# *Rekomendasi Artikel Terkait Pada Berita Online Menggunakan Teknik Text Mining*

Ridwan Rismanto
Jurusan Teknologi Informasi
Politeknik Negeri Malang
Malang
sngbrs@polinema.ac.id

Mustika Mentari
Jurusan Teknologi Informasi
Politeknik Negeri Malang
Malang
must.mentari@polinema.ac.id

Rahmadyan Nurwidhi Biddaris
Jurusan Teknologi Informasi
Politeknik Negeri Malang
Malang
rahmadyan98@gmail.com

***Abstrak*— Pengguna internet di Indonesia berkembang pesat mampu mengubah gaya hidup masyarakat. Masyarakat memanfaatkan internet untuk mengakses informasi pada berita online. Berita online yang beragam jenis dan kategori membuat pengguna layanan berita online harus menyediakan informasi sesuai dengan permintaan pengguna. Berdasarkan permasalahan tersebutpeneliti membuat sistem rekomendasi artikel dengan membandingkan artikel pengguna dengan artikel yang ada di database kemudian di ambil sesuai dengan kemiripan yang paling tinggi. Metode yang digunakan untuk menghitung nilai kemiripan adalah Cosine Similarity yang dapat menghitung nilai kemiripan antar kalimat dengan memodelkan dokumen teks sebagai vektor kata (*terms*). Nilai kesamaan antara dokumen acuan dan dokumen pembanding di urutkan dari yang tertinggi ke yang terendah. Skor kemiripan yang lebih tinggi berarti lebih banyak relevansi antara fokumen pembanding dan dokumen acuan.**

***Kata kunci—text mining; crawling; algoritma cosine similarity***

## I. PENDAHULUAN

Perkembangan internet mendorong tumbuhnya media pemberitaan online, sehingga menjadikan media online untuk menyampaikan informasi atau berita. Layanan yang terbaik yang di berikan salah satunya yaitu membantu pengguna untuk menawarkan pilihan berita yang disesuaikan dengan minat dan ketertarikan pembaca.

Pada umumnya, rekomendasi berita yang muncul di dalam media online masih kurang lengkap dan akurat juga belum sepenuhnya menggunakan teknik *text mining* dalam menentukan rekomendasi berita dengan membandingkan kesamaan artikel pembaca dan artikel yang ada di database.

Oleh karena itu penulis membuat penelitian mengenai sistem rekomendasi artikel terkait pada berita online menggunakan teknik *text mining*, *Cosine Similarity* adalah metode yang dapat menghitung nilai kemiripan antar kalimat dengan memodelkan dokumen teks sebagai vektor kata (*terms*). Nilai kemiripan antara dokumen acuan dan dokumen pembanding diurutkan dari yang tertinggi ke yang terendah. Hal ini yang menyebabkan metode *Cosine Similarity* menjadi salah satu pilihan yang tepat berharap hasil dari penelitian bisa menjadi solusi dari masalah dan memenuhi tujuan penelitian.

## II. DASAR TEORI

### A. *Text Mining*

*Text mining* adalah proses penemuan akan informasi atau trend baru yang sebelumnya tidak terungkap dengan memproses dan menganalisa data dalam jumlah besar. Dalam menganalisa sebagian atau keseluruhan *unstructured text*, *text mining* mencoba untuk mengasosiasikan satu bagian text dengan yang lainnya berdasarkan aturanaturan tertentu. Hasil yang di harapkan adalah informasi baru atau "insight" yang tidak terungkap jelas sebelumnya.

### B. *Crawling*

*Crawling* atau *Web Crawler* adalah program yang menelusuri World Wide Web dengan cara yang metodis, otomatis dan teratur. Istilah lain untuk *web crawler* adalah *ant*, *automatic indexer*, *bots*, *web spiders* atau *web robots* [3].

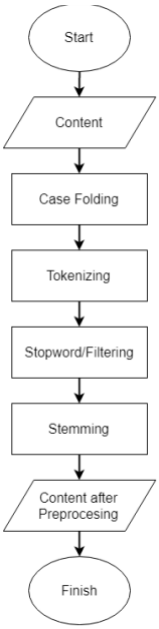
Proses crawling dimulai dengan list URL yang akan dikunjungi, disebut seeds. Kemudian web crawler akan mengunjungi URL tersebut satu per satu. Setiap page URL yang di kunjungi akan diitendifikasi apakah ada hyperlink di dalamnya. Jika ada maka akan ditambahkan kedalam list URL yang akan dikunjungi. Ini disebut crawl frontier.

### C. *Scrapy*

Scrapy adalah sebuah framework yang digunakan untuk melakukan proses crawling dan mengekstrak data yang terstruktur. Scrapy digunakan pada proses data mining, pemrosesan informasi dan pengarsipan history [4]. Scrapy dibangun dengan menggunakan python yang disupport dengan twisted (Jing Wang, 2012).

