ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

OPTIMASI MESIN PENCARI BUKU FIKSI BERDASARKAN PADA SEMANTIK IMPRESI

Rengga Asmara✉, Nur Rasyid Mubtada'i, Varidh Bimantara
Departemen Teknik Informatika dan Komputer, Politeknik Elektronika Negeri Surabaya, Indonesia
Email: rengga@pens.ac.id

DOI: https://doi.org/10.46880/jmika.Vol5No1.pp1-8

ABSTRACT

Fiction books are one of the most popular types of books in Indonesia. There are five most popular genres in fiction books, namely fantasy, mystery, romance, sci-fi, and thriller. Each genre gives a different impression and special interest for readers. It has become a common habit when people choose a fiction book based on the title, author, or publisher of the book. However, it does not provide precise search results. In this final project, an application system was developed to find out fiction books based on semantic impressions on the cover of the fiction book. The impression on each book cover is obtained through a survey of fiction book lovers in Indonesia. To get the results of the closeness between the user search and the impression survey data obtained through text mining, as well as the cosine similarity algorithm to calculate the most precise proximity value to the impression the user expects. The results of this system display a fiction book that has a closeness value with an error rate of 3.93% based on the impression expected by the user.

Keyword: Fiction, Impression, Text Mining, Cosine Similarity

ABSTRAK

Buku fiksi merupakan salah satu dari sekian jenis buku yang terpopuler di Indonesia. Ada lima genre paling popular didalam buku fiksi, yakni fantasy, misteri, romance, sci-fi dan thiller. Masing-masing genre memberikan kesan berbeda dan peminat tersendiri bagi pembaca. Sudah menjadi kebiasaan umum ketika orang memilih buku fiksi berdasarkan judul, penulis maupun penerbit buku tersebut. Namun hal tersebut belum memberikan hasil pencarian yang presisi. Pada proyek akhir ini dikembangkan sebuah system aplikasi untuk mengetahui buku fiksi berdasarkan semantic impresi yang terdapat pada sampul buku fiksi tersebut. Impresi pada setiap sampul buku didapatkan melalui survey kepada pecinta buku fiksi yang ada di Indonesia. Untuk mendapatkan hasil kedekatan antara pencarian user dengan data surver impresi didapatkan melalui teks mining, serta algoritma cosine similarity untuk menghitung nilai kedekatan paling presisi dengan impresi yang diharapkan user. Hasil dari system ini menampilkan buku fiksi yang memiliki nilai kedekatan paling presisi dengan impresi yang diharapkan oleh pengguna.

Kata Kunci: **Fiksi, Impresi, Teks Mining, *Cosine Similarity***

PENDAHULUAN

Fiksi merupakan suatu cerita atau latar yang berasal dari imajinasi, bukan secara nyata berdasarkan sejarah atau fakta (Adams, 2018). Awalnya fiksi lebih sering digunakan untuk bentuk sastra naratif seperti novel, novella, cerita pendek, dan sandiwara. Kini fiksi dapat divisualisasikan dalam berbagai format antara lain: tulisan, film, pertunjukan langsung, animasi, acara televisi, permainan video, dan permainan peran. Secara tradisional, fiksi mencakup cerita pendek, novel, fable, mitos, legenda, epik dan puisi naratif, dongeng, sandiwara (termasuk opera, teater musikal, permainan boneka, drama, serta berbagai jenis tarian teatrikal). Akan tetapi fiksi juga dapat mencakup kartun animasi, buku komik, anime, film, *stop motion,* manga, program radio, permainan video, program televisi (komedi dan drama), dan lain sebagainya.

Buku fiksi merupakan buku yang paling diminati di Indonesia. Hal ini dibuktikan oleh penelitian yang dilakukan picodi.com, hasil dari penelitian tersebut menunjukkan bahwa jumlah responden yang memilih jenis buku ini mencapai 75 persen dimana angka tersebut jauh lebih tinggi dari buku non-fiksi (41 persen), buku bisnis (33 persen), buku sains popular (31 persen), buku literatur hobi (24 persen), dan buku literatur sains serta buku teks (22 persen) (Iswara, 2019).

Saat ini ketika seseorang melakukan pencarian buku fiksi masih mengalami kesulitan dikarenakan kurangnya presisi dari hasil buku fiksi yang di cari.

