JAMI: Jurnal Ahli Muda Indonesia Vol. 3 No. 1 (2022)
Received 12th Oct 2020, revision 17th Nov 2020, accepted 20th Jan 2021, published 30th Jun 2022

Implementasi Text Mining Menggunakan Metode Cosine Similarity untuk Klasifikasi Konten Berita di Postingan Grup Facebook Info Lantas dan Kriminal Pasuruan

Rio Feriangga Kurniawan[1], Mochammad Firman Arif[2]

[1,2]Program Studi Informatika, Fakultas Teknologi Informasi, Universitas Merdeka Pasuruan,
e-mail: rioferianggakurniawan@gmail.com[1], mochammadfirmanarif @gmail.com[2]

Penulis Korespondensi. Rio Feriangga Kurniawan, Program Studi Informatika, Fakultas Teknologi Informasi, Universitas Merdeka Pasuruan,
e-mail: rioferianggakurniawan@gmail.com[1]

A B S T R A K

Objektif. Berita merupakan bentuk laporan tentang suatu kejadian yang sedang terjadi baru baru ini atau keterangan terbaru dari suatu peristiwa. Persebaran nya pun kini sudah mencapai ranah sosial media facebook, seperti berita yang berada di group Facebook Info lantas dan kriminal pasuruan, masih acak dan tidak ada pengelompokan beritanya sesuai dengan kemiripan tema atau isi.

Material and Metode. Penelitian ini bertujuan untuk menerapkan Teknik Text Mining menggunakan metode Cosine Similarity dalam melakukan klasifikasi konten berita berdasarkan kesamaan tema dan isi konten berita.

Hasil. Hasil dari penelitian menggunakan metode Cosine Similarity didapat bahwa nilai bobot kata pada tiap konten berita dapat mempengaruhi hasil klasifikasi dengan menggunakan algoritma Cosine Similarity.

Kesimpulan. Berdasarkan hasil dari penelitian tentang Implementasi Text Mining untuk Klasifikasi konten berita di grup Facebook Info Lantas Dan Kriminal Pasuruan menggunakan metode Cosine Similarity dapat disimpulkan bahwa nilai bobot kata pada tiap konten berita dapat mempengaruhi hasil klasifikasi dengan menggunakan algoritma Cosine Similarity. Dimana konten berita yang tidak berkategori akan dihitung bobot kata dan similarity nya dengan konten berita yang telah diketahui kategorinya sehingga muncul nilai similarity tertinggi sebagai kemiripan konten.

Kata kunci:
Berita, Text Mining, Cosine Similarity.

A B S T R A C T

Objective. News is a form of a report about an event that is happening recently or the latest information of an event. Its distribution has now reached the realm of Facebook social media, such as news in the Facebook Info group and Pasuruan crimes, which are still random, and there is no grouping of news according to similar themes or content.

Materials and Methods. This study aims to apply the Text Mining Technique using the Cosine Similarity method in classifying news content based on the similarity of themes and content of news content.

Result. The results of the study using the Cosine Similarity method showed that the word weight value of each news content could affect the classification results using the Cosine Similarity algorithm.

Conclusion. Based on the results of research on the implementation of Text Mining for the classification of news content in the Pasuruan Traffic Info and Criminal Facebook group using the Cosine Similarity method, it can be concluded that the word weight value of each news content can affect the classification results using the Cosine Similarity algorithm where news content that is not categorized will be calculated the weight of the word and its similarity with news content that has a known category so that the highest similarity value appears as a content similarity.

Keywords:
News, Text Mining, Cosine Similarity.

JAMI: Jurnal Ahli Muda Indonesia Vol. 3 No. 1 (2022)
Received 12 Oct 2020, revision 17 Nov 2020, accepted 20 Jan 2021, published 30 Jun 2022

1. PENDAHULUAN

Banyak instansi yang bergerak dalam penyaluran informasi masyarakat atau berita yang pada awalnya menyampaikan berita melalui media Televisi, Surat Kabar, Majalah atau Radio sudah mulai menggunakan sistem berbasis web untuk menyampaikan beritanya secara up to date (Fajar, Muhammad. 2008).