### D. *Pre-processing*

Dalam *text mining*, data text akan di proses menjadi data *numerik* agar dapat dilakukan proses lebih lanjut. Seingga dalam *text mining* ada istilah *preprocessing* data, yaitu pendahulu yang di terapkan terhadap data text yang bertujuan untuk menghasilkan data *numerik*. Tahap preprocessing dapat di lihat pada Gambar 1.

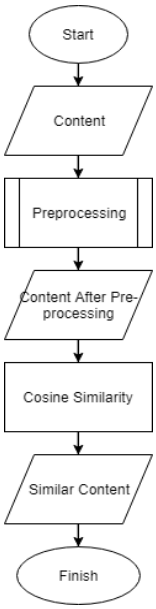



Gambar 1. *Flowchart* proses *pre-processing*

Tahap *preprocessing* yang digunakan dalam penelitian ini antara lain :

1. *Case Folding*
   Merupakan tahap perubahan huruf dari huruf kapital menjadi huruf kecil.
2. *Tokenizing*
   *Tokenizing* adalah proses memecah dokumen menjadi kumpulan kata. *Tokenization* dapat dilakukan dengan menghilangkan tanda baca dan memisahkannya per spasi. Tahapan ini juga menghilangkan karakter-karakter tertentu seperti tanda baca dan mengubah semua *token* ke bentuk huruf kecil (*lower case*) [5].
3. *Stopwords/Filtering*
   *Stopwords removal* merupakan proses penghilangan kata tidak penting pada deskripsi melalui pengecekan kata-kata hasil *parsing* deskripsi apakah termasuk di dalam daftar kata tidak penting (stoplist) atau tidak. Jika termasuk di dalam stoplist maka kata-kata tersebut akan di-remove dari deskripsi sehingga kata-kata yang tersisa di dalam deskripsi dianggap sebagai kata-kata penting atau keywords.
4. Stemming
   Stemming merupakan proses untuk mendapatkan root/stem atau kata dasar dari suatu kata dalam kalimat dengan cara memisahkan masing-masing kata dari kata dasar dan imbuhannya baik awalan (prefiks) maupun akhiran (sufiks).

Setelah melalui tahap preprocessing maka data akan disimpan dalam memori sementara dan nantinya akan diproses lebih lanjut ke dalam tahap analyzing menggunakan pembobotan TF-IDF dan klasifikasi dengan algoritma cosine similarity. Untuk lebih jelasnya proses klasifikasi dokumen dapat dilihat pada Gambar 2.




Gambar 3. *Flowchart* proses rekomendasi

Setelah dilakukan preprosesing dan dataset sudah mempunyai standar yang sama, proses selanjutnya adalah menghitung bobot kata pada masing-masing berita. Pembobotan kata yang digunakan menggunakan metode *Term Frequency* (TF). TF merupakan pembobotan dengan menghitung frekuensi kemunculan kata pada suatu dokumen. Untuk setiap kata dalam dokumen akan dihitung bobot menggunakan persamaan 1. Kata setiap dokumen $t_i$ akan dihitung dalam dokumen $d_j$ kemudian dihitung nilai TF nya :

$$W_{TF}\left(t_i, d_j\right) = f(t_i, d_j) \quad (1)$$

Sedangkan *Inverse Document Frequency* (IDF) melakukan pendekatan dengan menganggap kata yang sering muncul pada satu dokumen, tapi jarang muncul pada seluruh data set akan diberikan nilai bobot yang lebih tinggi. Dalam sebuah corpus yang terdiri dari D dokumen terdapat $d_{(ti)}$ dokumen yang mengandung kata *i*. Perhitungan IDF dari dokumen yang mengandung kata *i* dapat dilakukan dengan melihat persamaan.

$$W_{IDF}\left(t_i\right) = 1 + log\left(\frac{D}{d_{(ti)}}\right) \quad (2)$$

Perhitungan bobot TF.IDF dilakukan dengan melakukan perkalian antara persamaan 1 dengan 2 sehingga menghasilkan persamaan 3

$$W_{TF.IDF}\left(t_i.d_j\right) = f\left(t_i.d_j\right)x\left(1 + \log\left(\frac{D}{t_i}\right)\right) \quad (3)$$

Keterangan :

$W_{TF.IDF}\left(t_i.d_j\right)$ : pembobotan kata i pada dokumen j

$f\left(t_i.d_j\right)x$ : banyak kata atau term i pada dokumen j

$D$ : total dokumen dalam dataset

$t_i$ : total dokumen yang memunculkan kata i

*Consine Similarity* digunakan untuk melakukan perhitungan kesamaan dari dokumen. Rumus yang digunakan oleh *consine similarity* adalah [1].

