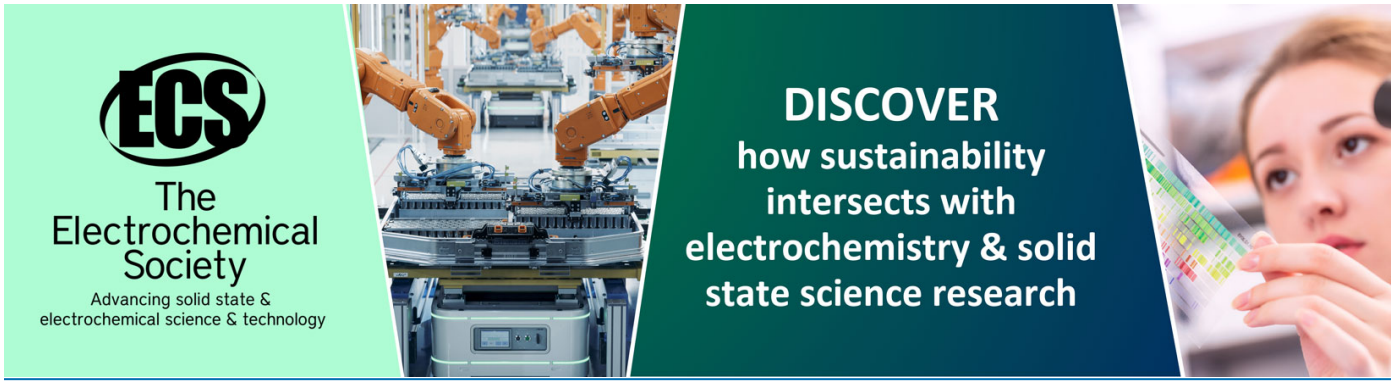
Hoax news validation using similarity algorithms



Hoax news validation using similarity algorithms

S Y Yuliani[1,2], S Sahib[2], M F Abdollah[2], Y S Wijaya, N H M Yusoff

[1]Information Security and Networking Research Group (InFORSNET), Faculty of Information Communication Technology, Universiti Teknikal Malaysia Melaka Malaka, Malaysia
[2]Informatic Department, Faculty Engineering, University Widyatama Bandung, Indonesia

Corresponding author: sy.yuliany@gmail.com

Abstract. News that is presented every day on social media dramatically affects the feelings, feelings, thoughts, or even actions of a person or group. Hoax News is one of them which is disturbing the public and raising noise in various fields, ranging from politics, culture, security, and order, to the economy. Inseparable from social media users. How every day, there is information on social media, which is not necessarily true so that people are provoked by hoax on social media. The news detection system in this study was designed using Unsupported Learning so that it does not require data training. The system was built using the Equation algorithm to calculate the validity of document similarity. Extraction results used to search for content related to user input using a detection engine, then the similarity value and the time needed to utilize hoax news are calculated. System validation testing by using a four text similarity algorithm called the Equation algorithm, the Levenshtein algorithm, the Smith-Waterman algorithm, the Damerau Levenshtein algorithm; this algorithm is used to find the best analytical solution of news hoaxes and submissions needed to find the news hoax password. The final results of the deception detection research using a script that has been done for Validation using an algorithm, get the value of accuracy in detection using the Smith-Waterman algorithm, which produces an accuracy value of text similarity of 99.29% and can be used a process of 6, 57 seconds, followed by the second sequence that is the similarity algorithm produces an accuracy of 75% and requires a processing time of 4.94 seconds, then the third sequence is the Levenshtein algorithm with an accuracy of 55.02% and requires a processing time of 5.49 seconds, and is used today is Damerau Levenshtein algorithm is 55.02% and requires a processing time of 7.54%. The results of research tests on this text can conclude the more text on the detection engine, the higher the verification value and the higher the time needed to process hoax news.