Fenomena yang digemari pada era sekarang yaitu sebuah jejaring sosial Facebook, baik kalangan remaja ataupun dewasa hampir semuanya memiliki akun Facebook dan biasanya tergabung dalam suatu grup yang bertujuan untuk saling memberi berita ataupun informasi terkini tentang keadaan yang sedang terjadi di suatu tempat. Pada umumnya berita yang di sebarkan membuat pembaca terkadang bingung karena tidak mengetahui kategori berita yang sedang dibaca.

Text Mining merupakan salah satu cabang ilmu data mining yang menganalisis data berupa dokumen teks. Menurut Han, Kamber, dan Pei (dalam Prilianti dan Wijaya, 2014), Text Mining adalah 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. Ide awal pembuatan Text Mining adalah untuk menemukan pola-pola informasi yang dapat digali dari suatu teks yang tidak terstruktur. Sebelum suatu data teks dianalisis menggunakan metode dalam Text Mining perlu dilakukan Preprocessing Text diantaranya adalah tokenizing, case folding, stopwords, dan stemming. Setelah dilakukan preprocessing maka selanjutnya dilakukan metode klasifikasi dalam mengelompokkan dalam masing-masing kategori (Prilianti, Wijaya. 2014).

Algoritma Cosine Similarity merupakan algoritma yang digunakan untuk menghitung similarity (tingkat kesamaan) antar dua buah objek. Secara umum penghitungan metode ini didasarkan pada vector space similarity measure yang dinyatakan dalam dua buah vector dengan menggunakan keywords (kata kunci) dari sebuah dokumen sebagai ukuran. Kelebihan metode Cosine Similarity ini adalah sederhana, efisien, mudah dipahami.

Maka dari itu, penulis tertarik untuk melakukan penelitian dengan judul "Implementasi Text Mining Menggunakan Metode Cosine Similarity Untuk Klasifikasi Konten Berita Di Postingan Grup Facebook Info Lantas Dan Kriminal Pasuruan".

2. MATERIAL DAN METODE

Sistem yang dibutuhkan untuk membangun aplikasi pengenalan bahasa isyarat yang mengandung kata kerja adalah sebagai berikut:

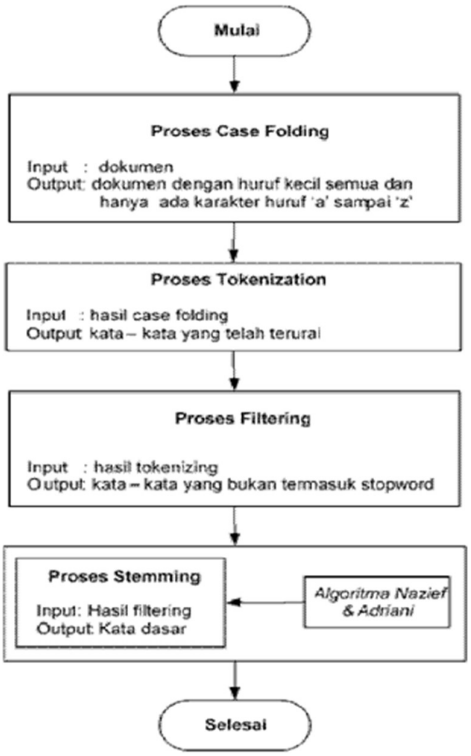
1. Perangkat Keras

Processor	: Intel Celeron 1.40GHz
RAM	: 2 GB
Hard Disk	: 500 GB

Monitor : 1366 x 768 px

2. Perangkat Lunak.
   Sistem Operasi : Windows 8.1 64 Bit Profesional
   Bahasa Pemrograman : PHP
   Code Editor : Sublime Text 3.1.1

Text Mining perlu dilakukan beberapa tahapan yang harus dilakukan untuk mengolah sumber data baik yang terstruktur, terstruktur sebagian dan yang tidak terstruktur dari beberapa sumber, maka data-data tersebut perlu dilakukan proses awal atau disebut sebagai Preprocessing Text yang bermaksud mengolah data awal yang masih bermacam – macam untuk dijadikan sebuah data teratur yang dapat dikenai atau diterapkan beberapa metode Text Mining yang ada. Sehingga preprocessing merupakan proses pengolahan data mentah sebelum masuk ke proses data mining, dimana hasilnya berupa dataset yang siap diolah sesuai metode yang akan dipakai.