Sejauh saat ini kebanyakan orang mencari buku fiksi berdasarkan judul, pengarang, maupun penerbit. Namun hal tersebut terkadang masih belum mewakili cerita yang diinginkan oleh pembaca. Ada permasalahan jika seorang pembaca merupakan orang yang baru ingin membeli buku fiksi, belum memiliki pengalaman dalam memilih atau belum banyak mengetahui tentang judul atau pengarang buku fiksi mana yang biasanya memiliki cerita menarik untuk dibaca.

Dilain sisi sampul buku fiksi memberikan kesan terhadap isi dari keseluruhan buku fiksi tersebut. Sampul pada buku dapat mewakili synopsis cerita yang dituangkan dalam bentuk gambar maupun typografi pada sampul buku fiksi. Sampul buku merupakan hal pertama yang akan dilihat seorang pembaca, begitu juga dengan buku fiksi. Kalimat "Don't judge the book by its cover" sepertinya bukan lagi menjadi pernyataan yang tepat untuk menggambarkan proses seorang pembaca dalam memilih buku yang akan ia baca. Pada kenyataannya, sampul buku memiliki peran yang penting dalam proses penerbitan sebuah buku, dimana di dalamnya mengandung penempatan obyek ilustrasi yang tepat, pemilihan font yang sesuai, hingga penggunaan warna yang menggambarkan emosi yang ingin penulis sampaikan lewat bukunya, dan yang penting adalah koherensi dengan cerita. Sebagai produk komoditas, buku dengan tampilan yang menarik tentunya akan meningkatkan kemampuan penjualannya.

Melalui jurnal dari Tess Adams "*Judging a Book By Its Cover: Are First Impressions Accurate?"* mengatakan bahwasanya ada korelasi yang timbul dari sebuah sampul buku dengan impresi. Melalui sampul buku, bisa didapatkan sebuah informasi tersirat (Adams, 2018). Sedangkan menurut jurnal Assist. Prof. Dr. Aslı Sezgin "The First Impression of "Best Selling" Agenda Books: Semiotic Analysis of Book Covers" menyampaikan dari hasil penelitian yang dilakukannya menyebutkan bahwasanya pada setiap sampul buku memiliki impresi masing-masing (Sezgin, 2014). Hal ini bisa menjadi sebuah rekomendasi cara baru bagi seorang pembeli buku dalam mencari buku yang dinginkan.

Saat ini penelitian mengenai sistem pencarian buku sudah banyak dilakukan dengan berbagai metode. Sedangkan proyek akhir ini berfokus pada pencarian buku fiksi saja dengan menggunakan metode semantic impresi (Zebua & Mustikasari, 2017). Salah satu bidang yang dapat diselesaikan dengan adanya penelitian ini adalah strategi markting. Pelanggan dapat menemukan buku fiksi yang di inginkannya melalui pendekatan yang berbeda, yakni melalui impresi. Dengan harapan dapat menemukan buku fiksi yang sesuai dengan keinginan pelanggan, yang juga akan mempengaruhi kenaikan penjualan.

KAJIAN LITERATUR

Fiksi

Karya fiksi merupakan sebuah karya yang menceritakan sesuatu yang bersifat rekaan, tidak nyata, khayalan, sesuatu yang tidak ada dan terjadi sungguh-sungguh sehingga seseorang tidak perlu mencari kebenarannya pada dunia nyata. Kebenaran sebuah cerita fiksi yang baik adalah kemungkinan, probabilitas atau kemasukakalannya.

Sesuai dengan nama dan sitatnya, cerita fiksi adalah karya kreatif-imajinatif yang tidak menyaratkan adanya verifikasi dengan kenyataan untuk memiliki kebenaran yang masuk akal. Bahkan sekalipun cerita fiksi salah mengutip fakta realitas, jika pengisahannya dapat membungkus kesalahan itu dengan cerita yang masuk akal, itu tidak akan merusak cerita.

Cerita masih dapat diterima oleh pembaca karena ia membawa alur logika sendiri. Kita pernbaca tentu menginginkan bahwa cerita yang dikisahkan itu benar. Namun, kebenaran itu hanya dapat tarjadi dalam dunia cerita itu yang dilakoni oleh tokoh dan peristiwa yang sengaja dibuat dan dikembangkan oleh penulis yang kemudian tercipta kembali dalarn imajinasi pembaca.

