E-ISSN : 2686-5238
P-ISSN : 2686-4916
JURNAL EKONOMI MANAJEMEN SISTEM INFORMASI (JEMSI)
https://dinastirev.org/JEMSI dinasti.info@gmail.com +62 811 7404 455
DOI: https://doi.org/10.31933/jemsi.v5i5
Received: 03 Maret 2024, **Revised:** 13 April 2024, **Publish:** 04 Mei 2024

Analisa Dokumen Menggunakan Metode TF-IDF

Arifman Afda*
President University, Indonesia

*Corresponding Author: arifman.afda@student.president.ac.id

Abstract: *Technology is more important for business processes in every single company. Day after day, the data is bigger than before, and we need effective searching methods to match it with an input user keyword. A TF IDF method can help the user find keyword comparisons with the document in the dataset. Furthermore, a suggestion document that shows for the user matches the input user keyword using CBOW methods.*

Keywords: *TF-IDF, CBOW, Dataset, BoW*

Abstrak: Teknologi tidak lepas dengan proses bisnis di setiap perusahaan ini. Semakin hari, data yang disimpan menjadi semakin besar sehingga dibutuhkan metode pencarian yang efektif sesuai dengan keyword yang dimasukan oleh pengguna. Dengan metode TF IDF dapat membantu pengguna menemukan perbandingan kata kunci dengan dokumen yang ada di dalam dataset. Oleh sebab itu, dokumen yang diharapkan muncul oleh pengguna sesuai dengan kata kunci yang dimasukan dengan menggunakan metode kesamaan CBOW.

Kata kunci: TF-IDF, CBOW, Dataset, BoW

PENDAHULUAN

Perkembangan teknologi informasi dari hari ke hari menjadi sangat pesat. Akibat nya data yang dikelola menjadi lebih besar. Bahkan hampir semua institusi, organisasi, dan industri bisnis lainnya menyimpan datanya secara elektronik. Sebagian besar data tersebut dapat diakses melalui internet dalam bentuk perpustakaan digital, repositori, dan informasi lainnya seperti blog, media sosial, dan email. Namun data dengan jumlah yang sangat besar memunculkan masalah baru yaitu menemukan cara untuk mengetahui pola dan tren yang tepat untuk mengekstraksi informasi penting dari data dalam jumlah besar ini. Cara tradisional tidak mampu menanganinya karena memerlukan waktu dan upaya untuk mengekstrak informasi tersebut.

Text mining merupakan proses mengekstraksi dari suatu sumber data dalam rangka menemukan informasi baru. Sumber data dapat dikumpulkan melalui *information retrieval, data mining, machine learning, statistics, and computational linguistics*. Text mining juga memliki konsep yang berelasi dengan metode lain seperti *summarization, classification, clustering* dll., dapat diterapkan untuk mengekstraksi suatu informasi. *Text mining* berkaitan juga dengan *natural language* yang disimpan dalam format *semi-structured and unstructured*.

Mengumpulkan *unstructured data* dari berbagai sumber yang tersedia dalam format file yang berbeda seperti teks biasa, halaman website, file pdf, dll. Namun proses ekstraksi informasi penting dari kumpulan dokumen berbeda merupakan proses yang panjang dan membutuhkan waktu yang lama. Dengan metode yang tepat untuk *text mining*, dapat mengurangi waktu dan dalam rangka menemukan cara yang relevan untuk proses analisis dan pengambilan keputusan. Tujuan dari jurnal ini adalah untuk membandingkan dokumen yang menggunakan metode tf-idf.

Tersedia berbagai metode *text mining* yang diterapkan untuk menganalisis pola dan prosesnya. Klasifikasi dokumen (klasifikasi teks, standardisasi dokumen), pengambilan informasi (pencarian kata kunci/query dan pengindeksan), pengelompokan dokumen (pengelompokan frasa), *natural laguage processing* (*spelling correction, lemmatization, grammatical parsing, and word sense disambiguation*), information extraction (relationship extraction / link analysis), and web mining.