Gambar 1 Flowchart Preprocessing

1. Tokenizing

Proses tokenization berguna untuk memecah setiap kalimat dari seluruh dokumen pengetahuan ke dalam kata-kata (term) dengan menggunakan pembatas tab dan karakter spasi.

2. Case Folding

Merupakan tahapan yang mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf "a" sampai "z" yang diterima.

3. Filtering / Stopword Removal

Stopword adalah kata-kata yang sering muncul dalam suatu dokumen yang kurang berguna dalam proses penggalian text. Proses Stopword removal yang berguna menghilangkan stopword, merupakan proses yang sangat penting dalam Text Mining.

4. Stemming

Proses stemming berguna untuk mengubah suatu kata menjadi kata dasarnya, misalnya kata 'mendapatkan' menjadi 'dapat'. Stemming akan meningkatkan klasifikasi teks dalam bahasa tertentu. Stemming pada penelitian ini juga sekaligus menghilangkan karakter tanda baca seperti tanda titik(.), koma (,), petik(“'), kurung(()), tanda tanya(?), tanda seru(!) dan tanda baca lainnya.

Metode TF-IDF merupakan cara pemberian bobot hubungan suatu kata terhadap dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu Term Frequency (TF) dan Inverse Document Frequency (IDF). TF merupakan frekuensi kemunculan kata (t) pada dokumen (d), Semakin besar jumlah kemunculan suatu term (TF tinggi) dalam dokumen, semakin besar pula bobotnya atau akan memberikan nilai kesesuaian yang semakin besar. Document Frequency (DF) merupakan banyaknya dokumen (d) dimana suatu kata term (t) muncul, DF menunjukkan seberapa umum kata tersebut. IDF (Inverse Document Frequency) merupakan sebuah perhitungan dari bagaimana term didistribusikan secara luas pada koleksi dokumen yang bersangkutan. IDF menunjukkan hubungan ketersediaan sebuah term dalam seluruh dokumen. Semakin sedikit jumlah dokumen yang mengandung term yang dimaksud, maka nilai IDF semakin besar. Bobot suatu kata term (t) akan semakin besar jika term sering muncul pada suatu dokumen dan akan semakin kecil jika term muncul dalam banyak dokumen. Rumus untuk

TF-IDF :

$$tf = 0{,}5 + 0{,}5 \text{ x } \frac{tf}{max(tf)}$$

$$idf_t = log\left(\frac{D}{df_t}\right)$$

$$W_{d.t} = tf_{d.t} \text{ x } idf_{d.t}$$

Keterangan:

D = dokumen ke-d

t = term ke-t dari dokumen

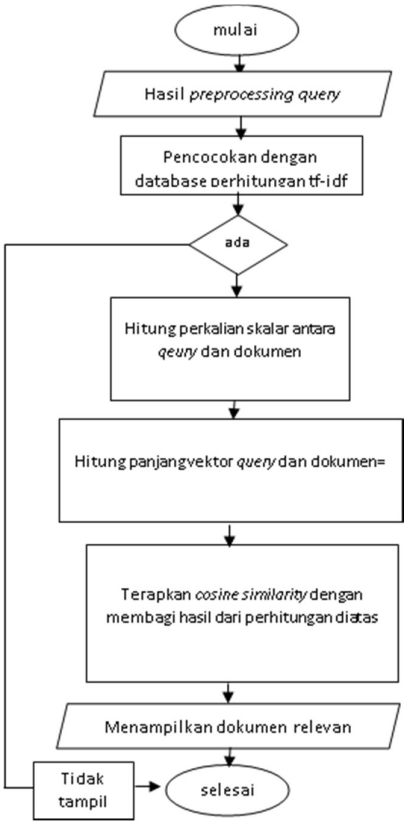
W = bobot dokumen ke-d terhadap term ke-t