1. Introduction

Sharing information is a positive thing, but not all information disseminated through social media is in the form of facts. There have been various cases of spreading the news that is not facts or often called hoaxes. Whereas hoax is critical information that misleads human perception by spreading false information but considered as valid, No wonder then the intensity of fake news and hoax news on social media is so viral on social media [1]. For personal and group benefit by spreading harmful content that caused unrest and mutual suspicion in the community [2]. The ease and speed of dissemination on social media make this hoax news known to many people in a relatively short period, and can to more people.

can be detrimental to parties related to the hoax news if the public reads the news, trusts it, and builds an unfavorable image of the parties related to the hoax news

The development of a hoax detection engine is a tool to limit the spread of hoax news, making the lowest possible spread rate. Application development is to build enough Hoax news datasets to make it easy to make hoax spread analysis detection. Hoax news detection results are expected to be accurate because social media users share hoax news posts without varying the authenticity and accuracy of the news. A new approach is needed to calculate the accuracy of hoax news validation [3], to produce better hoax news detection

The development of a hoax detection machine is a risk mitigation tool. How to suppress the spread of false news make the lowest possible numbers. Application development is to build a Hoax news database. With the database of deception, that there is sufficient base in analyzing the spread of hoax. Hoax news detection results are expected to be accurate because social media users share hoax news posts without varying the authenticity and accuracy of the news. A new approach is needed to calculate the accuracy of hoax news validation [3], to produce better hoax news detection

The purpose of this study is to provide an accurate measurement of hoax news detection for news on social media. The proposed system extracts the contents of hoax news items, searches hoax news on the internet to find similar articles in reliable online news sources, then matches it by extracting hoax news content with news site content and generates a certain level of hoax news detection. Processing techniques such as the web, crawling techniques, summarizing techniques and the similarity of words by using a similarity algorithm. Finally, a hoax news validation system produces a score that can explain the accuracy of a news post based on a measure of the similarity of words calculated between social media news posts and online news article content. Processing time is also a parameter for measuring the level of hoax news validation detection

2. Related study

Arjun [4] proposes an automatic hoax validation technique that classifies hoax news into several classes, right, mostly accurate, half-real, almost untrue, false and shorts. With models based on Convolutional Neural Networks (CNN) and Bi-directional Long Short Term Memory (Bi-LSTM). Representations obtained from these two models in the Multi-layer Perceptron Model (MLP) the final classification

Ishak et al. [5] proposed a Validation of a text-based deception detection system using the Levenshtein Distance algorithm. Then identifying the potential deception of the email content by comparing it with a database of deception, the required component consists of 3 main components: pre-processing of text, detection of deception, and detection of a new deception. For the Pre-processing text stage, collect emails that will or validity as genuine or fraudulent emails.

Shu et al. [6] propose the Tri-Relationship hoax detection validation, called TriFN, with data objects taken from social media. This technique explores the correlation of publisher bias, news establishments, and relevant user involvement simultaneously, and proposes Tri-Relationship. Shu et al. provide two comprehensive real-world fake news datasets to facilitate fake news research.

Tacchini et al. [7], have an automated online detection hoax system as a contribution that is by showing that Facebook posts can be classification with high accuracy as a hoax or not a hoax based on users who like them. System presenting two classification techniques, one based on logistic regression, the other based on new adaptations crowdsourcing Boolean algorithm. The dataset consists of 15,500. Facebook posts and 909,236 users, the research results obtained classification accuracy exceeding 99% even when the training set contained less than 1% of the post. Shows the power of the technique for which they purposed the system worked even to users who like hoaxes and Fact posts these results indicate that the diffusion pattern mapping information can be a useful component for automatic hoax detection system.

Veronica et al. [8] make a framework for the automatic identification of fake content in online news. Moreover, it has two contributions. The First Aviad Elyashar et al., The problem with this Research is how to divide the dataset into topic categories and authenticity in online discussions, the process used to retrieve their post accounts to train traditional ML groupings, and manual labels for label accounts.

Aviad Elyashar et al., propose an approach for the detection of email hoax, identifying link prediction-based features that are found useful for account classification [9].

Yoke Yie Chen et al., The problem is that misleading information is always a distortion of draft growth. Some hoaxes made in such a way that they can be private data provided they required for official purposes, Yoke Yie Chen et al. proposed developing a hoax detection system by incorporating text matching method using Levenshtein Distance measure, The proposed model is used to identify text-based hoax emails. Sensitivity and specificity are used to evaluate the accuracy of the system in identifying hoax emails [10].

