**Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)**

journal homepage: http://journal.lembagakita.org/index.php/jtik

# Text Mining Analysis untuk Identifikasi Artikel *Hoax* Menggunakan Algoritma *Cosine Similarity*

Yulianty Lasena [1], Husdi [*2], Maryam Hasan [3]

[1,2,3] *Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Ichsan Gorontalo.*

a r t i c l e i n f o

*Article history:*
Received 11 Okctober 2020
Received in revised form
11 November 2020
Accepted 22 November 2020
Available *online* November 2020

DOI:
https://doi.org/10.35870/jtik.v4i2.149

*Keywords:*
*Hoax, Articles, Cosine Similarity, Text Mining.*

*Kata Kunci:*
*Hoax, Artikel, Cosine Similarity, Text Mining.*

a b s t r a c t

The impact of significant technological developments in everyday life starts from simple activities to activities that require a high level of precision. The development of information technology also contributes to the dissemination of news. In Indonesia, Information Technology is also developing rapidly where internet users currently number 132.7 million or 52% of Indonesia's population. The exchange of information between people is a positive thing, but its dissemination through social media is not all facts. In a number of cases that have occurred, for example the spread of news that is not factual is often called a hoax. The latest technology that can help overcome this, one of which is the technology known as Text Mining. This is used to solve problems faced by internet users with fake information (hoax).

a b s t r a k

Dampak perkembangan teknologi yang signifikan dalam kehidupan sehari-hari dimulai dari kegiatan yang sederhana hingga kegiatan yang membutuhkan tingkat ketelitian yang tinggi. Perkembangan tekhnologi informasi juga turut serta mendorong dalam penyebaran berita. Di Indonesia teknologi informasi ikut berkembang pesat dimana pengguna internet saat ini berjumlah 132,7 juta atau 52% dari jumlah penduduk Indonesia. Pertukaran informasi antar sesama merupakan hal yang positif, namun penyebarannya melalui media sosial yang isinya tidak semua fakta. Dalam beberapa macam kasus yang telah terjadi misalkan penyebaran berita yang bukan fakta sering disebut hoax. Teknologi terkini yang bisa membantu mengatasi hal tersebut, salah satunya adalah teknologi yang dikenal dengan nama Text Mining (Penambangan Teks). Hal ini dimanfaatkan untuk memecahkan masalah yang di hadapi oleh pengguna internet terhadap informasi palsu (hoax).

*Corresponding author. Email: mr.husdi@unisan.ac.id [2].
© E-ISSN: 2580-1643.

## 1. Latar Belakang

Perkembangan teknologi saat ini sangat memberikan pengaruh dalalam kehidupan sehari-hari mulai dari kegiatan sederhana sampai pada kegiatan yang membutuhkan ketelitian tinggi, perkembangan teknologi informasi juga turut serta mendorong dalam penyebaran berita. Di Indonesia teknologi informasi ikut berkembang pesat dimana pengguna internet saat ini berjumlah 132,7 juta atau 52% dari jumlah penduduk Indonesia[1].

Pertukaran informasi merupakan hal yang positif, namun penyebarannya melalui media sosial yang isinya tidak semua fakta . Dalam beberapa kasus yang telah terjadi misalkan penyebaran berita yang bukan fakta sering disebut *Hoax*. Sedangkan *Hoax* adalah informasi berbahaya yang sering menyesatkan persepsi manusia dengan menyebarkan informasi yang salah namun dianggap sebagai kebenaran[2].

Salah satu dampak buruk dari informasi *Hoax* adalah kerusakan finansial dan menyakiti setiap penggunanya bahkan lebih buruknya *Hoax* memiliki kemampuan untuk mengumpulkan informasi dan meyakinkan penerima untuk menghadiri acara yang tidak pernah ada[3]. Menteri komunikasi dan informatika mengatakan sejauh ini sudah ada hampir 800 ribu situs yang menyebar *Hoax* di internet. Sehingga dari banyaknya jumlah situs *Hoax* yang tersebar di internet akan sangat sulit untuk mengelompokkan informasinya [4].