tf = banyaknya term i pada sebuah dokumen

idf = Inversed Document Frequency

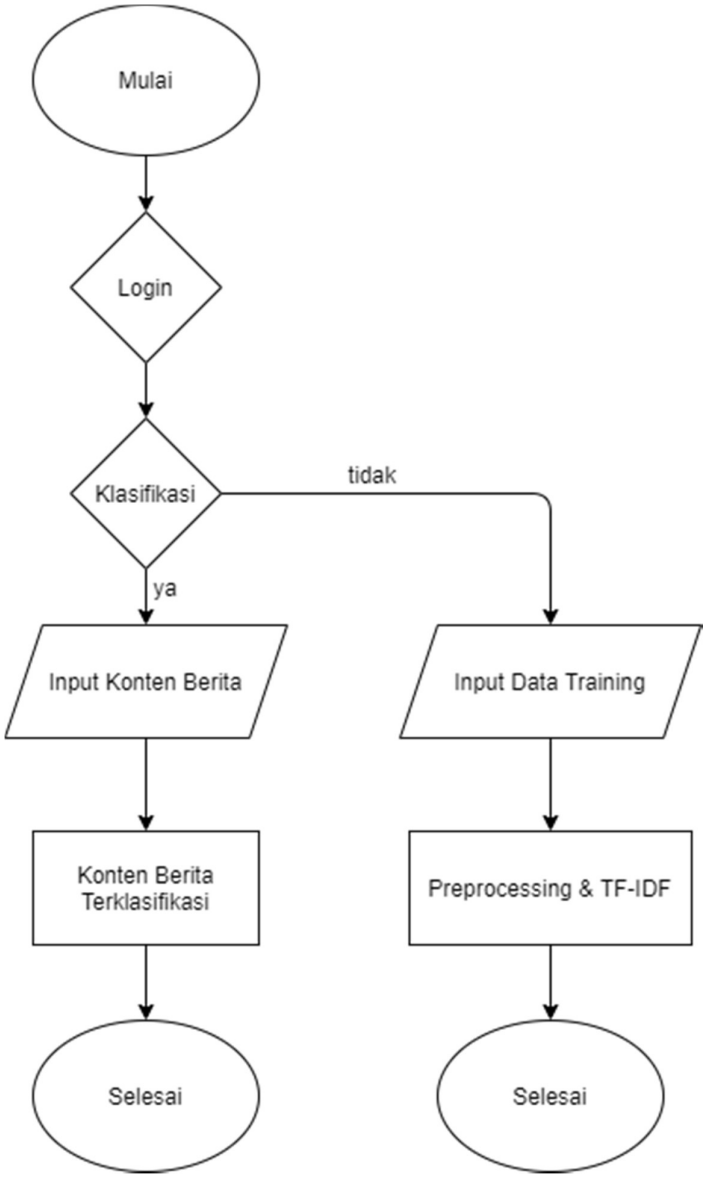
df = banyak dokumen yang mengandung term i

Metode Cosine Similarity merupakan metode yang digunakan untuk menghitung similarity (tingkat kesamaan) antar dua buah objek. Secara umum penghitungan metode ini didasarkan pada vector space similarity measure. Metode cosine similarity ini menghitung similarity antara dua buah objek (misalkan D1 dan D2) yang dinyatakan dalam dua buah vector dengan menggunakan keywords (kata kunci) dari sebuah dokumen sebagai ukuran.