Impresi

Sebagai seorang manusia, kita dapat dengan mudah melihat atau bereaksi terhadap seseorang, sesuatu, atau peristiwa. Persepsi yang diberikan merupakan gambaran dari proses observasi dan pengalaman yang telah kita alami. Ada pepatah Inggris yang mengatakan bahwa, *"Don't jugde the book by its cover",* kata-kata ini biasanya sering dihubungkan dengan persepsi dan pembentukan impresi seseorang. Apa yang indah di luar tidak selalu berarti di dalam, begitu pula sebaliknya. Ini menunjukkan bahwa apabila persepsi kita tidak dijelaskan dengan benar, maka akan berbeda dari situasi sebenarnya. Faktanya, pandangan yang kita berikan terkadang bisa salah, tersasar dan sebagainya.

Seperti yang kita ketahui Bersama, berdasarkan beberapa definisi presepsi, persepsi merupakan proses dasar yang penting untuk mengenali dan memahami orang lain atau peristiwa dalam kehidupan sehari-hari. Dalam konteks kita sebagai seorang kaunselor, ilmu di bidang ini sangat penting dan berguna untuk mempraktikkan pekerjaannya di sekolah nanti. Untuk

menjadi kaunselor yang mudah diingat, individu itu perlulah peka dalam membuat persepsi. Saat membentuk sebuah persepsi, petanda yang terlihat adalah petanda non-verbal. Ini bisa dilihat melalui ekspresi muka, kontak mata, gerak gerik tubuh dan sebagainya.

Pembentukan impresi merupakan proses seseorang dalam mengintegrasikan berbagai maklumat dari berbagai sumber untuk membentuk tanggapan, penilaian dan penjelasan menyeluruh kepada orang lain. Perlu ditekankan bahwas impresi pertama sangat penting untuk menentukan hubungan seterusnya.

Semantik

Makna semantik merupakan salah satu cabang linguistik yang mempelajari makna yang terkandung dalam bahasa, kode, atau jenis representasi lainnya. Singkatnya, semantik adalah strudi tentang makna. Istilah semantik mengacu pada berbagai ide yang sangat teknis dan populer. Biasanya sering digunakan dalam bahasa sehari-hari untuk mengungkapkan masalah pemahaman kata atau pilihan arti. Dalam jangka waktu yang panjang, masalah pemahaman seperti ini telah menjadi subjek banyak pertanyaan formal, terutama di bidang semangtik formal.

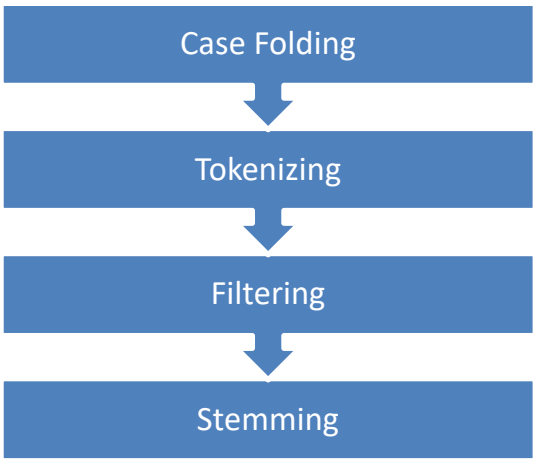
Dalam bidang linguistik, semantik adalah studi tentang interpretasi tanda-tanda atau simbol yang digunakan untuk agen atau masyarakat dalam situasi dan konteks tertentu (Khoctitah & Rahmawati, 2015). Dalam hal ini, suara, ekspresi wajah, bahasa tubuh, dan proxemics memiliki semantik konten (bermakna), dan setiap semantik mencakup beberapa cabang penelitian. Dalam bahasa tertulis, hal-hal seperti struktur ayat dan tanda baca memiliki kandungan semantik, dan bentuk bahasa lain juga memiliki konten semantik lainnya.

Semantik dalam linguistik merupakan sub bidang yang didedikasikan untuk studi makna, seperti yang melekat pada tingkat kata, frasa, kalimat, dan unit wacana yang lebih besar (disebut teks). Perhatian utamanya adalah bagaimana makna tersebut melekat pada potongan teks yang lebih besar, mungkin sebagai akibat dari komposisi dari unit yang lebih kecil dari makna. Secara tradisional, semantik mencakup studi tentang makna dan referensi denotatif, kondisi kebenaran, struktur argumen, peran tematik, analisis wacana, dan hubungan semua ini untuk sintaks.