Dalam *Cambridge Dictionari*, kata *Hoax* sendiri artinya tipuan atau lelucon. Kegiatan menipu, rencana menipu, trik menipu, disebut dengan *Hoax*.: Sedangkan menurut wikipedia, *Hoax* adalah usaha untuk menipu atau mengakali pembaca/pendengarnya untuk mempercayai sesuatu, *Hoax* tujuannya yakni membuat opini publik, menggiring opini, membentuk persepsi, juga untuk bersenang-senang yang menguji kecerdasan dan kecermatan pengguna internet dan media sosial.

Permasalahan utama yang diangkat dalam penelitian ini adalah maraknya artikel-artikel *Hoax* khususnya yang dimuat pada media *online* sehingga menyebabkan masyarakat kesulitan dalam membedakan dan menyaring berita yang *Hoax* dengan bukan. Apalagi ditengah *pandemic* seperti saat ini banyak sekali informasi simpang siur tentang *Covid-19* dimedia sosial sehingga banyak masyarakat termakan *Hoax* yang beredar. Selain itu di tahun politik juga sering dijumpai adanya berita-berita *Hoax* yang diperuntukkan untuk propaganda politik dalam menarik simpati masyarakat maupun untuk menjatuhkan lawan politiknya.

Teknologi terkini yang bisa membantu mengatasi hal tersebut, salah satunya adalah teknologi yang dikenal dengan nama *Text Mining* [5]. Berbeda dengan *Data Mining* yang menggali, menemukan atau menambang informasi dari kumpulan data (*dataset*)[6], *Text Mining* merupakan salah satu langkah dari analisis teks yang dilakukan secara otomatis oleh komputer untuk menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah dokumen [7, 8].

*Text Mining* merupakan teknologi yang dapat membantu memecahkan masalah kompleks yang didukung oleh perangkat–perangkat teknologi yang semakin maju dan sudah lumrah di gunakan oleh masyarakat umum. Hal ini dimanfaatkan untuk memecahkan masalah yang di hadapi oleh pengguna internet terhadap informasi palsu (*Hoax*). Dengan kemampuan text mining tersebut memberikan peluang bagi penulis untuk membuat aplikasi yang dapat membantu dalam mendeteksi Berita *Hoax*.

*Cosine similarity* adalah metode *similaritas* yang digunakan untuk menghitung kesamaan antara dua dokumen. Metode yang dipergunakan yaitu melakukan perhitungan ukuran kesamaan antara dua buah *vektor* dalam sebuah ruang dimensi yang didapat dari nilai *cosinus* sudut dari perkalian antara dua *vektor* yang dibandingkan karena cosinus dari 0 adalah 1 dan kurang dari 1 untuk nilai sudut yang lain[9].

Penelitian dengan menggunaan Algoritma *Cosine Similarity* diantaranya adalah:

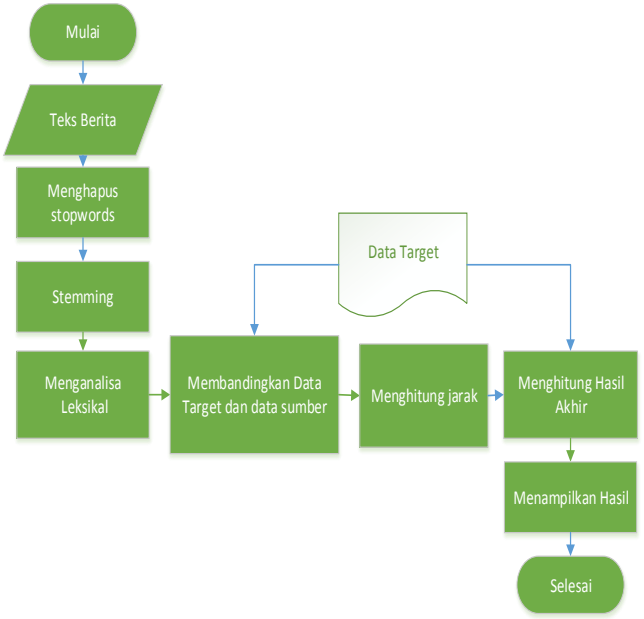
1. Peracangan Sistem Klasifikasi Surat Elektronik *(E-Mail)* Menggunakan Metode Cosine Similarity[10].
2. *Latent semantic analysis and cosine similarity for hadith search engine* [11].
3. *The optimalization of cosine similarity method in detecting similarity degree of final project by the college students*[12] dan lain-lain.

