Performance Evaluation of Similarity Measures on Similar and Dissimilar Text Retrieval

Victor U. Thompson, Christo Panchev and Michael Oakes
University of Sunderland, Edinburgh Building, Chester Rd, Sunderland SR1 3SD,
Department of Computing, Engineering and Technology, St. Peters campus, Sunderland, U.K.

Keywords: Candidate Selection, Information Retrieval, Similarity Measures, Textual Similarity.

Abstract: Many Information Retrieval (IR) and Natural language processing (NLP) systems require textual similarity measurement in order to function, and do so with the help of similarity measures. Similarity measures function differently, some measures which work better on highly similar texts do not always do so well on highly dissimilar texts. In this paper, we evaluated the performances of eight popular similarity measures on four levels (degree) of textual similarity using a corpus of plagiarised texts. The evaluation was carried out in the context of candidate selection for plagiarism detection. Performance was measured in terms of recall, and the best performed similarity measure(s) for each degree of textual similarity was identified. Results from our Experiments show that the performances of most of the measures were equal on highly similar texts, with the exception of Euclidean distance and Jensen-Shannon divergence which had poorer performances. Cosine similarity and Bhattacharryan coefficient performed best on lightly reviewed text, and on heavily reviewed texts, Cosine similarity and Pearson Correlation performed best and next best respectively. Pearson Correlation had the best performance on highly dissimilar texts. The results also show term weighing methods and n-gram document representations that best optimises the performance of each of the similarity measures on a particular level of intertextual similarity.

1 INTRODUCTION

Similarity measures are needed in many IR and NLP tasks such as document clustering (Huang, 2008), plagiarism detection (2003), text categorization (Bigi, 2003), and duplicate and near duplicate detection (Broder, 1997; Charika, 2002). The success of many IR systems to a large extent depends on similarity measures (Polettini, 2004). There are diverse kinds of similarity measures in the literature (Cha, 2007), and they all differ in terms of functionality; a similarity measure that is effective in addressing one measurement problem may not be effective in another. For example, Hoad and Zobel (2003) argue that Cosine similarity is not effective for detecting co-derivatives, and that Cosine is most effective when used for similarity measurement between texts of different lengths. Co- derivatives are documents that share significant portion of texts (i.e. when one document is derived from the other or both are derived from a third document Bernstein and Zobel, 2004). In a similar way, Jones and Furnas (1987) emphasized the importance of using the right similarity measure for a given textual similarity measurement task.

Several studies have been carried out in the literature to evaluate the performance of popular similarity measures (Strehl et al., 2000, White et al., 2004; Huang, 2008; Ljubesic et al., 2008; Forsyth and Sharoff, 2014). Most of these studies were either focused on the performance of single similarity measure in isolation, or on the performance of selected similarity measures in addressing only one level (degree) of textual similarity. What these studies failed to explore in detail is that there are different levels of intertextual similarity; some measures which work better on highly similar texts do not always work so well on highly dissimilar texts. Hence in this paper, we evaluated the performances of eight (8) popular similarity measures on four levels of textual similarity using a corpus of plagiarised texts (highly similar text, lightly reviewed texts, heavily reviewed texts and highly dissimilar texts). Our evaluation was carried out in the context of plagiarism detection (extrinsic plagiarism detection). Extrinsic or external plagiarism detection involves detecting overlapping

Thompson, V., Panchev, C. and Oakes, M..
Performance Evaluation of Similarity Measures on Similar and Dissimilar Text Retrieval.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 577-584
ISBN: 978-989-758-158-8

portions of texts in two documents (Potthast et al., 2009; Cloughs and Stevenson, 2011). The particular stage of plagiarism detection we used as our evaluation task was candidate selection; this involves selecting a list of source documents that are most similar to a suspicious one to be used for a later, more detailed analysis (Gollub et al., 2013). Source documents are the original documents from which text passages are removed (and altered in some cases) and placed in suspicious documents. Candidate selection is an important step in plagiarism detection because it reduces the workload on the next stage of the detection process; a stage that requires exhaustive search for overlapping portions of texts in compared documents, it is computationally expensive and time consuming.

We implemented the similarity measures using the vector space model (see section V. page [2] for details) in combination with the n-gram language model, as well as with different term weighting methods (TF-IDF, TF and Binary) to optimize performance. We measured performance in terms of recall.