Padahal, struktur dan fungsi sangat erat kaitannya dengan semantik. Dalam artian struktur tanpa makna dan makna tanpa struktur tidak mungkin ada. Sehingga bentuk atau struktur, fungsi dan makna merupakan satu kesatuan dalam meneliti atau mengkaji unsur-unsur bahasa.

Teks Mining

Data yang terstruktur dengan baik secara otomatis dapat memfasilitasi proses komputerisasi. Dalam Text Mining, informasi yang akan digali mengandung informasi dengan struktur yang tidak beraturan. Maka dari itu, diperlukan suatu proses untuk mengubah data menjadi terstruktur sesuai dengan kebutuhan proses data mining, biasanya nilai numerik. Proses ini disebut "Text Preprocessing". Setelah data menjadi terstruktur dan digunakan dalam bentuk numerik maka dapat digunakan sebagai sumber data yang dapat diolah lebih lanjut.



Gambar 1. Tahapan teks mining

Case folding adalah tahapan untuk mengubah semua huruf dalam dokumen menjadi huruf kecil. Karakter yang diterima hanya huruf A sampai dengan Z. Karakter selain huruf tersebut dihilangkan dan dianggap delimiter. Tahap selanjutnya adalah Tokenizing yaitu tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Kemudian Tahap filtering adalah tahap mengambil kata - kata penting dari hasil token. Bisa menggunakan algoritma stoplist (membuang kata yang kurang penting) atau wordlist (menyimpan kata penting). Stoplist / stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words. Tahap terakhir adalah stemming yaitu tahap mencari root kata dari tiap kata hasil filtering. Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama.

Pembobotan Tf-Id

Skema pembobotan jangka TF-IDF adalah skema yang paling berhasil dan banyak digunakan untuk menetapkan bobot jangka ke dokumen besar.

Saaat mencari informasi dari koleksi dokumen yang heterogen, pembobotan term perlu

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

dipertimbangkan. Sebuah term dapat berupa kata, frase atau unit indeks lainnya dalam dokumen yang dapat digunakan untuk mencari konteks dokumen. Oleh karena itu, untuk setiap kata diberikan insikator, yaitu *term weight* (Informatikalogi, 2017).

TF (Term Frequency) merupakan frekuensi kemunculan sebuah term dalam dokumen terkait. Semakin banyak term tersebut muncul (TF tinggi), semakin besar pula bobotnya atau nilai penerapannya.

Algoritma Cosine Similarity

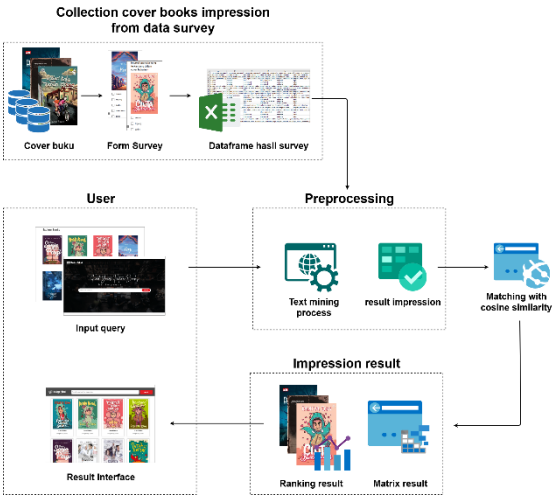
Algoritma cosine similarity merupakan ukuran kesamaan antara dua vektor bukan-nol dari ruang produk dalam yang mengukur kosinus sudut di antara mereka. Nilai sudut cosinus antara dua vektor menentukan kesamaan dua buah objek yang dibandingkan dimana nilai terkecil adalah 0 dan nilai terbesar adalah 1.

Cosine similarity mengukur kesamaan antara dua vektor ruang produk dalam. Ini diukur oleh cosinus sudut antara dua vektor dan menentukan apakah dua vektor menunjuk ke arah yang kira-kira sama. Ini sering digunakan untuk mengukur kesamaan dokumen dalam analisis teks. Sebuah dokumen dapat diwakili oleh ribuan atribut, masing-masing merekam frekuensi kata tertentu (seperti kata kunci) atau frasa dalam dokumen. Dengan demikian, setiap dokumen adalah objek yang diwakili oleh apa yang disebut vektor frekuensi-istilah.