Gambar 2 Flowchart Cosine Similarity

Pada gambar 2 merupakan flowchart pada proses Cosine Similarity menggambarkan proses untuk menemukan dokumen yang relevan dengan query user menggunakan metode cosine similarity, dimana query yang dimasukkan user dilakukan tahap preprocessing yang hasilnya dicocokan dengan database bobot yaitu hasil perhitungan tf•idf, apabila term ditemukan maka akan dihitung perkalian skalar antara term query dengan dokumen dengan rumus wqi x wdij, selanjutnya yaitu menghitung nilai panjang setiap dokumen termasuk query dengan mengkuadratkan bobot query dan bobot dokumen, jumlahkan nilai kuadrat dan selanjutnya diakarkan. Terakhir, membagi hasil dari perkalian skalar dan hasil panjang vektor yang sudah dihitung untuk menemukan hasil kemiripan antara query dengan dokumen, lalu sistem akan menampilkan dokumen yang relevan dengan query berdasarkan hasil perhitungan kemiripan dengan cosine similarity tersebut.



Gambar 3 Flowchart sistem

Pada gambar 3 merupakan flowchart sistem, dimana user harus login terlebih dahhulu kemudian jika user ingin menginput data training maka proses selanjutnya yaitu menginput data konten berita yang disalin, dan ditentukan kategorinya kemudian disimpan. Kemudian di proses dengan Preprocessing text & TF-IDF dimana tiap kata dalam beberapa konten berita yang tersimpan akan diberikan sesuai dengan perhitungan TF-IDF. Jika user ingin melakukan klasifikasi, langkah selanjutnya menginputkan berita yang ingin diklasifikasikan dengan hasil berupa Nilai Similarity pada tiap dokumen yang telah tersimpan. Hasil klasifikasi diambil nilai tertinggi dari beberapa dokumen. Rumus pada cosine similarity adalah

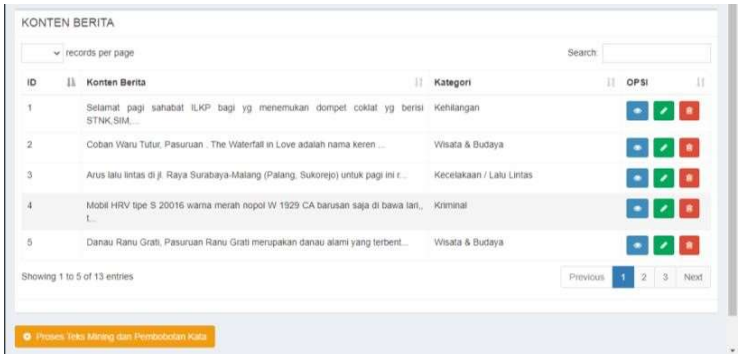
$$Cos\ \alpha = \frac{\mathbf{A \bullet B}}{|\mathbf{A}||\mathbf{B}|} = \frac{\sum_{i=1}^{n} \mathbf{A_i} \times \mathbf{B_i}}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

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 antara |A| dan |B

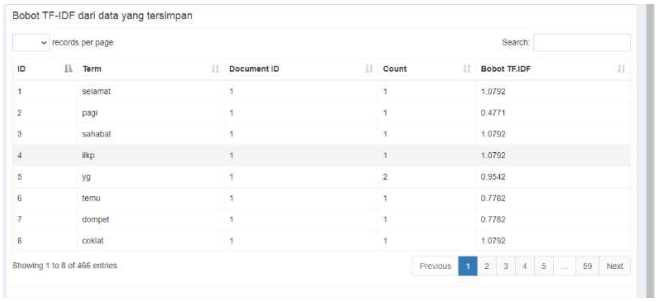
3. HASIL DAN PEMBAHASAN

Hasil penelitian dan implementasi dari aplikasi yang sudah dibuat secara keseluruhan, serta melakukan pengujian terhadap aplikasi yang sudah dibuat untuk mengetahui aplikasi tersebut telah dapat menyelesaikan permasalahan yang dihadapi sesuai dengan yang diharapkan:



Gambar 4 Data Training

Pada gambar 4 pengguna dapat melihat daftar konten berita yang tersimpan dengan kategori yang sudah ditentukan, menghapus, mengedit, serta menambah data dan melakukan proses pembobotan kata.