Pada penelitian tersebut Algoritma *Cosine Similarity* dapat digunakan untuk mengukur kesamaan teks yang berkonsepkan text mining. Berdasarkan permasalahan dan beberapa penelitin yang telah diuraikan maka pada penelitian ini akan menggunakan Algoritma *Cosine Similarity* untuk mengukur tingkat kemiripan dokumen. Dan hasil akhirnya diharapkan dapat memberikan persentase dari berita yang kemungkinan mengandung *Hoax* atau tidak.

## 2. Metode Penelitian

*Model Usulan*

Adapun Model yang diusulkan dalam penelitian ini adalah sebagai berikut :




Gambar 1. Diagram Sistem deteksi *Hoax* secara umum

Pada gambar 1 terdapat blok diagram model yang diusulkan pada penelitian ini dimana dijelaskan bahwa terdapat *PraProcessing* yaitu *tokenisasi* dan *Stemming* selanjutnya data target adalah artikel yang akan di analisis dengan menggunakan klasifikasi untuk mengetahui apakah artikel tersebut *Hoax* atau bukan.

*PraProses* data adalah langkah paling mendasar dalam *text mining* di mana data mentah ditransformasi menjadi bentuk yang lebih bermakna dan dapat dipahami. Hal ini disebabkan karena data tekstual yang diambil dari dokumen bersifat tidak terstruktur, tidak konsisten, dan banyak mengandung *noise* sehingga diperlukan upaya normalisasi agar data dapat diproses lebih lanjut[13]. Tiga Proses Utama dalam Pra-Pemrosesan data secara umum adalah[3] :

a. *Analisis Leksikal.*
b. *Penghapusan Stopwords.*
c. *Stemming.*

Salah Satu Algoritma yang sering digunakan dalam *information retrieval* untuk menentukan ukuran kesamaan adalah *Cosine similarity* [14] dan akan digunakan untuk melakukan perhitungan kesamaan dari dokumen pada penelitin ini. Adapun rumus yang digunakan pada *Cosine Similarity* adalah sbb. [15]

$$Cos\ a = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

Tabel 1. Keterangan Rumus Cosine Similarity [15]

| Variabel | Keterangan |
|---|---|
| A | Vektor A, yang akan dibandingkan kemiripannya |
| B | Vektor B, yang akan dibandingkan kemiripannya |
| A · B | *dot product* antara vektor A dan vektor B |
| \|A\| | = panjang vektor A |
| \|B\| | = panjang vektor B |
| \|A\|\|B\| | = *cross product* \|A\| dan \|B\| |

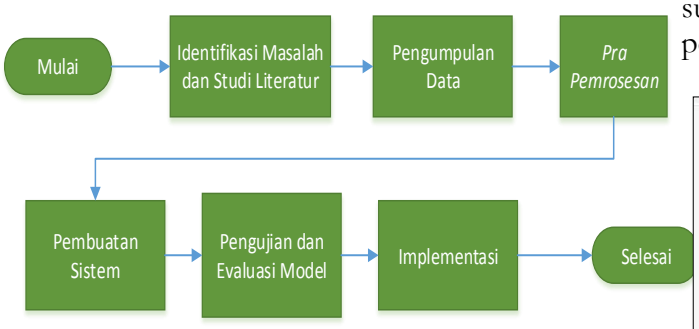
Pseudocode dari Algorithma *Cosine Similarity* adalah sebagai berikut [16]:

**Algorithm : Cosine Similarity**

```
1.  class Mapper
2.  method Map( docs )
3.  n = docs.length
4.  for i = 0 to docs.length
5.  for j = i+1 to docs.lengt
6.  write ( ( docs[i].id, docs[j].id ),( docs[i].tfidf,
    docs[j].tfidf ) )

7.  class Reducer
8.  method Reduce( ( docId_A, docId_B ),(
    docA.tfidf, docB.tfidf ) )
9.  A = docA.tfidf
10. B = docB.tfidf
11. cosine = sum( A×B )/ (sqrt( sum(A2) )
12. return ( (docId_A, docId_B), cosine )
```

*Tahapan Penelitian:*
Berikut merupakan diagram alir dari tahapan penelitian.




Gambar 2. Tahapan penelitian

*Tahap Pengumpulan Data*
Dalam penelitian ini, data yang digunakan adalah data berita *hoax*, baik itu dari sosial media ataupun dari situs web yang khusus menyajikan berita untuk khalayak umum.

*Tahap Pembuatan Aplikasi*
Merupakan tahapan di mana kita melakukan pengembangan, melakukan tahap pembangunan sesuai dari hasil analisa dan desain sistem yang sebelumnya, termasuk didalamnya membangun sebuah aplikasi, menulis code program dan membuatnya dalam bentuk sebuah antarmuka dan integrasi dari input, proses dan, output yang meruakan bagian-bagian dari system program. Adapun untuk pembuatan website menggunakan PHP dan MYSQL.

*Tahap Pengujian*
Tahap pengujian di lakukan setelah semua model selesai di buat, dan program dapat berjalan, di mana seluruh perangkat lunak, program tambahan, dan semua program yang terlibat dalam pembangunan sistem diuji untuk memastikan apakah sistem yang dirancang dapat berjalan dengan baik maka di lakukan pengujian whitebox untuk menguji prosesur-prosedur pada proses Selanjutnya pengujian blackbox untuk menguji event-event pada aplikasi yang di bangun.

## 3. Hasil dan Pembahasan

*Hasil Perancangan Sistem*
Berikut ini merupakan tampilan dari alikasi yang sudah dirancang dengan menggunakan bahasa pemrograman php.


Gambar 3. Halaman Awal Aplikasi

Halaman Awal Aplikasi pada gambar 3. merupakan tampilan awal pada saat aplikasi ini akan dibuka


Gambar 4. Halaman Tentang Aplikasi

Gambar 4 pada aplikasi ini menjelaskan tentang masalah yang akan di beritakan dikatakan *hoax* atau bukan *hoax*


Gambar 5. Cek Artikel *Hoax*

Gambar 5 pada aplikasi ini yaitu mengecek artikel *Hoax*. Tampilan ini dilakukan dengan cara memasukkan berita kemudian di pilih tombol cek artikel untuk mendapatkan hasil identifikasi tentang artikel *hoax*.


Gambar 6. Hasil Identifikasi

Setelah melakukan pengecekkan artikel pada gambar 6. di atas maka akan muncul hasil identifikasi apakah berita yang sudah dimasukkan tadi merupakan berita *hoax*, berita bukan *hoax* dan berita yang tidak dikenali

## 4. Kesimpulan

Berdasarkan dari hasil perancangan ssstem dan hasil penelian yang telah dilakukan diperoleh kesimpulan bahwa algoritma *cosine similarity* dapat digunakan untuk mendeteksi artikel *hoax* dan dapat dimanfaatkan untuk memecahkan masalah oleh pengguna internet tentang informasi palsu (*hoax*).

## 5. Ucapan Terima Kasih

Penelitian didanai oleh kementerian riset dan teknologi / Badan riset dan inovasi nasional pada Skim PDP (Penelitian Dosen Pemula) Tahun 2020.