Sakeena et al. found a problem with spreading fake online news that has identified as one of the main concerns of online abuse. This Research proposed evaluation of the effectiveness of algorithm(s), able to detect and filter to reasonable degree of accuracy what constitutes an online fake news, multi-layered evaluations technique to be built as an app where all information read online is associated with a tag, given a description of the facts about the contain [11].

RST extracts news style-based features by combines the vector space model and rhetorical structure theory. The SVM classifier applied for classification [12]. Castillo, This method predicts news veracity using social engagements. The features are extracting from user-profiles and friendship networks. To ensure a fair comparison, we also add the credibility score of users inferred in Sec as an additional feature [13].

2.1. Similarity technique detection of hoax

Text similarity measures play an increasingly important role in text-related Research and applications in tasks such as information retrieval, text classification, document clustering, topic detection, topic tracking, questions generation, question answering, essay scoring, short answer scoring, machine translation, text summarization, and others [14]. Validation Hoax news detection by comparing the similarity of documents is the detection of the similarity of some hoax news that results in the accuracy of text similarity with much time needed to process words. One of the detections of the similarity of this document can be done with several techniques, for example, information search techniques, statistical calculation techniques, or by using syntactic information from a word for word. Below is an explanation of some text similarity detection algorithms used.

2.1.1 Similarity. This algorithm is calculating the length of the longest string. By calculating the similarity between two strings, similar_text calculate the similarity in percent, by dividing the results of similar_text by the average length of the given string multiplied by 100. This example shows that swapping the first and second argument may yield different results.

```php
<?php
$sim = similar_text('bafoobar', 'barfoo', $perc);
echo "similarity: $sim ($perc %)\n";
$sim = similar_text('barfoo', 'bafoobar', $perc);
echo "similarity: $sim ($perc %)\n";
```

The above example output something similar to:
similarity: 5 (71.428571428571 %)
similarity: 3 (42.857142857143 %)

2.1.2 Levenshtein Distance. Levenshtein Distance is a matrix to measure the number of differences between two strings, the distance between strings is measured based on the number of characters added, deleting characters or replacing the characters needed to convert the source string into a target [5], [10]. Definition of Mathematically, the Levenshtein distance between two strings, a and b (of length |a| and |b| respectively), is given by lev a,b(|a|,|b|) where:

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Here, 1(ai≠bi) is the indicator function equal to 0 when ai≠bi and equal to 1 otherwise, and leva, b(i,j) is the distance between the first I characters of a and the first j characters of b. Note first element minimum corresponds to deletion (from a to b), the second to insertion and the third to match or mismatch, depending on whether the respective symbols are the same. So it can be concluded that the higher the similarity value produced, the greater the accuracy generated from hoax news detection

2.1.3 Smith-Waterman. The Smith-Waterman algorithm is a well-known dynamic programming algorithm for performing local sequence [15]

1. Determine the substitution matrix and the gap penalty scheme.
   - s(a,b)- Similarity score of the elements that constituted the two sequences
   - W_k- The penalty of a gap that has length k

2. Construct a scoring matrix H and initialize its first row and first column. The size of the scoring matrix is (n+1)*(m+1). The 0-based indexing.

$$H_{k0} = H_{01} = 0 < k < n \text{ and } 0 < l < l < m$$

3. Fill the scoring matrix using the equation below

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + s(a_i, b_j), \\ \max_{k \geq 1}\{H_{i-k,j} - W_k\}, \\ \max_{l \geq 1}\{H_{i,j-l} - W_l\}, \\ 0 \end{cases} \qquad (1 \leq i \leq n, 1 \leq j \leq m)$$

where

$H_{i-1,\ j-1} + s(a_i, b_j)$ is the score of aligning a_i *and* b_j
$H_{i-k,\ j-1} - W_k$ is the score if a1 is the end of a gap of length *K*,
$H_{i,\ j-1} - W_l$ is the score if $b\,j$ is the end of a gap of length *l*,
0 Means there is no similarity up to a_i and b_j

4. Traceback. Starting at the highest score in the scoring matrix *H* and ending at a matrix cell that has a score of 0, traceback based on the source of each score recursively to generate the best local alignment.

