Jurnal Teknik Informatika (JUTIF)
Vol. 3, No. 4, Agustus 2022, hlm. 975-982

DOI: https://doi.org/10.20884/1.jutif.2022.3.4.152
p-ISSN: 2723-3863
e-ISSN: 2723-3871

JARO WINKLER ALGORITHM FOR MEASURING SIMILARITY ONLINE NEWS

Teguh Efriyanto[1], Mardhiya Hayaty*[2]

[1,2]Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta, Indonesia
Email: [1]teguh.efriyanto@students.amikom.ac.id, [2]mardhiya_hayati@amikom.ac.id

(Naskah masuk: 8 Februari 2022, Revisi: 4 April 2022, diterbitkan: 20 Agustus 2022)

Abstract

Online news is a source of information for people; this impacts journalists as news writers who can find news information quickly and accurately every day. Journalists can plagiarise other journalists or take news material from other news media sites and use it to publish in the media without including the source. An algorithm is needed to measure the similarity of online news. This work proposed the Jaro Winkler algorithm, with the value obtained from the calculation normalised so that the value 0 means there is no resemblance, and one means it has the exact resemblance. The data used is 20 online news media sites in the Central Kalimantan area. The Scraping process utilised the Custome Search JSON API and used keywords to get the news on the same topic. The results of the calculation of news similarity with the Jaro Winkler algorithm obtained an average value of online news similarity of 74.49%, with 43 news data with severe plagiarism levels and 12 news data with moderate plagiarism levels. There are weaknesses in the Jaro Winkler algorithm in calculating the similarity value in the data obtained. Some undetected data should have a heavy plagiarism level but not severe and vice versa.

Keywords: *jaro winkler, online news, plagiarism, similarity, text preprocessing*

ALGORITMA JARO WINKLER UNTUK MENGUKUR SIMILARITY BERITA ONLINE

Abstrak

Berita online adalah sumber informasi bagi masyarakat; Hal ini berdampak pada jurnalis sebagai penulis berita yang dapat menemukan informasi berita dengan cepat dan akurat setiap hari. Jurnalis dapat menjiplak jurnalis lain atau mengambil materi berita dari situs media berita lain dan menggunakannya untuk dipublikasikan di media tanpa mencantumkan sumbernya. Diperlukan suatu algoritma untuk mengukur kesamaan berita online. Pada penelitian ini mengusulkan algoritma *Jaro Winkler*, dengan nilai yang didapat dari perhitungan dinormalisasi sehingga nilainya 0 artinya tidak ada kemiripan dan 1 artinya memiliki kemiripan yang sama persis. Data yang digunakan adalah data isi berita dari 20 situs media berita online daerah Kalimantan Tengah, yang diperoleh dengan proses *Scraping* yang disorting dengan *Google Custome Search Engine* dengan memanfaatkan *Custome Search JSON API* dan menggunakan *keyword* untuk mendapatkan berita dengan topik berita yang sama dan kemudian dilakukan proses *text preprocessing*. Hasil perhitungan *similarity* berita dengan algoritma *Jaro Winkler* dari 55 data berita yang diperoleh menghasilkan nilai rata-rata *similarity* berita online sebesar 74,49% dengan 43 data berita tingkat plagiarisme berat dan 12 data berita dengan tingkat plagiarisme sedang dan terdapat kelemahan pada algoritma *Jaro Winkler* dalam menghitung nilai *similarity* pada data yang diperoleh, yang mana terdapat beberapa data yang tidak terdeteksi yang seharunya tingkat plagiarismenya berat namun tidak berat dan sebaliknya.

Kata kunci: *berita online, jaro winkler, plagiarisme, similarity, teks preprocessing*

1. INTRODUCTION

The current development in this industrial era 4.0 has made online news media as an important thing in daily lives to find sources of information. This has an impact on online news media journalists to find out news information quickly and accurately, it is possible for journalists who work directly together to commit plagiarism to other journalists or take news material from other news media sites and use it to be published in their media without citing the source [1].

This of course violates the existing regulations because of plagiarism, plagiarism problem itself can be detected with several algorithms, such as in previous studies that used the Winnowing algorithm

to detect title's plagiarism of a student's final project, with 11 test data obtained an average similarity of 66 ,30%, but without going through the text preprocessing process and only detecting the title of the final project [2]. In the research that uses the Consine Similarity algorithms and TF-IDF, it is able to detect the similarity of abstract documents but the processing time is longer when using stemming and similarity results from documents that have high similarities, the similarity is only 0.56 with stemming and 0.46 without stemming and the average number of words only 127-157 words [3]. In a study that compared the Jaro Winkler algorithm and Latent Sematic Analysis for the detection of plagiarism in abstract documents with 5 test data, the average similarity of the Jaro Winkler algorithm was 75.7% better than the Latent Sematic Analysis algorithm and if the data compared were exactly the same, Jaro Winkler's algorithm will produce a plagiarism value of 100%, while the Latent Semantic Analysis method produces a plagiarism value of 97.14% [4].