METODE PENELITIAN

a. Respresentasi dokumen pada vector space model

Dalam text mining, setiap dokumen direpresentasikan sebagai vektor. Elemen-elemen dalam vektor menggambarkan frekuensi dalam dokumen, dan setiap kata adalah dimensi dan dokumen pada vektor. Setiap kata dalam dokumen mempunyai bobot. Bobot ini dapat terdiri dari dua jenis: bobot lokal dan global. Jika bobot lokal digunakan, maka biasanya bobot dinyatakan sebagai Term Frequencies (TF). Jika bobot global digunakan, maka biasa dinyakan dalam Inverse Document Frekuensi (IDF), nilai IDF memberikan bobot suatu data. Sehingga pembobotan dapat menjadi lebih baik jika dilakukan dengan metode mengalikan nilai TF dengan nilai IDF, dengan mempertimbangkan informasi bobot lokal dan global. Oleh karena itu bobot total dapat dinyata sebagai = TF*IDF. Hal ini biasa disebut dengan pembobotan "TF*IDF".

Kemudian dokumen tersebut akan diproses dengan <t, w>. t1, t2, t3…. tn mewakili jumlah yang muncul pada dokumen. Dan juga variabel sebagai koordinat dimensi dinyatakan dengan N. W1, W2, W3…, Wn menjelaskan nilai yang relevan untuk koordinat. Jadi setiap dokumen (d) direlasikan dengan target sebagai vektor V(d)=(t1,w1,t2,w2,t3,w3….tn, wn). Sebelum pemrosesan data, langkah yang paling penting adalah memproses sumber data dan juga menentukan vektor. Bobot dapat digunakan sebagai kriteria pemilihan. Nilai elemen vektor Wi untuk dokumen d dihitung sebagai kombinasi statistik TF (t, d) dan DF(t). Istilah frekuensi TF (t, d) adalah jumlah kata t yang muncul dalam dokumen d. Frekuensi dokumen DF (t) meurujuk pada jumlah dokumen di mana kata t muncul setidaknya satu kali. IDF adalah frekuensi dokumen terbalik (t) dapat dihitung dengan mencari frekuensi dokumen. Log (||)/ DFtDIDFt|D| |D| adalah jumlah keseluruhan dokumen. Frekuensi dokumen IDF pada suatu kata akan bernilai rendah jika muncul di banyak dokumen dan menghasilkan nilai yang tinggi jika kata muncul hanya di satu dokumen. Nilai Wi fitur Ti untuk dokumen d kemudian dihitung sebagai hasil kali (,) ( )(2) iii W=TF td. IDF t wi disebut bobot kata ti pada dokumen d. Secara istilah kata, pembobotan kata ti merupakan istilah pengindeksan kata yang penting untuk dokumen d jika sering muncul. Kata yang muncul di banyak dokumen dinilai kurang penting dalam istilah pengindeksan karena frekuensi IDF yang rendah dan juga digunakan untuk mencari t. IDF frekuensi dokumen inverse (t) dapat dihitung dari frekuensi dokumen.

Tf(t)	=	(Jumlah t yang muncul pada dokumen)/(Total jumlah dokumen)
IDF(t)	=	log_e (Total Jumlah Dokumen/Jumlah dokumen yang terdapat kata yang dibandingkan)
W(t)	=	tf*idf

b. Pembuatan Corpus

Sebelum proses pembuatan Corpus, pengumpulan data perlu dilakukan dan dipersiapkan agar dapat dijadikan sebagai data penelitian. Persiapkan tools yang mendukung penelitian ini, dalam hal ini menggunakan googlecollab dengan bahasa pemograman Python. Pertama, mari kita mulai dengan memuat data, dapat berupa teks atau kumpulan teks (biasa disebut korpus) yang ingin kita ekstrak untuk mendapatkan informasi penting. Corpus adalah kumpulan n dokumen dengan masing-masing dokumen ini didefinisikan sebagai kumpulan m (kata atau himpunan kata)

c. Preprocessing

Selanjutnya memproses corpus terlebih dahulu dengan menghilangkan tanda baca, angka, menghapus spasi, mengubah teks menjadi huruf kecil, menghapus stopwords, stemming, dll. Hal ini perlu dilakukan karena ada beberapa karakter tidak terlalu penting untuk diproses dan beberapa juga sulit untuk diproses. Misalnya stopwords dalam bahasa Inggris seperti “the”, “is”, “myself”, “about”, “I” dll tidak akan memberikan banyak informasi tentang isi teks yang akan dibandikan. Langkah ini dilakukan untuk membersihkan data guna meningkatkan efisiensi dan akurasi hasil dengan menghilangkan kata-kata yang tidak diperlukan. Berikut adalah pre-processing yang dilakukan pada penelitian ini:

1. Convert to lower case
2. Remove punctuation
3. Remove numbers
4. Stemming
5. Stop words

d. Text feature extraction TF-IDF

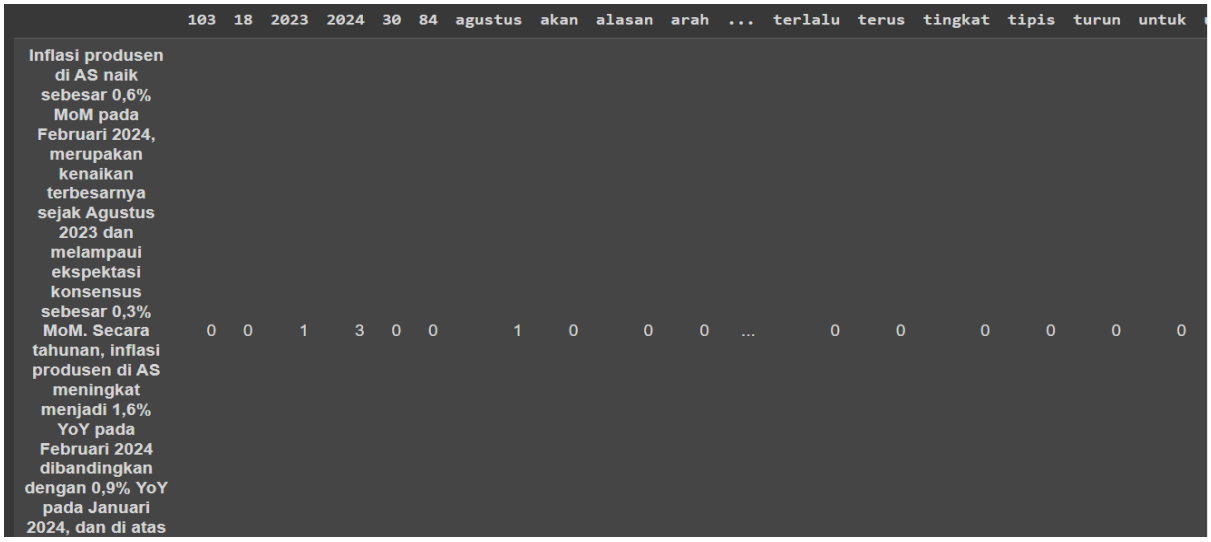
Istilah frekuensi digunakan untuk merepresentasikan informasi pada teks ruang vektor. Namun, masalah utama adalah pendekatan term-frekuensi itu sendiri bahwa pendekatan ini meningkatkan suku kata yang sering muncul dan memperkecil suku kata yang jarang muncul yang secara kontekstual lebih informatif dari pada suku kata yang berfrekuensi tinggi. Pada dasarnya kata yang sering muncul di banyak dokumen bukanlah hasil yang baik, dan benar-benar dapat diterima (setidaknya dalam banyak uji coba penelitian). Sehingga pertanyaan penting saat ini adalah mengapa pada setiap masalah klasifikasi, selalu mengedepankan untuk menemukan kata yang hampir ada di seluruh kumpulan dokumen.

Pembobotan dengan metode tf-idf hadir untuk mengatasi masalah ini. Metode tf-idf adalah metode yang menjelaskan pentingnya sebuah kata bagi sebuah dokumen dalam suatu koleksi data, dan itulah mengapa tf-idf menggabungkan parameter lokal dan global, karena mempertimbangkan tidak hanya istilah yang terisolasi tetapi juga istilah dalam koleksi dokumen. Lalu kemudian tf-idf juga dapat digunakan untuk mengatasi masalah tersebut dengan memperkecil suku kata yang sering muncul dan meningkatkan suku kata yang jarang; suatu istilah yang muncul 10 kali lebih banyak dari yang lain, tidak akan menjadi 10 kali lebih penting dari pada yang lain, itu sebabnya tf-idf menggunakan skala log untuk melakukannya proses perbandingannya.