$$Sim(\vec{q},\vec{d}) = \frac{\vec{q}.\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\sum_{k=1}^{t} w_{qk} \times w_{dk}}{\sqrt{\sum_{k=1}^{t}(w_{qk})^2}.\sqrt{\sum_{k=1}^{t}(w_{dk})^2}} \quad (4)$$

Keterangan :

$\vec{q}$ : vektor q

$\vec{d}$ : vektor d

$w_{qk}$ : bobot term *q* dalam blok $W_{qk}$

$w_{dk}$ : bobot term *d* dalam blok $W_{dk}$

*k* : jumlah *term* dalam kalimat

*t* : jumlah vektor

## III. METODE PENELITIAN

### A. Bahan Penelitian

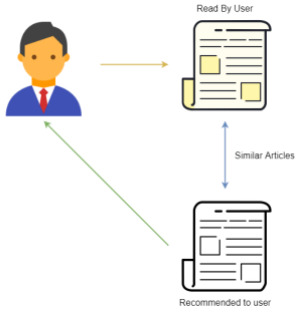
Data yang diambil merupakan data sample dari 3 situs berita online menggunakan teknik *crawling* yang sudah di jelaskan di dasar teori, data yang sudah di peroleh diolah pada sistem berbasi web yang telah disusun oleh organisasi pemrograman text mining.

### B. Deskripsi Sistem

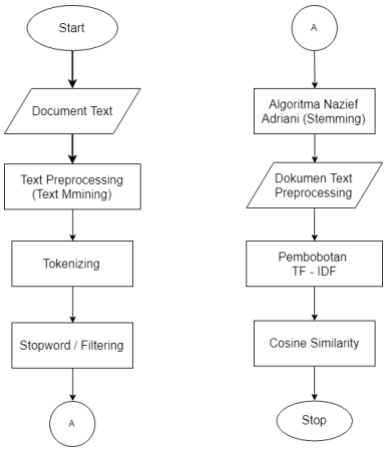
Sistem merupakan aplikasi berbasis website yang dibangun menggunakan bahasa pemrograman *Python* dan *framework flask*, sistem mengambil banyak data artikel di 3 media online dengan kategori politik, data yang sudah di simpan dapat di tampilkan melalui interface aplikasi yang dapat di akses oleh user, ketika user mengakses artikel berita tersebut sistem dapat memberikan rekomendasi artikel terkait sesuai dengan dengan artikel yang di baca oleh user melalui perhitungan kemiripan *cosine similarity*. Untuk lebih jelasnya dapat dilihat pada Gambar 3.

### C. Arsitektur Sistem

Sistem rekomendasi artikel dengan algoritma *cosine similarity* sebagai suatu sistem memiliki beberapa proses yang membangun sistem secara keseluruhan. Proses sistem rekomendasi artikel terdiri dari : proses preprocessing, pembobotan TF-IDF, dan cosine similarity [8]. Secara lengkap rancangan dari modul deteksi kesamaan dokumen dapat dilihat pada Gambar 4.




Gambar 3. Proses penyajian artikel




Gambar 4. Arsitektur Sistem

## IV. HASIL DAN PEMBAHASAN

Metode penelitian dibagi menjadi beberapa tahapan. Tahap pertama adalah mengambil data artikel di 3 situs media online dengan kategori berita politik dan di simpan di database MySQl, pengambilan data menggunakan teknik web crawling.

Tahap kedua adalah data yang ada di database di proses lagi dengan dilakukan proses *Pre-Processing* yaitu dengan dilakukanya *Case Folding*, *Tokenizing*, *Stopword*, dan *Stemming*.

Tahap ketiga, data yang sudah melalui proses text *Pre-Processing* dilakukan pembobotan *Term Frequency* dan *Inverse Document Frequency*. Dan yang terakhir adalah perhitungan dengan metode *Cosine Similarity* untuk menentukan angka kemiripan pada setiap dokumen yang di bandingkan.

Hasil pengujian dari penerapan *Text Mining* ditunjukkan pada Tabel I. Pengamatan dilanjutkan pada pencocokan query Q dengan 6 dokumen D1 – D6 yang sudah dilakukan proses *Preprocessing* yaitu *tokenisasi, stopword, dan stemming*.

Hasil *Pre-Processing* dan perhitungan *Cosine Similarity* di tunjukan pada pada Tabel II dan di urutkan berdasarkan hasil perhitungan dari yang terbesar pada Tabel III.