Therefore, in this research, the Jaro Winkler algorithm is proposed to measure similarity with online news data which has also been conducted by other researchers with different data [5]-[13].

2. RELATED WORK

Various approaches to calculating similarity in the text have been conducted by many researchers, such as the researcher [14] using the Text Extraction technique and Jaro Winkler, to monitor of prices by filtering commodity price data from sources twitter.com, detik.com and liputan6.com that suitable with the keyword and text preprocessing was carried out, Jaro Winkler's accuracy for detecting commodity prices in text was 75%. The plagiarism checker in the final project data management system performed by [15] using the Jaro Winkler algorithm resulted in 40% accuracy, 33.3% precision, 100% recall and 35% F-Measure.

Similarity measurements for detection of duplicate database records have been conducted by [16] using 7 different algorithms, and produce a better Jaro Winkler algorithm with 95% accuracy, but testing without text preprocessing, transposition in Jaro Winkler allows reducing computational complexity, saving time and increasing accuracy Jaro Winkler. Detection of similarity of URL text to find out which one is more relevant to the input query to conduct content mining is carried out by [17] using 5 different algorithms, resulting in Jaro Winkler's algorithm being better with a precision of 0.90 and a recall of 0.87.

Detection of plagiarism in thesis documents is conducted by [18] using the Single Link Clustering method and Jaro Winkler, where the clustering stage aims to categorize related titles, from the test obtained good results with an average precision of 84.37% and recall of 84.37%. Comparison of string search algorithms in the string matching method carried out by [19] identify typing errors by comparing 4 algorithms produces the best Jaro Winkler algorithm with a MAP value of 0.87.

Prediction of real-time activity of automatic data flow segmentation is carried out by [20] using Jaro Winkler to make comparisons between training datasets and data streams resulting in Jaro Winkler algorithm accuracy of 76%. The auto-correction system for essay scoring conducted by [21] using the Jaro Winkler algorithm with a word library consisting of 97 synonym words and 204 function words resulting in an average accuracy of 85.246%.

Text matching in computer voice recognition for reading the Al-Qur'an which was carried out [22] using the Jaro Winkler algorithm resulted in good precision in text matching, with an accuracy of 91%. The grouping of online store products based on product names is carried out [23] to make it easier to find out products on many E-Commerce sites, using the Jaro Winkler algorithm by categorizing 5 products on different sites, from 20 trials the Jaro Winkler algorithm accuracy results are 81.27%.

2.1. Plagiarism

Plagiarism is an act of misuse, theft, publishing, or a statement that tells that a work is a thought, idea, writing, or creation of one's own which actually belongs to someone else [24].

Plagiarism can be considered as a criminal offense because of stealing someone else's copyright. In journalism, perpetrators of plagiarism will be punished based on Article 2 of the Journalistic Code of Ethics, plagiarism among journalists called cloning journalism [1].

In the case of plagiarism, there are several classifications of plagiarism that can be used as a reference. The classification can be calculated based on the proportion or percentage of hijacked words, sentences, paragraphs [24] :

- Low Plagiarism : <30%
- Mediun Plagiarism : 30-70%
- High Plagiarism : >70%

2.2. Text Preprocessing

Text Preprocessing is the first stage in text processing. This stage includes all routines, and processes for preparing text into data that is ready to be processed in the knowledge discovery operation of the text processing system [25]. Text preprocessing in this study conducted through 6 processes, namely:

A. Remove HTML Tag, the stage of removing the HTML tag and only taking the contents of the news text,

B. Remove Special Caracter, step of removing journalists' names, news portal website links' names that existed in the news text,
C. Case Folding, step of changing capital letters to lowercase and removing punctuations,
D. Number Removal, step of deleting the assumptions in the news text,
E. Filtering, step of removing conjunctions or stopword removal, Stopwords are words that are not unique in the document. Examples are "oleh", "pada", "di", "karena", "sebuah", etc. Stopwords will be removed to obtain more structured text data [26], and
F. Stemming, step of changing words into basic words or removing affixes, using Sastrawi Stemmer from Nazief-Andriani's algorithm [27].

2.3. Jaro Winkler Algorithm

Jaro Winkler is an algorithm for measuring the similarity between two strings which is a variant of the Jaro distance metric, and this algorithm is usually used in duplicate detection [28]. The higher the Jaro Winkler value on two strings, the more similar the strings are. The value 1 indicates the similarity between the strings and the value 0 indicates the inequality between the strings [9].