HASIL DAN PEMBAHASAN

Langkah awal yang dilakukan pada penelitian ini adalah menyiapkan data. Dataset yang digunakan adalah dataset internal dari organisasi tempat penulis bekerja. Terdapat 400 data yang berisi tentang berita harga pasar instrumen investasi di Indonesia. Pada masing-masing dokumen terdapat ratusan kata dengan informasi penting yang dibutuhkan oleh pengguna sehingga dibutuhkan query pencarian data yang efektif dengan memanfaatkan metode TF IDF untuk mendeteksi kesamaan kata pada semua dokumen.

Data yang telah diolah dan dipersiapkan untuk penelitian ini harus melalui *preprocessing* dengan beberapa tahap *Convert to lower case, Remove punctuation, Remove numbers, Stemming dan Stop words.* Selanjutnya, data yang telah melalui tahapan *preprocessing* diubah menjadi token dengan fungsi tokenizing dan mulai diproses dengan menggunakan metod Bag Of Word (BoW) sebagaiman hasil pada gambar 1.

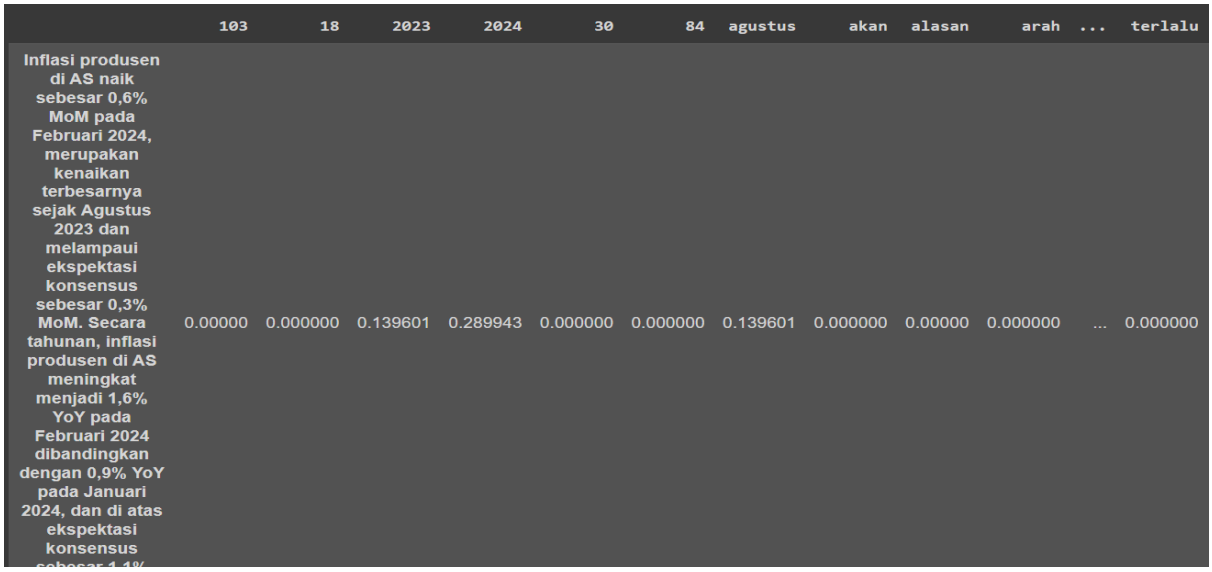


	103	18	2023	2024	30	84	agustus	akan	alasan	arah	...	terlalu	terus	tingkat	tipis	turun	untuk
Inflasi produsen di AS naik sebesar 0,6% MoM pada Februari 2024, merupakan kenaikan terbesarnya sejak Agustus 2023 dan melampaui ekspektasi konsensus sebesar 0,3% MoM. Secara tahunan, inflasi produsen di AS meningkat menjadi 1,6% YoY pada Februari 2024 dibandingkan dengan 0,9% YoY pada Januari 2024, dan di atas	0	0	1	3	0	0	1	0	0	0	...	0	0	0	0	0	0

Gambar 1. Contoh hasil salah satu dokumen dengan menggunakan metode BoW

Kolom menunjukkan setiap kata yang ada di dalam dataset dan bari menunjukkan dokumen yang ada di dalam dataset. Isian setiap kolom menjelaskan jumlah kata yang muncul pada setiap dokumen sehingga dapat dipahami dengan mudah. Contoh pada gambar 1. Kata agustus muncul 3 kali pada dokumen tersebut.