2.1.4 Damerau Levenshtein. The Damerau–Levenshtein distance is a string metric for measuring the edit distance between two sequences. Informally, the Damerau–Levenshtein distance between two words is the minimum number of operations yang consisting of insertions, deletions or substitutions of a single character, or transposition of two adjacent characters, this required to change one word into the other. The Damerau–Levenshtein distance differs from the classical Levenshtein distance by including transpositions among its allowable operations in addition to the three classical single-character edit operations deletions, insertions, and substitutions [16]

Damerau stated that more than 80% of all human misspellings could be express by a single error of one of the four types [17]. Damerau' sDamerau's paper considered only misspellings that could be correct with at most one edit operation while the original motivation was to measure the distance between human misspellings to improve applications such as spell checkers.

2.2. Methods in validation hoax news detection

Validation to detect hoax news in this Research using a similarity method [18]. Most existing approaches consider the hoax news problem as a classification problem that predicts whether a news article is a hoax

or not [19]. Problems encountered in text mining are large amounts of data, high dimensions, data and structures, constantly changing data, and data noise [20]. Overcoming unstructured data we need to evaluate the accuracy of words. A similarity algorithm is used in this evaluation to detect hoax news [21]. The similarity algorithm calculates the similarity between words; the complexity of this algorithm is the length of the longest word. The parameter used is. The first word. The second word. The third is to calculate the similarity in percent. The similar text calculates the similarity in percent, by dividing the results of similar text by the average length of the given word. Finding the longest general substring first, and then doing this for the prefix and suffix, recursively calculate the number of matching characters, below are the stages method of the hoax news detection validation :

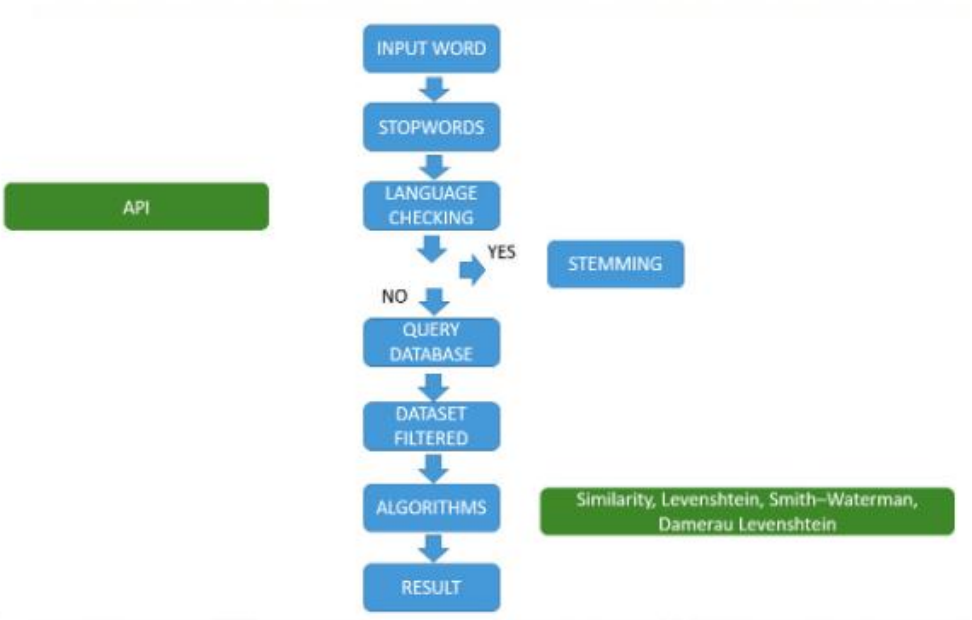


Figure 1. Methods in hoax news detection.

Stages of Validation of hoax news similarity detection:

1. Word Input: Input by comparing some of the words contained in the hoax news dataset contained in the hoax detection engine. Document reading: reading text documents
2. Stop word: Optimization by removing all words that classified as stop words. Stop words are common words that usually appear in large numbers and are considered to have no meaning. Stop words are generally used in information retrieval, stop words for English include, among others, the, while for Indonesian, among others are, at, too.
3. Language checking: Check the language using Multi-Language using an application that is already available in the API
4. Stemming: return various kinds of word formations to the representation of essential words, stemming method requires input in the form of words contained in a document, by producing output in the form of essential words. Basic word search for words that influence Indonesian
5. Database Query: Search for words contained in the hoax news database
6. Algorithm: The selection is carried out one by one testing of four algorithms namely similarity algorithm, Levenshtein algorithm, Smith-Waterman algorithm, Damerau Levenshtein algorithm
7. Result: Shown Results of the similarity of words in percent and the amount of time needed to process in seconds.