The Jaro-Winkler Distance algorithm has three basic sections [29], namely:

1. Calculate the length of the string.
2. Specifies the same character of both strings.
3. Determine the number of transpositions.

The Jaro-Winkler Distance algorithm uses a formula to calculate the distance (d_j) between the two strings, they are s_1 and s_2, which is shown by equation 1.

$$d_j = \frac{1}{3} x \left(\frac{m}{s_1} + \frac{m}{s_2} + \frac{m-t}{m}\right) \quad (1)$$

Notes:

m : The same number of characters in both strings
s_1 : Length of string 1
s_2 : Length of string 2
t : Number of transposition

Two characters having the same theoretical distance that can be justified if they do not exceed the equation 2.

$$\left(\frac{max(s_1,s_2)}{2}\right) - 1 \quad (2)$$

However, if it refers to the value that will be produced by the Jaro Winkler algorithm, then 1 is the maximum distance value which indicates the similarity of the strings compared to reaching one hundred percent or exactly equal. Prefix scale (p), used in Jaro Winkler algorithm, with the standard value for constant according to Winkler is p = 0,1. For prefix length (l), is known the length of the same character until an inequality is found (maximum 4 characters), the calculation for the Jaro-Winkler Distance value d_w can be shown by equation 3.

$$d_w = d_j + \left(lp(1 - d_j)\right) \quad (3)$$

Notes:

d_w : Value of jaro-winkler distance,
d_j : Value of jaro distance for untuk string s_1 and s_2 value,
l : The length of the common prefix at the beginning of the string with a maximum value is 4 characters,
p : Scaling factor constant (according to Winkler, the standard value for this constant is p = 0,1).

3. METHOD

The following are the steps of the research method carried out in this research :

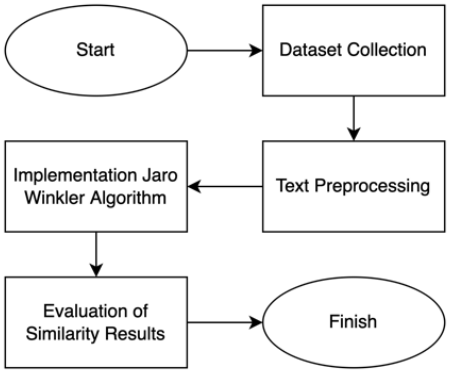

Figure 1. Research Steps.

3.1. Dataset

The dataset used in this study is news text data from 20 online news sites in Central Kalimantan. Where the dataset is obtained from the test data input process and from this process the news keyword data contained in the test data will be obtained, then the keywords are used to perform a Custom Search JSON API by retrieving the top 4 news data in every 20 news data sorted by Google Programmable Search Engine, the keyword itself is used to obtain the same news topic as the test data. And the dataset in this study is dynamic so that the data obtained can change.

3.2. Text Preprocessing

Text Preprocessing carried out on s_1 and s_2 data before algorithm Jaro Winkler calculation process is carried out with the following steps Remove HTML

Tag, Remove Special Caracter, Case Folding, Number Removal, Filtering, dan Stemming.

3.3. Implementation Jaro Winkler Algorithm

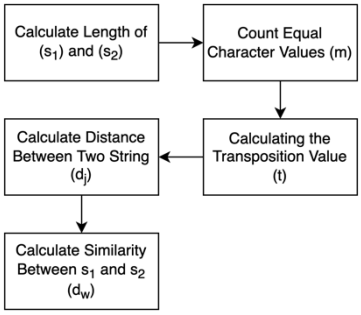

Figure 2. Steps of Algoritma Jaro Winkler

The implementation phase of the Jaro Winkler algorithm is carried out when the data has passed the text preprocessing process and is ready to be processed. after the data is ready to be processed, the next step is to implement the Jaro Winkler algorithm with the initial stages of calculating the length of the string, calculating the value of the same character, calculating the transposition, calculating the distance between two strings and calculating the similarity value between s_1 and s_2. The implementation process is carried out using the PHP programming language with stages as shown in Figure 2.

3.4. Evalution Similarity Results

Online news in the plagiarism category or not is measured by the document similarity value. The values is light, medium or heavy categories

4. RESULT AND DISCUSSION

4.1. Dataset

In this study, the testing process is carried out by inputting the test data news Url and taking keywords from the test data news url. From the test data url https://www.klikkalteng.id/baca/2021/10/15/88390/kepala-pdam-kotim-bantah-kenaikan-tarif-hingga-100-persen obtained the keyword *Pdam Tirta Mentaya Tarif*, and the data obtained online news with the same news topic as many as 55 news data from 20 regional news portals in Central Borneo.