The rest of this paper is divided as follows: section II highlights related work on the evaluation of text similarity measures. Section III discusses similarity measures and their basic properties in relation to text similarity measurement, and then outlines the eight similarity measures used this study. Section IV is a concise description of the levels of textual similarity the measures we used for evaluation. Section V describes the evaluation task used in this study, Section VI discusses the methods used in accomplishing the evaluation task. Section VII describes the experiments carried out; the corpus used and the experimental procedures. The results we obtained are presented and discussed in section VIII, section IX concludes this paper with a brief summary of the contributions, and points out areas for future work.

2 RELATED WORK

In an attempt to measure the impact of similarity measures on web-based clustering (web-document categorization), Strehl et al; (2000) evaluated the performances of four similarity measures (Cosine similarity, Euclidean distance, Jaccard index and Pearson Correlation) on several clustering algorithms. The intuition behind this study was that accurate similarity measurement results in better clustering. Experimental results from the Strehl et al., (2000) study showed that Cosine and Jaccard similarity performed best, while Euclidean distance performed the least. White and Jose (2004) evaluated the performance of eight similarity measures according to how well they could classify documents by topic. Results from White and Jose's experiment show that correlation coefficient outperformed the other measures by outputting predictions (topic similarity) that aligned more closely with human judgment. In an attempt to extract semantic similarity from text documents, Ljubesic et al; (2008) experimented with eight different similarity measures, and found that Jenssen-Shannon divergence, Manhattan distance (L1) and Euclidean distance (L2) performed best, outperforming standard IR measures like Cosine and Jaccard similarity. Huang (2008) extended the works of Strehl et al. by including an additional similarity measure (Kullback–Leibler divergence) and using the k-means clustering algorithm with n=1. Results from the Huang experiments show that clustering based on Pearson correlation and Kullback–Leibler divergence was more accurate, while document clustering based on Pearson correlation and Jaccard similarity were more cohesive (where cohesive means similarity or closeness between members of a cluster). The worst performance came from Euclidean distance. In a more recent study, Forsyth and Sharoff (2014) proposed an approach for measuring the performance of similarity measures against human judgment as reference. Results obtained from their study revealed that Pearson correlation outperformed standard measures like Cosine and Kullback–Leibler divergence (KLD).

This study differs from the above study in that it focuses on evaluating similarity measures on different levels of intertextual similarity. This is due to the fact that a similarity measure that is effective on one level of textual similarity may not be effective on another. In addition, the above studies did not consider the effect of term weighting methods and document representation models on performance. These are optimization factors that should be carefully chosen; knowing which term weighting method and document representation model to use with a particular similarity measure on a particular level of intertextual similarity is important for retrieval performance. This paper addresses both problems empirically.

3 SIMILARITY MEASURES

Similarity measures are functional tools used for measuring the similarity between objects. When

used for measurement, the output of a similarity measure is a numeric value usually in the range of 0 and 1, where 0 means completely dissimilar and 1 means exactly similar. A proper similarity (or distance) measure is defined by the following properties; a similarity measure (1) must be symmetrical (2) must satisfy the triangular inequality (3) must satisfy the similarity property. For details on these properties, see Oakes (2014).

According to the literature, there are three major groups of similarity measures, they include string-based, corpus-based and knowledge-based (Mihalcea et al., 2006; Gomaa and Fahmy, 2013). The string-based group is further divided into character-based and term-based. Knowledge-based and corpus-based similarity measures apply semantic similarity measurement techniques such as Latent semantic analysis (LSA) (Deerwester, 1990 et al), pointwise mutual information (Turney, 2001) or lexical databases such as WordNet for measuring textual similarity. In this study, the main focus is on term-based similarity measures because they are relatively more efficiency on high dimensional data such as documents, and for the most part, they are standard in IR for addressing many document similarity measurement problems. Term based similarity measures use statistics derived from texts to compute their similarity. Such statistics include Term frequency, inverse document frequency, document length etc.

