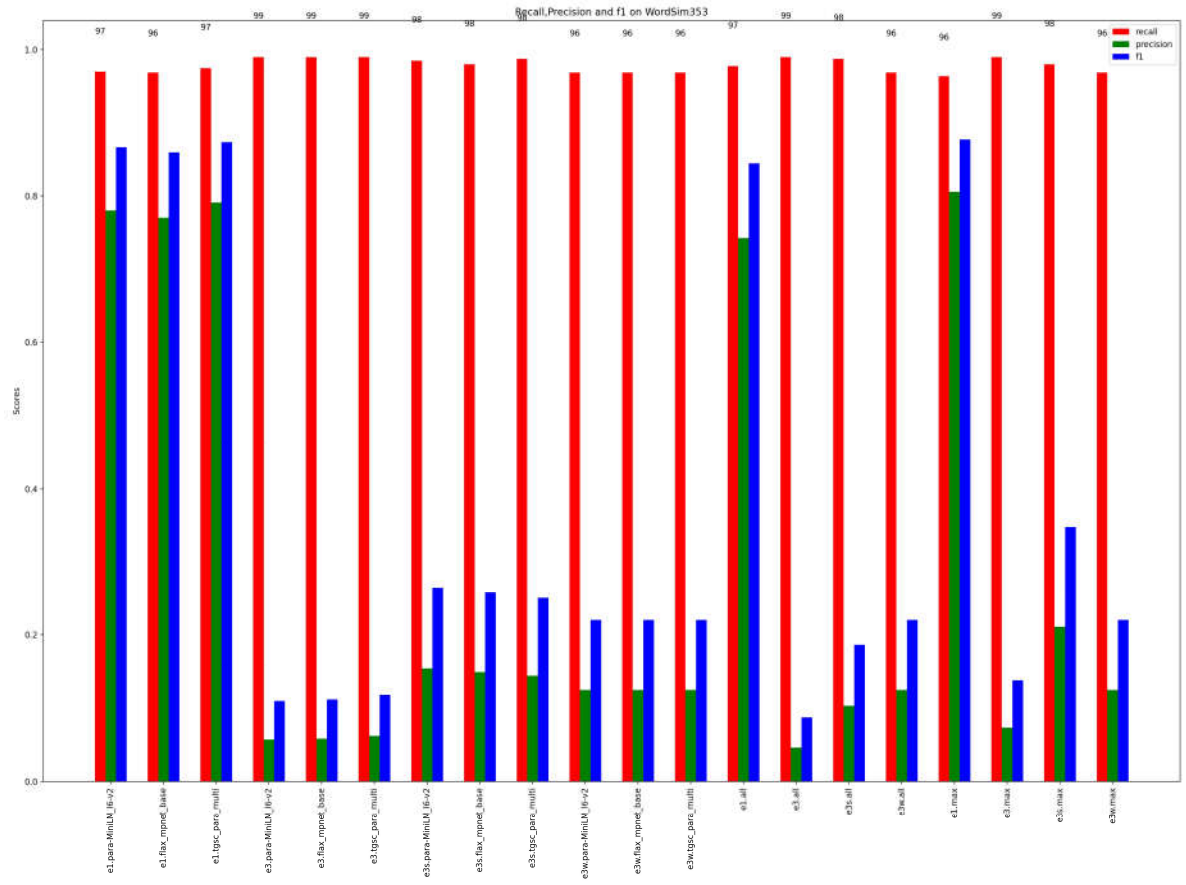

Figure 13. ensemble methods result for WordSim353. Labels are those of Table 3b.

It is evident that when considering the top 3 results, the recall achieves its peak values, whereas the precision remains relatively low. This discrepancy arises because the identification of the most similar tags results in an expansion of the number of tags, leading to a decrease in precision. However, in the context of the study, this is not a concern since the objective is to retrieve all tags that are even remotely related, making it more inclined towards a "recall-oriented" approach rather than precision-focused.

The experimental setup has been implemented in Python 3.10, using standard packages like Numpy, Matplotlib, Pandas and other more specific ones for processing of textual data such as NLTK, Gensim, and Sklearn, together with Pytorch, Transformers, Spacy and deep_translate and some experimental packages in GitHub. We used various pretrained models taken from HuggingFace and datasets from GitHub.

6. Conclusions and future works

In this article, we have presented an approach for evaluating similarity between tags, both single and compound words, belonging to different datasets. This approach integrates various state-of-the-art methods and techniques. Specifically, we have tested methods that utilize pretrained language models, methods based on the hierarchical structure of WordNet, and string-based similarity measures used solely as reinforcement in majority voting mechanisms for ensemble models. These measures were subsequently composed using ensemble methods that showed the effectiveness of the method. Experiments were conducted on various datasets characterized by different challenges, such as multilingual datasets, Italian-only datasets, or datasets with highly specialized terms. In all the experiments, the best results were achieved by integrating the outcomes of different single similarity methods, both within each language model and globally across all models. The recall exceeded 90% for the top three results. Experimentation using the gold standard also produced optimal results.