4.2. Text Preprocessing

From the 55 data, the text preprocessing process was carried out with the several steps Remove HTML Tag, Remove Special Caracter, Case Folding, Number Removal, Filtering, dan Stemming. An example of the results of the text preprocessing process on the data data is shown in Table I.

Table 1. Result Of Text Preprocessing Data 18

Klikkalteng	Beritakalteng
kepala usaha daerah air minum pdam tirta mentaya kabupaten kotawaringin timur kotim firdaus herman ranggan ban naik tarif persen isu edar masyarakat atas dasar naik sampai persen firdaus jumat ungkap susul banyak isu edar masyarakat tarif pdam capai persen susul laku tarif langgan pdam tirta mentaya kotim oktober naik tarif timbul polemik masyarakat langgan pdam kendati masyarakat mudah hasut isu edar ecek benar isu contoh isu naik tarif persen turut sebab sangkut air kapasitas dampak naik…*dst*	dewan wakil rakyat daerah kabupaten kotawaringin timur kotim kelar rapat dengar dapat rdp bapemperda komisi iv perintah daerah usaha daerah air minum pdam tirta mentaya dewan awas pdam rdp kait naik tarif air langgan keluh masyarakat selasa gelar rapat dengar dapat perintah daerah pdam rekomendasi pdam tirta mentaya tinjau sesuai tarif keluh masyarakat ketua badan bentuk atur daerah bapemperda dprd kotim handoyo j wibowo selasa salah poin hasil rapat dengar dapat tinjau naik tarif…*dst*

4.3. Implementation Jaro Winkler Algorithm

Calculation of the similarity of the data 18 with the Jaro Winkler algorithm with data s_1 being the news data of Klikkalteng and s_2 being the news data of the news of Beritakalteng, the parameter values obtained are as follows :

$$s_1 = 1492$$
$$s_2 = 1642$$
$$m = 1441$$
$$t = 650.0$$

$$d_j = \frac{1}{3}\ x\ \left(\frac{1441}{1492} + \frac{1441}{1642} + \frac{1441 - 650}{1441}\right)$$
$$d_j = \frac{1}{3}\ x\ 2{,}3923303594$$
$$d_j = \ 0.7974434531$$

The result of the distance value d_j, is used to calculate similarity with the formula d_w.

$$d_j = \ 0.7974434531$$
$$l = 0$$
$$p = 0{,}1$$

$$d_w = 0.7974434531 + (0x0{,}1(1 - 0.7974434531))$$
$$d_w = 0.7974434531 + 0$$
$$d_w = 0.7974434531$$

The results of the calculation of Jaro Winkler Distance d_w with Klikkalteng s_1 news data and s_2 BeritaKalteng news obtained similarity of 0.7974434531 x 100 = 79,74434531 or 80% rounded up with the *round php* function.

4.4. Evaluation Similarity Results

The tests conducted in this research calculate the similarity of news s_1 with s_2, where the s_2 news data is news data obtained from the test data input process and 55 news data is obtained, from 55 news data obtained, the similarity calculation process with s_1 test data is carried out, and the results of the similarity of the 55 news data are shown in Table 2.