## 6. Daftar Pustaka

[1] R Pakpahan, "Menurut Pakpahan (2017), Teknologi Informasi untuk Indonesia sendiri ikut berkembang pesat dimana pengguna internet di Indonesia saat ini berjumlah 132,7 juta atau 52% dari jumlah penduduk Indonesia," Konf. Nas. Ilmu Sos. dan Teknol., 2017.

[2] A. Errissya Rasywir and Purwarianti, "Eksperimen pada Sistem Klasifikasi Berita *Hoax* Berbahasa Indonesia Berbasis Pembelajaran Mesin," 2015.

[3] G. W. Frista, "Deteksi Konten *Hoax* Berbahasa Indonesia Pada Media Sosial Menggunakan Metode Levenshtein Distance," pp. 1–78, 2018.

[4] D. Maulina and R. Sagara, "Klasifikasi Artikel *Hoax* Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency – Inverse Document Frequency," vol. 2, no. 1, pp. 35–40, 2018.

[5] A. G. TAMMAM, "DETEKSI HOAKS PADA MEDIA SOSIAL BERBASIS TEXT MINING CLASSIFICATION SYSTEM," Univ. Nusant. PGRI KEDIRI, vol. 15, no. 2, pp. 2017–2019, 2018.

[6] H. Husdi and Y. Lasena, "Real Time Analisys Berbasis Internet Of Things Untuk Prediksi Iklim Lahan Pertanian," J. MEDIA Inform. BUDIDARMA, vol. 4, no. 3, pp. 834–840, 2020.

[7] M. P. R. Putra and K. R. N. Wardani, "Penerapan Text Mining Dalam Menganalisis Kepribadian Pengguna Media Sosial," JUTIM (Jurnal Tek. Inform. Musirawas), vol. 5, no. 1, pp. 63–71, 2020.

[8] S. N. Asiyah and K. Fithriasari, "Klasifikasi Berita *Online* Menggunakan Metode Support Vector Machine Dan K-Nearest Neighbor," J. Sains dan Seni ITS, vol. 5, no. 2, pp. 317–322, 2016.

[9] M. D. R. Wahyudi, "Penerapan Algoritma Cosine Similarity pada Text Mining Terjemah Al-Qur'an Berdasarkan Keterkaitan Topik," Semesta Tek., vol. 22, no. 1, pp. 41–50, 2019.

[10] S. Syamsuddin, Ahyuna, and K. Alloto'dang, "Peracangan Sistem Klasifikasi Surat Elektronik (E-Mail) Menggunakan Metode Cosine Similarity," J. Syntax Admiration, vol. 1, no. 5, pp. 594–606, 2020.

[11] W. Darmalaksana, C. Slamet, W. B. Zulfikar, I. F. Fadillah, D. S. adillah Maylawati, and H. Ali, "Latent semantic analysis and cosine similarity for hadith search engine," Telkomnika (Telecommunication Comput. Electron. Control., vol. 18, no. 1, pp. 217–227, 2020.

[12] R. A. Purba, S. Suparno, and M. Giatman, "The optimalization of cosine similarity method in detecting similarity degree of final project by the college students," IOP Conf. Ser. Mater. Sci. Eng., vol. 830, no. 3, pp. 1–7, 2020.

[13] W. Hardi, "Pengelompokan Topik Dokumen Berbasis Text Mining Dengan Algoritme K-Means : Studi Kasus Pada Dokumen Kedutaan Besar Australia Jakarta," vol. 21, no. 1, pp. 67–76, 2019.

[14] R. Ahmad, "E-learning Automated Essay Scoring System Menggunakan Metode Searching Text Similarity Matching Text," J. Penelit. Enj., vol. 22, no. 1, pp. 38–43, 2019.

[15] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Jurnal Teknik Elektro," J. Tek. Elektro, vol. 9, no. 1, pp. 18–23, 2017.

[16] G. S. Victor, P. Antonia, and S. Spyros, "CSMR: A scalable algorithm for text clustering with cosine similarity and MapReduce," IFIP Adv. Inf. Commun. Technol., vol. 437, no. September, pp. 211–220, 2014.