The uniqueness of this approach lies in its fully unsupervised nature, once the most suitable language models have been identified. It can be applied, almost without adaptation, to different datasets in real-world scenarios. Another characteristic of the method is its ability to adapt to multilingual datasets, thanks to automatic language detection. This feature is beneficial for both semantic models, as it aids in selecting the appropriate language model, and for WordNet-based models. The automatic language detection enables the method to seamlessly handle different languages within the dataset.

The results obtained so far have been encouraging in terms of recall and precision. The next step is to integrate it into QueryLab to fully evaluate its functionality, usability, and usefulness. This integration will provide users with additional flexibility in accessing and searching for information, enhancing their overall experience, and allowing for more effective queries.

References

1. Atoum, I., & Otoom, A. A Comprehensive Comparative Study of Word and Sentence Similarity Measures. International Journal of Computer Applications, 975, 8887.
2. Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *international journal of Computer Applications*, *68*(13), 13-18.
3. Gupta, A., Kumar, A., Gautam, J., Gupta, A., Kumar, M. A., & Gautam, J. (2017). A survey on semantic similarity measures. *International Journal for Innovative Research in Science & Technology*, *3*(12), 243-247.
4. Sunilkumar, P., & Shaji, A. P. (2019, December). A survey on semantic similarity. In 2019 International Conference on Advances in Computing, Communication and Control (ICAC3) (pp. 1-8). IEEE.
5. Wang, J., & Dong, Y. (2020). Measurement of text similarity: a survey. *Information*, *11*(9), 421.
6. Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology*, *6*(1), 1-12.
7. Atoum, I., & Otoom, A. (2016). Efficient hybrid semantic text similarity using WordNet and a corpus. International Journal of Advanced Computer Science and Applications, 7(9).
8. Ensor, T. M., MacMillan, M. B., Neath, I., & Surprenant, A. M. (2021). Calculating semantic relatedness of lists of nouns using WordNet path length. *Behavior Research Methods*, 1-9.
9. Kenter, T., & De Rijke, M. (2015, October). Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 1411-1420).
10. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint.
11. Chandrasekaran, D., & Mago, V. (2021). Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, *54*(2), 1-37.
12. Zad, S., Heidari, M., Hajibabaee, P., & Malekzadeh, M. (2021, October). A survey of deep learning methods on semantic similarity and sentence modeling. In *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 0466-0472). IEEE.
13. Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyandé, T. F., Klein, J., & Goujon, A. (2021, April). A comparison of pre-trained language models for multi-class text classification in the financial domain. In *Companion Proceedings of the Web Conference 2021* (pp. 260-268).
14. Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H., & Luo, Y. (2023). A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, *30*(2), 340-347.
15. Wang, H., Li, J., Wu, H., Hovy, E., & Sun, Y. (2022). Pre-Trained Language Models and Their Applications. *Engineering*.
16. Guo, T. (2021). A Comprehensive Comparison of Pre-training Language Models (Version 7). TechRxiv. https://doi.org/10.36227/techrxiv.14820348.v7
17. Artese, M. T., & Gagliardi, I. (2022). Integrating, Indexing and Querying the Tangible and Intangible Cultural Heritage Available Online: The QueryLab Portal. Information, 13(5), 260.
18. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Xu, C. (2020). Transformers: State-of-the-art natural language processing. Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (p. 38-45). Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-demos.6
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
20. Fellbaum, C. (2010). WordNet. In *Theory and applications of ontology: computer applications* (pp. 231-243). Dordrecht: Springer Netherlands.
21. Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

22. Hassan, B., Abdelrahman, S. E., Bahgat, R., & Farag, I. (2019). UESTS: An unsupervised ensemble semantic textual similarity method. *IEEE Access*, *7*, 85462-85482.
23. Rychalska, B., Pakulska, K., Chodorowska, K., Walczak, W., & Andruszkiewicz, P. (2016, June). Samsung Poland NLP Team at SemEval-2016 Task 1: Necessity for diversity; combining recursive autoencoders, WordNet and ensemble methods to measure semantic similarity. In *Proceedings of the 10th international workshop on Semantic Evaluation (SemEval-2016)* (pp. 602-608).
24. Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich (2008). Introduction to Information Retrieval. Cambridge University Press.
25. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001, April). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web* (pp. 406-414).
26. Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665-695.
27. Artese, M. T., & Gagliardi, I. (2022). Methods, Models and Tools for Improving the Quality of Textual Annotations. Modelling, 3(2), 224-242.
28. Artese, M. T., & Gagliardi, I. (2017). Inventorying intangible cultural heritage on the web: a life-cycle approach. International Journal of Intangible Heritage(12), 112-138.
29. Unesco ICH. (2023, May). Retrieved from Browse the Lists of Intangible Cultural Heritage and the Register of good safeguarding practices: https://ich.unesco.org/en/lists