Table 2. Similarity Result

No	News	Similarity
1	Mulai 1 Oktober, PDAM Tirta Mentaya Kotim Berlakukan Tarif Baru	79%
2	Menaikan Tarif Pengamat Ekonomi Tantang Pdam Tirta Mentaya ...	67%
3	Kepala PDAM Kotim Bantah Kenaikan Tarif Hingga 100 Persen	100%
4	Warga Tarif Pdam Kok Naik Padahal Sumbernya Dari Sungai Ini ...	77%
5	Ini Komentar Direktur PDAM Dharma Tirta Mentaya Sampit Terkait ...	63%
6	Warga Keluhkan Penaikan Tarif PDAM di Sampit	63%
7	Pelanggan PDAM Kotim Keluhkan Kenaikan Tarif di Tengah Pandemi	73%
8	Kenaikan Tarif PDAM Dinilai Wajar	73%
9	Kepala PDAM: Sedih Kalau Ada yang Mengaku Tak Mampu Bayar ...	69%
10	Penjelasan Bupati Kotim di Balik Naiknya Tarif PDAM Sampit ...	73%
11	Siap-Siap, Tarif PDAM Bakal Naik untuk Pelanggan Tertentu ...	69%
12	Penyertaan Modal Bisa Jadi Solusi Kenaikan Tarif PDAM Sampit ...	70%
13	Harga Produksi Air Bersih Tinggi, Tarif PDAM di Sampit Naik	66%
14	Lakukan RDP, Dewan Minta Pemerintah Tinjau Kembali Kenaikan ...	72%
15	Lakukan RDP, Dewan Minta Pemerintah Tinjau Kembali Kenaikan ...	76%
16	Mantan Ketua DPRD Kotim Sebut Kenaikan Tarif PDAM Tidak Tepat…	78%
17	Ekonomi Lesu Jadi Pertimbangan Kenaikan Tarif PDAM ...	76%
18	Rencana Kenaikan Tarif PDAM Jadi Perhatian Serius Para Wakil ...	80%
19	Dewan Bakal RDP tentang Kenaikan Tarif Air 100 Persen ...	78%
20	Momen Menaikan Tarif PDAM Dinilai Kurang Tepat – BeritaKalteng ...	80%
21	Kenaikan tarif air PDAM di tengah pandemi COVID-19 dikeluhkan ...	73%
22	DPRD Kotim sarankan PDAM tinjau kenaikan tarif - ANTARA News ...	76%
23	Legislator Kotim minta PDAM pertimbangkan kondisi ekonomi ...	80%
24	PDAM Tirta Mentaya Sampit ditarget harus sudah untung 2020 ...	79%
25	Sempat Minta Ditunda Komisi IV Sayangkan Kenaikan Tarif PDAM ...	77%
26	PDAM Diminta Harus Pastikan Pasokan Air Bersih dan Lancar ...	75%
27	NAH..!! Legislator Siap ”Kuliti” Kejanggalan Tarif Baru PDAM Sampit ...	79%
28	Tarif Janggal PDAM, Harusnya Belum Naik, Warga Sudah Bayar ...	77%
29	Begini Penjelasan Bupati Kotim di Balik Naiknya Tarif PDAM Sampit	82%
30	Legislator Apresiasi Langkah Pemkab Kotim Gratiskan Tagihan PDAM	73%
31	Pembangunan Instalasi Pengelolaan Air PDAM Tirta Barito Segera ...	74%
32	Revisi Perbup Kenaikan Tarif Air PDAM akan Dipelajari Lebih Lanjut	64%
33	Batalkan Kenaikan Tarif Air PDAM - Kalteng.co	80%
34	Minta Kenaikan Tarif Air PDAM Ditinjau Lagi - Kalteng.co	82%
35	Terkait Kenaikan Tarif Air 100 Persen, Dewan RDP dengan PDAM ...	77%
36	Dewan Minta Kenaikan Tarif PDAM di Kotim Ditinjau Ulang – Ini..	73%
37	Kenaikan Tarif PDAM Agar Perusahaan Tidak Bangkrut - Ini Kalteng	71%
38	Dirut PDAM Tirta Mentaya Sudah Saatnya Diganti - Ini Kalteng	78%
39	Fraksi Golkar Sarankan Pemkab Kotim Perhatikan Hal Ini Terkait…	73%
40	Tunda Dulu Kenaikan Tarif PDAM - KaltengOnline.com DPRD ...	79%
41	Rencana PDAM Menaikkan Tarif kepada Pelanggan Harus ...	66%
42	Bupati Siap Revisi Kenaikan Tarif PDAM – 101KPFM ...	73%
43	Hari Ini Dewan RDP dengan PDAM – 101KPFM PALANGKARAYA	78%
44	Soal Kenaikan Tarif PDAM Tirta Mentaya Sampit, Dewan ...	73%
45	Dewan Minta PDAM Tirta Mentaya Sampit Sosialisasikan Tarif Baru ...	75%
46	Usai di RDP, Ini Harapan Mantan Anggota Dewan Terkait PDAM ...	71%
47	DPRD Kotim Minta Review Perbup No 19 Tahun 2021 | Berita Sampit	71%
48	Pengusaha Ternak di Gumas Belum Minari RPH, Padahal Tarifnya…	80%
49	PDAM Sampit akan Memberlakukan Penyesuaian Tarif Kenaikan ...	73%
50	Wakil Ketua Komisi IV DPRD Kotim: Kebijakan Menaikan Tarif ...	61%
51	H Ary Dewar: Sarankan PDAM Tingkatkan Koordinasi dengan ...	68%
52	DPRD Kotim sarankan PDAM tinjau kenaikan tarif	76%
53	*Anggota dprd kotim minta pdam pertimbangkan kondisi ekonomi ...*	80%
54	PDAM disarankan tingkatkan koordinasi dengan DPRD Kotim	64%
55	Legislator Kotim Berharap Fungsi Sosial PDAM Tetap Diutamakan..	84%
	Average	74,49%

In the results of 55 data, the average value of online news similarity is 74.49%, and 43 data with a high plagiarism level above 70% are obtained, and 12 data with a moderate plagiarism level between 30% - 70%, and there are 100% similarity results found in data 3 this is because the test data and data 3 are the same news data, this certainly shows that the Jaro Winkler algorithm can detect text with the actual same level of similarity without errors as in the study [4]. However, from the results of the news similarity, there are inaccurate results which are found in data 31, 38 and 48, where the news topics obtained are different with the topics of test data s_1, but obtain similarity values of 74%, 78% and 80% or high. An example of the data 31 with the results of text preprocessing is shown in Table 3.