TABEL I. TABEL DATA SAMPLE

| Dokumen | Term yang mewakili dokumen |
|---|---|
| Q | universitas trunojoyo |
| D1 | komisi yudisial universitas jalin kerjasama berantas mafia adil |
| D2 | sar trunojoyo diklat bumi kemah wisata air terjun mojokerto bantu rector |
| D3 | roadshow speedy trunojoyo seminar internet sehat cangkruk komunitas workshop lomba band |
| D4 | perintah kabupaten pamekasan henti program bantu beasiswa mahasiswa pamekasan universitas trunojoyo |
| D5 | 11 staf universitas trunojoyo magang fakultas teknik industri uii |
| D6 | perpus universitas airlangga datang tamu 2 guru tinggi staf perpus universitas trunojoyo staf perpus universitas gunadarma |

TABEL III. HASIL PRE-PROCESSING DAN PERHITUNGAN COSINE SIMILARITY

| D1 | D2 | D3 | D4 | D5 | D6 |
|---|---|---|---|---|---|
| 0.059 | 0.0102 | 0.0102 | 0.0599 | 0.073 | 0.048 |

TABEL III. URUTAN HASIL PERHITUNGAN

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| D5 | D4 | D1 | D6 | D2 | D3 |

Dari hasil table III, dokumen yang relevan dengan Query "universitas trunojoyo" yaitu dokumen D5 dan D4

## V. PENUTUP

### A. Kesimpulan

Sistem dibangun dengan membandingkan hasil crawling data dari 3 situs berita online. Proses yang dilakukan dalam ekstraksi teks yaitu tokenisasi, stopword removal, stemming, dan pembobotan. Hasil ekstraksi lalu dibandingkan dengan menggunakan metode kemiripan *cosine similarity*. Semakin besar nilai *cosine-similarity* yang di hasilkan, maka semakin mirip kedua dokumen tersebut, sehingga rekomendasi artikel akan didasarkan pada nilai *cosine-similarity* terbesar antara dokumen artikel yang dibaca user dengan artikel yang ada di database.

Sistem yang di bangun telah memenuhi kebutuhan fungsionalitas yang menjawab hasil dari analisis permasalahan yang di tentukan pada awal penelitian. Hal tersebut sesuai dengan hasil pengujian fungsionalitas sistem yang di lakukan menggunakan *blackbox testing*. Sistem telah sesuai dengan perancangan.

### B. Saran

Penelitian yang telah diselesaikan ini memebuka beberapa penyempurnaan untuk menjadikan sistem rekomendasi ini menjadi lebih baik lagi. Penelitian lanjutan yang dapat dilakukan diantaranya:

1. Berdasarkan implementasi sistem, ditemukan bahwa teknik crawling yang dipakai masih belum dilakukan secara berkala atau otomatis, hal ini tentunya menurunkan efisiensi kinerja sistem. Untuk itu diperlukan penelitian untuk melakukan crawling data secara otomatis dengan tujuan data yang di ambil akan selalu terupdate.
2. Sistem rekomendasi artikel ini hanya di bandingkan melalui perbandingan isi content berita, kedepanya akan sangat optimal, jika sistem rekomendasi artikel dibandingkan melalui perbandingan seperti judul, pengarang, dan kategori yang lain sehingga lebih memberikan tingkat akurasi yang tinggi.

## Referensi


[1] F.Rahutomo, T.kitasuka, and M.Aritsugi, "Semantic Cosine Similarity,"
Semant. Sch, vol. 2, no. 4, pp. 4-5, 2012

[2] I. Adiwijaya, "Text Mining dan Knowledge Discovery," Kolok. Bersama komunitas datamining indones. Soft-computing Indones., pp. 1-9, 2006.

[3] A. Halim et al., "Perancangan Aplikasi Web Crawler Untuk Menghasilkan Dokumen Teks Pada Domain," vol.1, no.2, pp.2-5, 2017

[4] Fathurrahman, I.M., Nurjanah, D., & Rismala, R. (2017). Sistem Rekomendasi pada Buku dengan Menggunakan Metode Trust-Aware Recommendation. e-Proceeding of Engineering, 4(3), 4966-4977

[5] S. Vijayarani, J. Ilamathi, and Nithya, "Preprocssing Techniques for Text Mining – An Overview," int. j. Comput. Sci. Commun Networks, vol. 5, no. 1, pp. 7-16, 2015

[6] F. Rahutomo and A. R. T. H. Ririd, "Evaluasi Daftar Stopword Bahasa Indonesia," J. Teknol. Inf. dan Ilmu Komput, vol. 6, no. 1, p.41, 2019.

[7] N. C. Wibowo, D. Satria, Y. Kartika, and S. R. Wardhana, "Pengkategorian Berita Online Secara Otomatis Menggunakan Metode PLSA," vol. 3, no. 1, pp. 45-51, 2018.

[8] H. Zisopoulos, S. Karagiannidis, and S. Antaris, "Content-Based Recommendation Systems," no. June 2014, 2008.