2.3. Tool in validation hoax news detection

The tool used in this study is a separate computer. The following hardware specifications used in this study are as follows: PC, Intel Xeon Scalable (Skylake) 2.0 GHz processor, vCPU, 3.75 GB memory. The software used in this study is as follows, Java Script, XAMPP 5.6.12, Sqlite Browser PHP, Maria DB. The dataset used was 9814 hoax news collected by crawling from ten hoax news fact-checking websites,

3. Testing and validation

In testing and Validation of hoax news detection, five hoax news documents compared. The five documents have different word lengths. The documents used are taken from the hoax news dataset. This document is named W1, W2, W3, W4, and W5. The W1 document contains <=5 news hoax words, the W2 document contains <=10 news hoax words, the W3 document contains <=50 news hoax words, the W4 document contains <=100, and the W5 document contains <=150 words. Each document is tested one by one using an algorithm, and it is how long it takes to detect hoax news, for each document W1, W2, W3, W4, and W5. Examples of data tested presented in table 1:

Table 1. Validation of hoax news detection

Algorithms	Accuracy (%)					Time (Sec)				
	W1	W2	W3	W4	W5	W1	W2	W3	W4	W5
Similarity	6,93	12,8	31,9	75.00	70,40	2.44	2.47	2.87	4.95	5.09
Levenshtein	3,42	6,84	18.97	40.92	17.82	2.46	2.55	3.04	5.49	10.71
Smith-Waterman	100	100	100	99.29	100	2.59	2.84	3.71	6.57	12.88
Damerau Levenshtein	3.42	6.84	18.80	55.02	54.61	2.58	2.86	3.99	7.45	19.99

The results of testing and Validation using a hoax detection engine, the results of testing by comparing four algorithms namely similarity, smith waterman, Levenshtein, and demerau Levenshtein, resulting in the percentage accuracy of text similarity and the amount of time needed to process hoax news detection. The results of the accuracy of text similarity that has the highest results found in the smith-waterman algorithm that is 100% apply to W1, W2, W3, W5, and 99.29% for W4, while the time required to detect hoax news is 2.59 seconds to 12.88 seconds. The test results with the lowest accuracy results are in the Levenshtein algorithm, which is 3.42% to 40.92%, and the time required to detect hoax news is 2.46 to 10.71. We can see the results of the Validation of text similarity in figure 2.

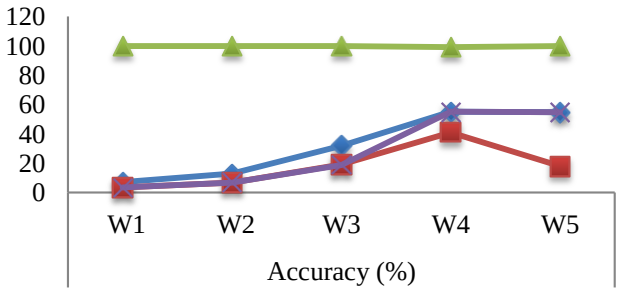


Figure 2. Accuracy of hoax news detection

The test results for accuracy can look in the picture above that the more words that are input on the detection engine, the accuracy level of text similarity is higher but the time needed to make the hoax news detection process becomes a long time, this can be seen in figure 3.