Selanjutnya data diproses menggunakan metode *Term Frequency - Inverse Document Frequency (TF-IDF).* TF IDF dapat diartikan dengan BoW * IDF dengan memanfaatkan fitur *scaling* untuk mendektsi kata yang sering muncul bisa saja menjadi kata yang terlalu penting, misalnya *stopwords.* Dengan menggunakan metode TF IDF dapat menemukan kecocokan masing-masing kata dengan dokumen yang ada di dalam dataset sesuai gambar 2.



	103	18	2023	2024	30	84	agustus	akan	alasan	arah	...	terlalu
Inflasi produsen di AS naik sebesar 0,6% MoM pada Februari 2024, merupakan kenaikan terbesarnya sejak Agustus 2023 dan melampaui ekspektasi konsensus sebesar 0,3% MoM. Secara tahunan, inflasi produsen di AS meningkat menjadi 1,6% YoY pada Februari 2024 dibandingkan dengan 0,9% YoY pada Januari 2024, dan di atas ekspektasi konsensus	0.00000	0.000000	0.139601	0.289943	0.000000	0.000000	0.139601	0.000000	0.00000	0.000000	...	0.000000

Gambar 2. Contoh hasil salah satu dokumen dengan menggunakan metode TF IDF

Sama seperti penjelasan sebelumnya mengenai gambar 2 adalah hasil dari menggunakan metode TF IDF. Kolom menunjukkan setiap kata yang ada di dalam dataset dan bari menunjukkan dokumen yang ada di dalam dataset. Isian setiap kolom menjelaskan bobot kemiripan satu kata pada suatu dokumen.

Lalu untuk mendapatkan bobot kemiripan satu kata dengan kata lainnya dapat menggunakan Word2Vec. Word2Vec adalah formula berbasis *neural network* untuk menghasilkan *word embedding*. Word2Vec terbagi menjadi dua jenis, Continuous Bag of Words (CBOW) and Skip-Gram (SG). CBOW dapat memprediksi berdasarkan kata saat ini dalam satu himpunan kata sedangkan SG dapat memprediksi kata selanjutnya dengan menggunakan kata saat ini.

```
Cosine similarity between 'sektor' and 'jasa' - CBOW :  -0.031075697
Cosine similarity between 'sektor' and 'suku' - CBOW :  -0.058002703
```

Gambar 3. Hasil Pencocokan Salah Satu Kata dengan manggunakan Metode CBOW

```
[('oleh', 0.23916706442832947),
 ('para', 0.2173747569322586),
 ('prospek', 0.19434289634227753),
 ('pertumbuhan', 0.17292821407318115),
 ('pemangkasan', 0.1697288602590561),
 ('kekhawatiran', 0.16425928473472595),
 ('4,5', 0.16269434988498688),
 ('melakukan', 0.16083787381649017),
 ('januari', 0.1601407676935196),
 ('mom', 0.15898124873638153)]
```

Gambar 4. Hasil top 10 kemiripan kata "sektor"

Pada penelitian ini, CBOW digunakan unuk menganalisa kemiripan satu kata dengan kata lainnya. Dengan mengambil salah satu kata yang ada di dalam dokumen lalu dibandingkan dengan kata lainnya seperti gambar 3. Analisa selanjutnya juga dapat dilakukan dengan mencari top 10 kata yang memiliki kesamaan dengan kata kunci yang dicari seperti gambar 4.

KESIMPULAN

Dengan menggunakan metode TF IDF dapat membantu perusahaan dalam menemukan dokumen yang diharapkan dengan kata kunci yang dimasukan. Banyaknya dokumen pada suatu perusahaan tidak menjadi halangan karena dengan menggunakan metode CBOW dapat menemukan kemiripan kata di dalam dataset dengan kata kunci yang dimasukan ke dalam sistem.

REFERENSI

Dr.M. Umadevi. (2020). *Document Comparison Based On Tf-Idf Metric.* Guntur: IRJET

Feinerer. (2008). *An Introduction to Text Mining in R*

Dan Munteanu. (2007). *Vector Space Model for Document Representation in Information Retrieval*. University of Galati Fascicle

Hinrich Schutze. (2011). *Introduction to Information Retrieval. Institute for Natural Language Processing*. University of Stuttgart

Mazid, Mohammad. (2009). *A comparison between rule based and Association rule mining algorithm*. Queensland