Table 3. Result Of Text Preprocessing Data 31

Klikkalteng	Kaltengtoday
kepala usaha daerah air minum pdam tirta mentaya kabupaten kotawaringin timur kotim firdaus herman ranggan ban naik tarif persen isu edar masyarakat atas dasar naik sampai persen firdaus jumat ungkap susul banyak isu edar masyarakat tarif pdam capai persen susul laku tarif langgan pdam tirta mentaya kotim oktober naik tarif timbul polemik masyarakat langgan pdam kendati masyarakat mudah hasut isu edar ecek benar isu contoh isu naik tarif persen turut sebab sangkut air kapasitas dampak naik…*dst*	kalteng today buntok bangun instalasi kelola air water treatment plant usaha daerah air minum pdam tirta barito kabupaten barito selatan selesai program sumber dana pusat dana alokasi khusus dak tahap kerja direktur pdam tirta barito rona cipta sambut datang sekretaris daerah sekda barsel eddy purwanto kunjung tinjau langsung bangun wtp selasa tahap bangun wtp selesai kerja rona proyek tahap bangun tahap proyek laksana dinas kerja rumah rakyat dpupr bagai laksana lapang bagai terima manfaat…*dst*

In Table 3, Klikklateng news and Kaltengtoday news have different news topics, where the Kaltengtoday news has the topic of Pdam Tirta Barito but has a similarity level of 74% or high. This is because the number of transpositions in data 31 is 406.0 so that it is possible to increase similarity because letters that are swapped in position will be counted [16], and the prefix length in the news data of Klikkalteng and Beritakalteng has the number of prefix length = 1 because there are the same letters at the beginning of the sentence, namely in data S1 it has the letter ***k*** at the beginning of the sentence in the word ***k****alteng* and in data S2 also has the letter ***k*** at the beginning of the sentence in the word ***k****epala*, This of course can increase the similarity results from the Jaro Winkler algorithm because the same letters at the beginning of the sentence will be counted [21], [13], [17].

After that, there are also medium similarity results, but have the same news topic as the s_1 test data contained in the data 13 with 66% similarity. The results of the text preprocessing text from the data 13 are shown in Table 4.

Table 4. Result Of Text Preprocessing Data 31

Klikkalteng	Matakalteng
kepala usaha daerah air minum pdam tirta mentaya kabupaten kotawaringin timur kotim firdaus herman ranggan ban naik tarif persen isu edar masyarakat atas dasar naik sampai persen firdaus jumat ungkap susul banyak isu edar masyarakat tarif pdam capai persen susul laku tarif langgan pdam tirta mentaya kotim oktober naik tarif timbul polemik masyarakat langgan pdam kendati masyarakat mudah hasut isu edar ecek benar isu contoh isu naik tarif persen turut sebab sangkut air kapasitas dampak naik…*dst*	naik tarif rencana usaha daerah air minum pdam tirta mentaya oktober pengaruh tinggi harga produksi air bersih direktur pdam tirta mentaya firdaus herman ranggan tarif kait produksi air bersih naik sesuai harga bbm listrik bahan kimia mati bakteri air harga atas tarif tahan harga sulit adakan harga sesuai senin september tarif sesuai layan masyarakat jalan lantar biaya produksi air salur masyarakat asal dapat perintah pns tunjang perintah biaya asal dapat sesuai firdaus naik tarif semenjak tarif naik tarif bilang rendah banding…*dst*

In Table 4, the news of Klikkalteng and Matakalteng have the same news topic but only has a similarity level of 66%. This is because the string length between the two data is different where the Klikkalteng data has a string length of 1492 and the Matakalteng data has a string length of only 647, where the Jaro Winkler algorithm states that the two strings compared are considered equal if the distance between the two strings is not more than the theoretical value distance [21], [17]. So that with very different string lengths it is possible to reduce the similarity level of the Jaro Winkler algorithm. And the Jaro Winkler algorithm does not pay attention to the meaning or synonyms of words so that the similarity of the algorithm can decrease if there are synonyms in both data [21], [15]. Such as in the Matakalteng data where there is the sentence *"biaya"* which is a synonym for the word *"tarif"* in the Klikkalteng data.