The following similarity measures were implemented in this study; Cosine similarity, Jaccard similarity, Bhattacharryyan coefficient, Dice coefficient, Pearson correlation coefficient (PCC(R)), Euclidean distance, Kullback–Leibler divergence and Jensen-Shannon divergence. Similarity measures are usually implement on vectors, hence given any two document vectors $\vec{U},\vec{V}$ over vector space $\{i\text{----}z\}$, the similarity of $\vec{U},\vec{V}$ could be computed using any of the following similarity measures;

$$\text{Bhat}\left(\vec{U},\vec{V}\right)=\sum_{i=1}^{z}\sqrt{\left(\sum u_i\cdot\sum v_i\right)}$$

$$\text{Cosine similarity}\left(\vec{U},\vec{V}\right)=\frac{\sum_{i=1}^{z}u_i v_i}{\sqrt{\sum_{i=1}^{z}(ui)^2}\sqrt{\sum_{i=1}^{z}(vi)^2}}$$

$$\text{Euclidean distance}\left(\vec{U},\vec{V}\right)=\sqrt{\sum_{i=1}^{z}\left(u_i-v_i\right)^2}$$

$$\text{Ext Jaccard}\left(\vec{U},\vec{V}\right)=\frac{\sum_{i=1}^{z}u_i v_i}{\sum_{i=1}^{z}(ui)^2+\sum_{i=1}^{z}(vi)^2-\sum_{i=1}^{z}u_i v_i}$$

$$\text{KLD}\left(\vec{U},\|\vec{V}\right)=\sum_{i}u_i*\log\frac{u_i}{v_i}$$

$$\text{JSD}\left(\vec{U},\vec{V}\right)=\text{dKLD}(U\|\frac{U+V}{2})+\text{dKLD}(V\|\frac{U+V}{2})$$

$$\text{PCC}\left(\vec{U},\vec{V}\right)=\frac{\sum_{i=1}^{z}(Ui-\text{mean}(U))(Vi-\text{mean}(V))}{\sqrt{\sum_{i=1}^{z}(Ui-\text{mean}(U))^2}\sqrt{\sum_{i=1}^{z}(Vi-\text{mean}(V))^2}}$$

Note: There are two measurement variables that determines the similarity between objects in a vector space, they include vector length (extent of similarity) and direction/angle (content /topic similarity) (Zhang and Korfhage (1999). A query and a document vector are exactly similar if they have equal length and zero angular distance between them, and they are completely different if one is orthogonal to the other. However, in terms of document similarity, angular distance matters most as it is a clear reflection of content similarity.

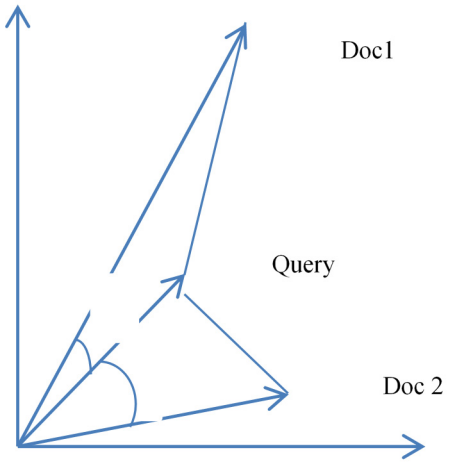

Figure 1: Comparing a query and documents using angular distances between vectors and distances between vector lengths, where the length of a vector is indicated by the arrow sign.at the peak.

4 DESCRIPTION OF TASK

Candidate selection in external plagiarism detection is a typical retrieval task where documents are ranked according to their relevance to a query document (suspicious document). The task involves comparing a suspicious document with a collection

(database) of source documents using relevant similarity measures. After the comparison, documents are sorted by their similarity scores, and the top *K* documents (with highest similarity scores) are selected as candidates for further analysis. Just like most IR tasks, the similarity measures are employed to capture semantic relationships between text documents.