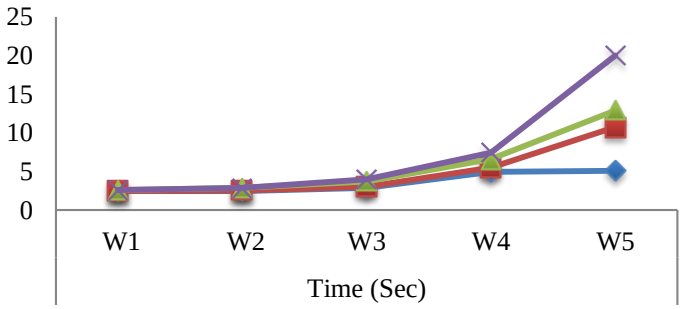


Figure 3. Processing time of hoax news detection

4. Conclusion

Measurement of accuracy and Validation of text similarity in hoax news detection by comparing four similarity algorithms, smith waterman, Levenshtein, and demerau Levenshtein, were made to assist in checking the similarity and news of hoaxes. Determination of the similarity of hoax news based on structural similarity in a word and sentence. After testing in this study, it can conclude that the hoax news detection system using the smith waterman algorithm produces a more accurate similarity value than the three liane algorithms and detects some hoax news with a percentage of 99.29%, and can use a process of 6.57 seconds. Validation testing using the Smith-Waterman Algorithm produces pretty good accurate results, but the lack of this algorithm is that the more words entered to produce the more significant the accuracy value and the more time needed to detect hoax news. Future work that can be used to develop this Research is that the synonym database can be productive by taking the Indonesian language thesaurus. It is necessary to develop a system that is not only able to detect documents containing text, but also contains graphics, tables and images

References

[1] S Zannettou, M Sirivianos, J Blackburn, and N Kourtellis *The Web of False Information : Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans* pp. 1–26
[2] S Y Yuliani, S Sahib, M F Abdollah, M N Al-mhiqani and A R Atmadja 2018 *Review Study of Hoax Email Characteristic* vol. 7 pp. 778–782
[3] R Chandrathlake, L Ranathunga, S Wijethunge and P Wijerathne 2018 *3rd Int. Conf. Inf. Technol. Res.* pp. 1–6
[4] A Roy, K Basak, A Ekbal and P Bhattacharyya 2017 *A Deep Ensemble Framework for Fake News Detection and Classification*
[5] A Ishak, Y Y Chen and S Yong 2012 *Distance-based Hoax Detection System* pp. 215–220
[6] K Shu, A Sliva, S Wang, J Tang and H Liu 2016 *Fake News Detection on Social Media : A Data Mining Perspective* vol. 19 no. 1 pp. 22–36
[7] E Tacchini, G Ballarin, M L Della Vedova, S Moret and L de Alfaro 2017 *Some Like it Hoax: Automated Fake News Detection in Social Networks*
[8] B Kleinberg, A Lefevre and R Mihalcea 2017 *Automatic Detection of Fake News*, no. August
[9] A Elyashar, J Bendahan and R Puzis *Is the News Deceptive? Fake News Detection using Topic Authenticity*
[10] Y Y Chen, S Yong and A Ishak 2014 *Email Hoax Detection System Using Levenshtein Distance Method* vol. 9 no. 2 pp. 441–446
[11] S M Sirajudeen, N F A Azmi and A I Abubakar 2107 *J. Theor. Appl. Inf. Technol*
[12] V L Rubin, Y Chen and N J Conroy 2015 *Proc. Assoc. Inf. Sci. Technol.* vol. 52 no. 1 pp. 1–4
[13] L Akoglu and C Faloutsos 2012 *Opinion Fraud Detection in Online Reviews by Network Effects*, pp. 2–11
[14] W H Gomaa 2013 *A Survey of Text Similarity Approaches* vol. 68 no. 13 pp. 13–18

[15] R W Irving *Plagiarism and Collusion Detection using the Smith-Waterman Algorithm*, pp. 1–22
[16] G V Bard 2005 *Spelling-Error Tolerant, Order-Independent Pass-Phrases via the Damerau-Levenshtein String-Edit Distance Metric*
[17] F J Damerau, I B M Corporation and Y Heights 1964 *A Technique for Computer Detection and Correction of Spelling Errors* no. 3 pp. 171–176
[18] Z Su, B Ahn, K Eom, M Kang, J Kim and M Kim 2008 *Plagiarism Detection Using the Levenshtein Distance and Smith-Waterman Algorithm* pp. 0–3
[19] K Shu, A Sliva, S Wang, J Tang and H Liu 2016 *Fake News Detection on Social Media : A Data Mining Perspective* no. i
[20] A Huang 2008 *Similarity Measures for Text Document Clustering* no. April
[21] M Vuković, K Pripužić and H Belani 2009 *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* vol. 5711 LNAI no. PART 1 pp. 318–325