5. CONCLUSION

This research using the Jaro Winkler algorithm produces an average online news similarity value of 74,49% with 55 data, 43 data with high plagiarism levels and 12 data with medium plagiarism levels and there are weaknesses in some undetected data which should have heavy plagiarism levels but not heavy and vice versa, therefore there is need to be a more in-depth study to overcome these problems. Suggestions from researchers for future research, it is necessary to apply a method or add a synonym dictionary to handle synonym sentences in the text data so that the results of the Jaro Winkler algorithm are even better.

REFERENCES

[1] E. Kartinawati, "Jurnalisme Kloning di Kalangan Wartawan Kota Surakarta," *J. Messenger*, vol. 9, no. 1, p. 91, Jan. 2017, doi: 10.26623/themessenger.v9i1.432.

[2] N. I. Kurniati, A. Rahmatulloh, and R. N. Qomar, "Web Scraping and Winnowing Algorithms for Plagiarism Detection of Final Project Titles," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 10, no. 2, p. 73, Aug. 2019, doi: 10.24843/LKJITI.2019.v10.i02.p02.

[3] M. Z. Naf'an, A. Burhanuddin, and A. Riyani, "Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen," *J. Linguist. Komputasional*, vol. 2, no. 1, p. 23, Mar. 2019, doi: 10.26418/jlk.v2i1.17.

[4] T. Tinaliah and T. Elizabeth, "Perbandingan Hasil Deteksi Plagiarisme Dokumen dengan Metode Jaro-Winkler Distance dan Metode Latent Semantic Analysis," *J. Teknol. dan Sist. Komput.*, vol. 6, no. 1, pp. 7–12, Jan.

2018, doi: 10.14710/jtsiskom.6.1.2018.7-12.

[5] S. C. Cahyono, "Comparison of document similarity measurements in scientific writing using Jaro-Winkler Distance method and Paragraph Vector method," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 662, no. 5, p. 052016, Nov. 2019, doi: 10.1088/1757-899X/662/5/052016.

[6] M. A. Yulianto and N. Nurhasanah, "The Hybrid of Jaro-Winkler and Rabin-Karp Algorithm in Detecting Indonesian Text Similarity," *J. Online Inform.*, vol. 6, no. 1, p. 88, Jun. 2021, doi: 10.15575/join.v6i1.640.

[7] B. Leonardo and S. Hansun, "Text Documents Plagiarism Detection using Rabin-Karp and Jaro-Winkler Distance Algorithms," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 5, no. 2, p. 462, Feb. 2017, doi: 10.11591/ijeecs.v5.i2.pp462-471.

[8] S. Christina, E. D. Oktaviyani, and B. Famungkas, "Mendeteksi Plagiarism Pada Dokumen Proposal Skripsi Menggunakan Algoritma Jaro Winkler Distance," *J. SAINTEKOM*, vol. 8, no. 2, p. 143, Sep. 2018, doi: 10.33020/saintekom.v8i2.68.

[9] P. Novantara, "Implementasi Algoritma Jaro-Winkler Distance Untuk Sistem Pendeteksi Plagiarisme Pada Dokumen Skripsi," *Buffer Inform.*, vol. 3, no. 1, Apr. 2018, doi: 10.25134/buffer.v3i2.960.

[10] C. Varol and H. M. T. Abdulhadi, "Comparision of String Matching Algorithms on Spam Email Detection," in *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, Dec. 2018, pp. 6–11. doi: 10.1109/IBIGDELFT.2018.8625317.

[11] H. A. Rouf, A. Wijayanto, and A. Aziz, "Deteksi Plagiarisme Skripsi Mahasiswa dengan Metode Single-link Clustering dan Jaro-Winkler Distance," *J. PILAR Teknol. J. Ilm. Ilmu Ilmu Tek.*, vol. 5, no. 1, Jun. 2020, doi: 10.33319/piltek.v5i1.50.

[12] R. A. Salim, M. R. D. Septian, S. Suhartini, D. Anggraini, and Q. Qomariyah, "Aplikasi Pendeteksi Kesamaan Dokumen Dengan Menggunakan Algoritma Jarak Jaro Winkler Dan Levenshtein," *Sebatik*, vol. 25, no. 1, pp. 35–41, Jun. 2021, doi: 10.46984/sebatik.v25i1.1309.

[13] I. E. Agbehadji, H. Yang, S. Fong, and R. Millham, "The Comparative Analysis of Smith-Waterman Algorithm with Jaro-Winkler Algorithm for the Detection of Duplicate Health Related Records," *2018 Int. Conf. Adv. Big Data, Comput. Data Commun. Syst. icABCD 2018*, pp. 1–10, 2018, doi: 10.1109/ICABCD.2018.8465458.

[14] V. Nurcahyawati and Z. Mustaffa, "Online Media as a Price Monitor: Text Analysis using Text Extraction Technique and Jaro-Winkler Similarity Algorithm," in *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, Dec. 2020, pp. 1–6. doi: 10.1109/ETCCE51779.2020.9350898.