5 METHODS

The task of Candidate selection is very similar to document ranking in IR. The two most popular approaches for ranking documents in IR are; the vector space model (VSM) and the probabilistic model (Baeza-Yates and Ribeiro-Neto, 1999; Manning et al., 2008). In the VSM (Salton et al., 1975), documents and queries are represented as vectors in space; ranking is done based on the relevance of documents to a query using similarity scores. The VSM can therefore be implemented with similarity measures, as well as with different term weight methods. Probabilistic models such as the binary independent model (BIM) (Robertson, 1977) or the recently proposed language model (Ponte and Croft, 1998; Hiemstra and De Vries, 2000), represent documents as probability distributions and rank them based on their probability to a query. Probabilistic models are usually not implemented with similarity measures and models such as BIM are based on naïve assumptions, and hence not suitable for this study. However, the language model has proven to be effective, and has even outperformed the VSM in some studies (Hiemstra and De Vries, 2000). One unique characteristics of the language model is that it uses n-grams as index terms. N-grams preserve word order and discriminate documents based on overlapping phrases, which makes them really relevant to this study as plagiarised texts often occur as phrases, and detecting plagiarised documents involves searching for local similarity (Oakes, 2014). Hence in this research, the VSM was adopted, but in conjunction with the n-gram capability of the language model to optimize retrieval performance.

5.1 Transformation of Documents to N-gram Vector Space Model

In order to implement the VSM, each document must be transformed into a vector by indexing and assigning weights to indexed terms (words or sequence of words). Indexing allows for rapid comparison of documents (Baeza-Yates and Ribeiro-Neto, 1999). Assigning weights to indexed terms ensures that terms are well represented according to their discriminatory power in a document. In this study, we transformed documents to vectors using the following steps; data pre-processing, transformation to n-grams and term-weighting.

5.1.1 Data Pre-Processing

Data pre-processing tokenizes texts and removes unwanted terms. Steps in data pre-processing includes tokenization, stop-word removal and stemming (Manning et al., 2008). Tokenization is the process of parsing texts into tokens (bag-of-words). Stop-words are commonly found words in documents (such as ‘a’, ‘the’, ‘and’, ‘what’). Their contribution to document comparison is almost insignificant; hence they are often removed. Stemming reduces words to their root form thereby increasing the chances of overlap (i.e. ‘friendly’, ‘friendship’, ‘friend’ all reduced to ‘friend’) and precision. Stemming can result in an increase in algorithm efficiency (Manning et al., 2008).

5.1.2 Transformation of Documents to N-Grams

N-gram document models enable similarity to be measured on the basis of overlapping sequence of words (phrases) rather than individual words. N-grams capture some form of syntactic similarity between documents and avoid the drawback of word-independence assumption that limits the bag-of-word model (Johnson and Zhang, 2015). N-grams were basically used in this study to discriminate and categorise documents based on similar n-gram sizes. For example majority of highly similar documents can be detected using higher order n-grams (n-grams of longer lengths) than lightly reviewed documents. Hence using n-grams of certain lengths (size) can help discriminate one class of similar documents from another. However the size of an n-gram model should be carefully chosen to avoid bypassing potential plagiarised documents or detecting documents of nearby categories resulting in false positives and a decrease in performance. In this study we tested n-grams of different lengths in order to obtain the best n-gram for a particular category.

5.1.3 Term Weighting

After pre-processing, documents are transformed to vectors by assigning weights to the terms in a

document. Popular term weighting methods include term frequency (TF), Term frequency inverse document frequency (TF-IDF) and binary weighting (Salton and Buckley, 1998) Term frequency is the number of occurrences of a term in a document. TF-IDF is a global weighting method (meaning that the weighting takes into consideration other documents in the corpus) where rare terms with higher discriminating power are assigned more weights than commonly found terms in a corpus (Manning et al., 2008). TF-IDF is simply the multiplication of term frequency (TF) and the inverse document frequency (IDF) (Sparck Jones, 1972; Robertson, 2004). The TF of a term (t) can be derived as described above, while the IDF of (t) can be derived by dividing the corpus size (the number of documents in the corpus) by the number of documents in which the term occurs. Both TF and TF-IDF are often normalised by document's length. Length normalisation helps in cancelling out any bias in favour of longer documents (Singhal et al., 1996A). We used Cosine length normalisation in this study because it is very popular and has had remarkable success with the VSM (Singhal et al., 1996B). The binary weighting method is one that assigns a weight of one to each term found in a document as long as it appears once or more; terms which do not appear at all are given a weight of 0.

5.2 Document Comparison

Term weighting completes the transformation to vectors; document comparison can then be carried out between a query vector and a collection of source document vectors in a vector space using a similarity measure. For each comparison, documents are ranked in decreasing order of similarity based on their similarity scores, and the top K documents can then be selected as candidate set.