Cosine similarity umumnya digunakan sebagai metrik untuk mengukur jarak ketika besarnya vektor tidak masalah. Ini terjadi misalnya ketika bekerja dengan data teks yang diwakili oleh jumlah kata. Kita dapat mengasumsikan bahwa ketika sebuah kata (mis. Sains) muncul lebih sering di dokumen 1 daripada di dokumen 2, dokumen 1 itu lebih terkait dengan topik sains. Namun, bisa juga karena kami bekerja dengan dokumen yang panjangnya tidak merata (artikel Wikipedia misalnya). Kemudian, sains mungkin terjadi lebih banyak dalam dokumen 1 hanya karena itu jauh lebih lama daripada dokumen 2. Kesamaan cosinus mengoreksi untuk ini.

METODE PENELITIAN

Untuk membangun sistem pencarian buku fiksi bedasarkan semantic impresi ini penulis menggunakan alur proses seperti yang ditunjukkan pada gambar 2.

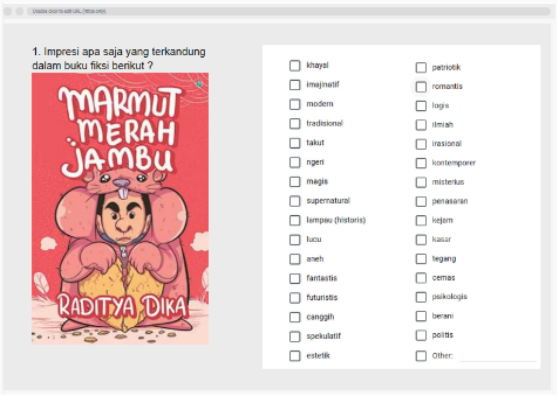


Gambar 2. Desain Sistem

Pengumpulan Impresi Sampul Buku Dari Survei

Pada proses ini bermula dari data yang telah dikumpulkan berupa beberapa cover buku fiksi. Dari cover buku fiksi tersebut akan dibuat form survey, dimana isi dari form survey tersebut adalah cover buku fiksi dan pilihan multiple choice impresi. Form survey ini nantinya akan di sebar kepada sepuluh komunitas pecinta buku yang ada pada facebook.

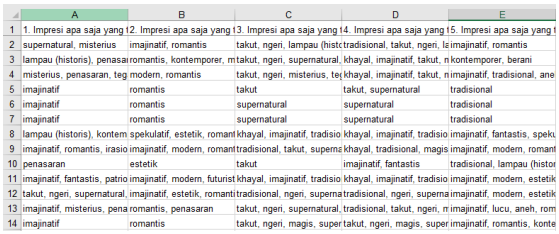
Tahapan pertama dari pengumpulan data survey disini adalah dari surveyor nantinya akan melihat dari masing-masing buku fiksi yang ada pada form survey, dalam proses ini dicontohkan pengambilan data survey pada buku berjudul "marmut merah jambu" seperti pada gambar 3. Untuk tahap selanjutnya surveyor akan memilih impresi apa saja yang terkandung dalam desain cover buku fiksi tersebut. Pemilihan impresi dalam satu buku dapat diisi lebih dari satu. Data survey tersebut akan dikumpulkan dalam bentuk matriks.



Gambar 3. Form penilaian impresi cover

Setelah memperoleh data impresi dari form survey, hasil dataframe tersebut akan disimpan dalam bentuk excel seperti pada gambar 3. Dan selanjutnya akan

dilakukan pengolahan dataframe sehingga dapat diolah kedalam sistem.