[15] M. H. P. Swari, C. A. Putra, and I. P. S. Handika, "Plagiarsm Checker pada Sistem Manajemen Data Tugas Akhir," *J. Sains dan Inform.*, vol. 7, no. 2, pp. 192–201, Dec. 2021, doi: 10.34128/jsi.v7i2.338.

[16] S. R. Alenazi, K. Ahmad, and A. Olowolayemo, "A review of similarity measurement for record duplication detection," in *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*, Nov. 2017, vol. 2017-Novem, pp. 1–6. doi: 10.1109/ICEEI.2017.8312386.

[17] D. Hu and A. Yin, "Efficient fuzzy keyword search scheme over encrypted data in cloud computing based on Bed-tree index structure," *J. Intell. Fuzzy Syst.*, pp. 1–13, Aug. 2021, doi: 10.3233/JIFS-202844.

[18] Sandhya and U. Ghose, "san_sim: Factual and efficient URL text similarity algorithm," in *2017 3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, Dec. 2017, pp. 359–364. doi: 10.1109/ICATCCT.2017.8389161.

[19] Y. Rochmawati and R. Kusumaningrum, "Studi Perbandingan Algoritma Pencarian String dalam Metode Approximate String Matching untuk Identifikasi Kesalahan Pengetikan Teks," *J. Buana Inform.*, vol. 7, no. 2, pp. 125–134, Jan. 2016, doi: 10.24002/jbi.v7i2.491.

[20] H. Cho, J. An, I. Hong, and Y. Lee, "Automatic Sensor Data Stream Segmentation for Real-time Activity Prediction in Smart Spaces," in *Proceedings of the 2015 Workshop on IoT challenges in Mobile and Industrial Systems*, May 2015, pp. 13–18. doi: 10.1145/2753476.2753484.

[21] A. A. P. Ratna, R. Sanjaya, T. Wirianata, and P. Dewi Purnamasari, "Word level auto-correction for latent semantic analysis based essay grading system," in *2017 15th International Conference on Quality in Research (QiR) : International Symposium*

on Electrical and Computer Engineering, Jul. 2017, vol. 2017-Decem, pp. 235–240. doi: 10.1109/QIR.2017.8168488.

[22] Y. A. Gerhana *et al.*, "Computer speech recognition to text for recite Holy Quran," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 434, no. 1, p. 012044, Dec. 2018, doi: 10.1088/1757-899X/434/1/012044.

[23] M. Elveny, S. M. Hardi, I. Jaya, and P. Gundari, "Web-based E-Commerce Products Grouping," *J. Phys. Conf. Ser.*, vol. 1898, no. 1, p. 012018, Jun. 2021, doi: 10.1088/1742-6596/1898/1/012018.

[24] S. Sastroasmoro, "Beberapa Catatan tentang Plagiarisme *," *Maj. Kedokt. Indones.*, vol. Volum: 57, pp. 239–244, 2007.

[25] R. Feldman and J. Sanger, *The Text Mining Handbook*. Cambridge: Cambridge University Press, 2006. doi: 10.1017/CBO9780511546914.

[26] P. Wahyuningtias, H. W. Utami, U. A. Raihan, and H. N. Hanifah, "COMPARISON OF RANDOM FOREST AND SUPPORT VECTOR MACHINE METHODS ON TWITTER SENTIMENT ANALYSIS (CASE STUDY : INTERNET SELEBGRAM RACHEL VENNYA ESCAPE FROM QUARANTINE) PERBANDINGAN METODE RANDOM FOREST DAN SUPPORT VECTOR MACHINE PADA ANALISIS SENTIMEN TWITT," vol. 3, no. 1, pp. 141–145, 2022, doi: 10.20884/1.jutif.2022.3.1.168.

[27] P. M. Prihatini, I. K. G. D. Putra, I. A. D. Giriantari, and M. Sudarma, "Stemming Algorithm for Indonesian Digital News Text Processing," *Int. J. Eng. Emerg. Technol.*, vol. 2, no. 2, pp. 1–7, 2017.

[28] Y. Wang, J. Qin, and W. Wang, "Efficient approximate entity matching using Jaro-Winkler distance," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10569 LNCS, pp. 231–239, 2017, doi: 10.1007/978-3-319-68783-4_16.

[29] A. Prasetyo, W. M. Baihaqi, and I. S. Had, "Algoritma Jaro-Winkler Distance: Fitur Autocorrect dan Spelling Suggestion pada Penulisan Naskah Bahasa Indonesia di BMS TV," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, p. 435, Oct. 2018, doi: 10.25126/jtiik.201854780.