6 EXPERIMENTS

Corpus: The corpus used in this experiment is the PAN@Clef 2012 text alignment corpus. It is artificially generated and comprises of 6500 documents; of which 3000 are suspicious documents (plagiarised at different degrees) and the remaining 3500 are source documents (the original documents where the plagiarised passages were taken from). The corpus is made up of six categories of textual similarity, however, only four of these categories are relevant to this study, namely: no obfuscation (highly similar), low-obfuscation (lightly reviewed), high-obfuscation (heavily reviewed) and no-plagiarism (highly dissimilar). Each category was created by removing one or more passages from a source document and altering them by paraphrasing, replacing some of the texts with their synonyms etc. before pasting the passages in a suspicious document. The alteration was done with different intensity to separate one level of textual similarity from the other. Hence all the suspicious documents in the same category were altered with the same intensity. The corpus comes with a ground-truth for evaluation purpose; pairs of documents and their appropriate categories according to human Judgement.

6.1 Description of Experiments

The similarity measures were implemented using the vector space document representation with TFIDF, TF and Binary weighting methods and different lengths of word n-grams. The measures were evaluated on the four categories of textual similarity mentioned above, one rewrite level at a time.

Half of the corpus was used to develop algorithms for implementing the similarity measures. In the algorithm development stage (training stage), the best term weighting method and n-gram level for each similarity measure and for each category of textual similarity were determined. In determining the best term weighting method for a particular similarity measure, we implemented the similarity measure with TF, TFIDF and binary weighting methods, and the term weighting method that resulted in the best performance was noted as the most suitable term weighting to use with that similarity measure on that particular category. The same procedure was used to determine the best n-gram document model; different sizes of n-grams were run (starting with one gram and progresses upwards). The n-grams and term weighting parameters were then used to run the similarity measures on the other half of the corpus. Each suspicious document was compared with the collection of source documents in the corpus using the selected similarity measures. For every query document (suspicious document) run, recall was measured at retrieval intervals of [1,5,10,20,30,50,60,70]; for example, recall was measured when only the highest ranking document was retrieved, and again when the top 5 documents were retrieved, and so on until the top 70 documents were retrieved. Performance was measured in terms of recall; where recall is the number of relevant documents retrieved divided by the total number of

relevant documents expected. Performance was measured in recall because in candidate selection what really matters is the retrieval of relevant candidate documents, and not how precise the measurement algorithm is.

7 RESULTS AND DISCUSSION

Tables 1-5 display the results from the experiments carried out according to levels of textual similarity. Each table contains the similarity measures and their respective performances measured in recall. The tables also show the weighting methods and word n-grams used.

For highly similar texts, the performances of the measures were high and equal, except for Euclidean distance and JSD with lower performances. For lightly reviewed texts, Cosine similarity and the Bhattacharryan coefficient performed best, they both have a recall of 0.96. For heavily reviewed texts, Cosine similarity and PCC(R) outperformed the others both having a recall of 0.88, the second best performance was 0.81; this suggests that increase in textual rewriting does not affect the performance of Cosine and PCC(R)as much as it does for the other measures. For highly dissimilar texts, PCC(R) emerged as the best performer with a recall of 0.787. The performances of the measures were lowest on highly dissimilar texts.

The main reason for the high performance for majority of the similarity measures on the highly similar category is primarily due to the absence of alterations, and the application of n-grams to clearly discriminate documents. A closer look at the results revealed that the performance of the similarity measure decreased with increase in rewriting (paraphrasing) of the texts. This trend is consistent with Cloughs and Stevenson (2011) findings, and suggests that the more texts are rewritten, the more difficult it is to accurately measure their similarity. As textual alteration increases, the chances of retrieving a false document that happens to share some common terms with a query document increases as well. This ultimately results in increase in false positive, and a decrease in performance. While the above is true for all similarity measure, the results reveal that some similarity measures tend to cope better with altered cases of plagiarism.

The results also show that the similarity measures performed better on highly similar texts when implemented with higher order n-grams than with lower order ones. One can therefore conclude that when the degree of inter-textual similarity is high, to achieve optimum performance, higher order n-grams should be used, and when low, lower order n-grams should be used. The result also show that most of the similarity measures performed well on highly similar texts when combined with TF, while TFIDF seems relatively better for measuring lower

Table 1: Recall for Highly Similar Texts.