Gambar 5 Hasil TF-IDF

Pada gambar 5 menampilkan bobot kata dari data training, dimana bobot ini akan diproses pada menu klasifikasi dengan menggunakan algoritma cosine similarity.

Gambar 6 Hasil Klasifikasi

Pada gambar 6 menampilkan konten berita yang belum diketahui kategori kemudian telah diklasifikasikan, proses menentukan kategori berita dengan cara mengambil nilai similarity tertinggi. Jika nilai similarity tertinggi diatas bernilai 0,166 masuk kedalam kategori kehilangan maka berita yang diuji merupakan anggota dari kategori berita kehilangan.

4. KESIMPULAN

Berdasarkan hasil dari penelitian tentang Implementasi Text Mining untuk Klasifikasi konten berita di grup Facebook Info Lantas Dan Kriminal Pasuruan menggunakan metode Cosine Similarity dapat disimpulkan bahwa nilai bobot kata pada tiap konten berita dapat mempengaruhi hasil klasifikasi dengan menggunakan algoritma Cosine Similarity. Dimana konten berita yang tidak berkategori akan dihitung bobot kata dan similarity nya dengan konten berita yang telah diketahui kategorinya sehingga muncul nilai similarity tertinggi sebagai kemiripan konten.

UCAPAN TERIMAKASIH

Ucapan terima kasih disampaikan kepada semua pihak yang berperan dalam penyelesaian penelitian tentang implementasi text mining untuk klasifikasi konten berita di grup facebook info lantas dan kriminal pasuruan menggunakan metode cosine similarity sehingga penelitian ini dapat diselesaikan dengan baik dalam bentuk tulisan dan diharapkan bermanfaat untuk kedepanya.

DAFTAR PUSTAKA

Fajar, Muhammad. 2008. Media cetak era digital.

Hamzah, A. (2012). Klasifikasi Teks dengan Naïve Bayes Classifier (NBC)untuk Pengelompokan Teks Berita dan Abstract Akademis. InProsiding Seminar Nasional.

Muhammad Sholeh hudin, M Ali Fauzi, Sigit Adinugroho (2008). “Implementasi Metode *Text Mining* dan *K-Means Cluster*ing untuk Pengelompokan Dokumen Skripsi (Studi Kasus: Universitas Brawijaya)”

Nyoman Gede Yudiarta, Made Sudarma, Wayan Gede Ariastina (2018). “Penerapan Metode *Cluster*ing *Text Mining* Untuk Pengelompoan Berita Pada Unstructured Textual Data

Elly Indrayuni (2019) dengan judul "Klasifikasi *Text Mining* Review ProdukKosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma NaiveBayes "

Eko Yulian (2018), "Text *Mining* dengan *K-Means Cluste*ring padaTema LGBT dalam Arsip *Tweet* Masyarakat Kota Bandung "

Kestrilia Rega Prilianti, Hendra Wijaya, (2014), "Aplikasi *Text Mining*

Untuk Automasi Penentuan Tren Topik Skripsi Dengan Metode KMeans *Cluste*ring," Jurnal Cybermatika, Vol. 2 No. 1.

Rizky Sam Pratama. (2018) "PERBANDINGAN ANALISIS ALGORITMA*K-MEANS* DAN FUZZY C-MEANS UNTUK PENGELOMPOKAN HADITS TERJEMAHAN BAHASA INDONESIA"

G. A. Pradnyana dan N. A. Sanjaya,"*Cosine Similarity*", Perancangan Dan Implementasi Automated Document Integration Dengan Menggunakan Algoritma *Complete Linkage Agglomerative Hierarchical Clustering*,vol. 5, (2), pp. 1-10, September 2012.

Arief, M. Rudianto. (2011). Pemrograman Web Dinamis Menggunakan Php dan Mysql.

Faridl, Miftah. 2015. Fitur Dahsyat Sublime Text 3. Surabaya: LUG STIKOM