Similarity measures	Term weighting	N-grams	Number of retrieved documents (highest ranking documents	
			1	5
Cosine similarity	Binary/TF/TFIDF	10	0.97	1.0
KLD	Binary/TFIDF/TF	12	0.97	1.0
Dice coefficient	Binary/TF	12	0.96	1.0
Jack -index	Binary/TF	12	0.96	1.0
Bhayttacharyan	BinaryTF/TFIDF	12	0.95	1.0
PCC(R)	TF/Binary	10	0.94	1.0
JSD	Binary/TF/TFIDF	10	0.83	0.897
Euclidean distance	TF/Binary	8	0.68	0.73

Table 2: Recall for Lightly Reviewed Similar Texts.

Similarity measures	Term weighting	N-grams	Number of retrieved documents (highest ranking documents)						
			1	5	10	20	30	40	50
Bhayttacharyan	TF	3	0.913	0.933	0.94	0.947	0.96	0.96	0.96
Cosine similarity	TFIDF	3	0.893	0.933	0.94	0.94	0.953	0.953	0.96
Dice coefficient	Binary	3	0.893	0.927	0.933	0.94	0.947	0.947	0.947
Jaccard index	Binary	3	0.893	0.927	0.933	0.94	0947	0.947	0.947
KLD	TF	3	0.853	0.873	0.913	0.933	0.933	0.947	0.947
PCC(R)	TFIDF	1	0.66	0.727	0.807	0.827	0.86	0.873	0.90
JSD	TF	3	0.56	0.633	0.667	0.69	0.697	0.72	0.74
Euclidean distance	TF/IDF	3	0.51	0.613	0.62	0.627	0.633	0.633	0.63

Table 3: Recall for Heavily Reviewed Texts.

Similarity measures	Term weighting	N-grams	Number of retrieved documents (highest ranking documents)								
			1	5	10	20	30	40	50	60	70
Cosine similarity	TFIDF	3	0.527	0.60	0.653	0.66	0.687	0.707	0.733	0.793	0.88
PCC (R)	TFIDF	1	0.50	0.567	0.65	0.72	0.753	0.78	0.827	0.86	0.88
Dice coefficient	Binary	2	0.513	0.60	0.647	0.693	0.713	0.727	0.753	0.78	0.81
Jaccard_index	Binary	2	0.513	0.60	0.647	0.693	0.713	0.727	0.753	0.78	0.81
Bhayttacharyan	TF	2	0.50	0.567	0.613	0.633	0.66	0.713	0.747	0.753	0.78
KLD	TF	3	0.48	0.513	0.607	0.627	0.66	0.673	0.72	0.753	0.78
JSD	TF	3	0.38	0.447	0.487	0.507	0.52	0.54	0.573	0.587	0.587
Euclidean distance	TFIDF	1	0.313	0.373	0.447	0.493	0.527	0.533	0.55	0.577	0.577

Table 4: Recall for Highly Dissimilar Texts.

Similarity measures	Term weighting	N-grams	Number of retrieved documents (highest ranking documents)								
			1	5	10	20	30	40	50	60	70
PCC(R)	TFIDF	1	0.367	0.433	0.50	0.58	0.62	0.66	0.70	0.733	0.787
Cosine similarity	TFIDF	1	0.313	0.40	0.46	0.513	0.58	0.64	0.66	0.673	0.733
Dice coefficient	Binary	2	0.273	0.34	0.46	0.54	0.573	0.61	0.653	0.68	0.707
Jaccard index	Binary	2	0.273	0.34	0.46	0.54	0.573	0.61	0.653	0.68	0.707
Bhayttacharyan	TF	2	0.333	0.301	0.467	0.513	0.547	0.567	0.613	0.667	0.667
KLD	TF	2	0.287	0.347	0.373	0.373	0.407	0.46	0.493	0.527	0.58
JSD	TF	2	0.27	0.32	0.34	0.367	0.38	0.40	0.44	0.487	0.54
Euclidean distance	TF	1	0.247	0.267	0.293	0.313	0.333	0.353	0.387	0.407	0.44

levels of intertextual similarity because very few plagiarised texts are often left after heavy alteration of a plagiarised passage, and such terms should be weighted higher in order to improve their discriminating power, which is exactly what the TFIDF weighting scheme does.

8 CONCLUSION

We evaluated the performances of eight popular similarity measures and determine the best performed measures to be used on four levels of textual similarity (highly similar, lightly reviewed, heavily reviewed and non-plagiarised texts. We determined and confirmed the most suitable term weighting methods to use with each of the similarity measures for optimum performance. This was achieved by implementing each measure with three popular term weighting methods used in IR (Term frequency (TF), Term Frequency Inverse Document Frequency (TFIDF) and Binary). We also determine the best n-gram document representations to use on each level of textual similarity.

Future work will be focused on improving the effectiveness of the similarity measures, with more emphasis on highly altered plagiarised texts using both semantic similarity measurement and other relevant techniques.

REFERENCES

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). New York: ACM press.

Bernstein, Y., & Zobel, J. (2004, January). A scalable system for identifying co-derivative documents. In *String Processing and Information Retrieval* (pp. 55-67). Springer Berlin Heidelberg.

Bigi, B. (2003). *Using Kullback-Leibler distance for text categorization* (pp. 305-319). Springer Berlin Heidelberg.

Broder, A. Z. (1997, June). On the resemblance and containment of documents. In *Compression and Complexity of Sequences Proceedings* (pp. 21-29). IEEE

Cha S. (2007), "Comprehensive survey on distance/similarity measures between probability density functions." *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1, Issue 4, pp. 300-307.

Charikar, M. S. (2002, May). Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing* (pp. 380-388). ACM.

Clough, P., & Stevenson, M. (2011). Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, *45*(1), pp.5-24.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JAsIs*, *41*(6), 391-407.

Eiselt, M. P. B. S. A., & Rosso, A. B. C. P. (2009).

Overview of the 1st international competition on plagiarism detection. In *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse* (p. 1).

Forsyth, R. S., & Sharoff, S. (2014). Document dissimilarity within and across languages: A benchmarking study. *Literary and Linguistic Computing*, *29*(1), 6-22.

Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P.. & Stein, B. (2013). Recent trends in digital text forensics and its evaluation. In*Information Access Evaluation. Multilinguality, Multimodality, and Visualization* (pp. 282-302). Springer Berlin Heidelberg.M.

Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches.*International Journal of Computer Applications*, *68*(13), pp. 13-18.

Hiemstra, D., & De Vries, A. P. (2000). Relating the new language models of information retrieval to the traditional retrieval models.

Hoad, T. C., & Zobel, J. (2003). Methods for identifying versioned and plagiarized documents. *Journal of the American society for information science and technology*, *54*(3), pp.203-215.

Huang, A. (2008, April). Similarity measures for text document clustering. In*Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand* (pp. 49-56).

Johnson, R., & Zhang, T. (2014). Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*.

Jones, W. P., & Furnas, G. W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American society for information science*, *38*(6), 420-442.

Ljubešić, N., Boras, D., Bakarić, N., & Njavro, J. (2008, June). Comparing measures of semantic similarity. In *Information Technology Interfaces, 2008. ITI 2008. 30th International Conference on* (pp. 675-682). IEEE

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1, p. 496). Cambridge: Cambridge university press.

Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI* (Vol. 6, pp. 775-780).

Oakes, M. P. (2014). *Literary Detective Work on the Computer* (Vol. 12). John Benjamins Publishing Company.

Polettini, N. (2004). The vector space model in information retrieval-term weighting problem. *Entropy*, 1-9.

Ponte, J. M., & Croft, W. B. (1998, August). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 275-281). ACM.

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of documentation*, *33*(4), pp.294-304.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation*, *60*(5), 503-520.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), pp.513-523.

Singhal, A., Salton, G., Mitra, M., & Buckley, C. (1996A). Document length normalization. *Information Processing & Management*, *32*(5), 619-633.

Singhal, A., Buckley, C., & Mitra, M. (1996B). Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 21-29). ACM.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), pp.11-21.

Strehl, A., Ghosh, J., & Mooney, R. (2000, July). Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)* (pp. 58-64).

Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL.

Turney, P. & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*(1), pp.141-188.

White, R. W., & Jose, J. M. (2004, July). A study of topic similarity measures. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 520-521). ACM.

Zhang, J., & Korfhage, R. R. (1999). A distance and angle similarity measure method. *Journal of the American Society for Information Science*, *50*(9), pp. 772-778.