	A	B	C	D	E
1	1. Impresi apa saja yang t	2. Impresi apa saja yang t	3. Impresi apa saja yang t	4. Impresi apa saja yang t	5. Impresi apa saja yang t
2	supernatural, misterius	imajinatif, romantis	takut, ngeri, lampau (histc	tradisional, takut, ngeri, la	imajinatif, romantis
3	lampau (historis), penasa	romantis, kontemporer, m	takut, ngeri, supernatural,	khayal, imajinatif, takut, n	kontemporer, berani
4	misterius, penasaran, teg	modern, romantis	takut, ngeri, misterius, te	khayal, imajinatif, takut, n	imajinatif, tradisional, ane
5	imajinatif	romantis	takut	takut, supernatural	tradisional
6	imajinatif	romantis	supernatural	supernatural	tradisional
7	imajinatif	romantis	supernatural	supernatural	tradisional
8	lampau (historis), kontem	spekulatif, estetik, romant	khayal, imajinatif, tradisio	khayal, imajinatif, tradisio	imajinatif, fantastis, speku
9	imajinatif, romantis, irasio	imajinatif, modern, romant	tradisional, takut, supernat	khayal, tradisional, magis	imajinatif, modern, romant
10	penasaran	estetik	takut	imajinatif, fantastis	tradisional, lampau (histor
11	imajinatif, fantastis, patrio	imajinatif, modern, futurist	khayal, imajinatif, tradisio	khayal, imajinatif, tradisio	imajinatif, modern, estetik
12	takut, ngeri, supernatural,	imajinatif, estetik, romanti	tradisional, ngeri, supernat	tradisional, ngeri, supernat	imajinatif, modern, estetik
13	imajinatif, misterius, pena	romantis, penasaran	takut, ngeri, supernatural,	tradisional, takut, ngeri, n	imajinatif, lucu, aneh, rom
14	imajinatif	romantis	takut, ngeri, magis, super	takut, ngeri, magis, super	imajinatif, romantis, konte

Gambar 4. Hasil penilaian survey impresi

Preprocessing Query

Pada tahap ini, queri yang di inputkan oleh user akan diproses pada teks mining. Teks mining disini akan memproses queri yang di inputkan oleh user dan dataframe impresi yang diperoleh dari survey. Dalam proses teks mining ada beberapa tahapan yang dilakukan yakni:

a. Tokenizing
   Pada tahap tokenizing, dataframe impresi dan queri impresi dari user akan dijadikan satu dataframe. Dataframe pada impresi survey hanya akan diambil impresinya saja. Queri impresi yang di inputkan oleh user akan masuk kedalam satu dataframe dengan letak baris paling bawah. Semua kata-kata akan dibaca perbaris, kata-kata tersebut akan dipisah dengan karater spasi sebagai pemisahnya. Sehingga hasil yang keluar dari proses filtering ini adalah setiap kata akan terpisah pada setiap baris.
b. Filtering
   Tahap filtering merupakan tahap pengambilan kata-kata penting menggunakan algoritma stopwords. Pada setiap baris hasil output dari tokenizing akan dicek setiap kata, apabila kata tersebut terdapat pada dataframe stopword maka kata tesebut akan dihapus, akan tetapi apabila kata tersebut tidak ada pada dataframe stopword maka kata tersebut akan disimpan pada table koleksi.
c. Stemming
   Pada tahap stemming disini merupakan tahap untuk mendapatkan kata dasar dari setiap kata hasil filtering. Proses dilakukan dengan menganalisa setiap baris, jika kata pada setiap baris tersebut mengandung awalan atau akhiran maka kata awalan atau akhiran tesebut akan di hapus, sehingga output dari proses ini adalah kata dasar dari setiap kata pada masing-masing baris.
d. Pembobotan Tf-Idf
   Pada tahap pembobotan Tf-Idf disini, semua kata output dari proses teks mining akan dihitung banyaknya pada setiap baris. Baris disini dimaksudkan setiap kepemilikan sampul buku. Setiap baris akan dihitung nilai impresi yang diperoleh oleh masing-masing sampul buku. Hasil pembobotan dari table Tf diatas selanjutnya akan dikalikan dengan nilai Idf.

Matching With Cosine Similarity

Pada tahap ini, matriks dari impresi data survey akan dilakukan matching dengan kata impresi queri yang di inputkan oleh user. Proses matching disini menggunakan metode cosine similarity. Jadi dari queri akan di hitung nilai kedekatannya dengan nilai yang ada pada matriks impresi. Dari proses pencarian nilai cosine similarity diatas maka didapatkan masing-masing nilai cosine similarity dari tiap sampul buku.

Impression Result

Hasil dari proses matcing dengan menggunakan metode cosine similarity adalah matriks dua dimensi, dimana pada masing-masing fitur menyimpan nilai kedekatan antara queri dengan impresi pada cover buku. Hasil matriks tersebut akan dilakukan sorting berdasarka nilai tertinggi, dan kemudian ditampilkan pada layar user. Tampilan output yang dikirim pada user adalah berupa list cover buku fiksi yang memiliki kedekatan dengan queri yang di inputkan oleh user.

HASIL DAN PEMBAHASAN

Penulis akan menguji system dengan beberapa tahap ujian yang pada akhirnya mendapatkan ketepatan query pencarian dengan data hasil survey. Tahap pengujian pertama dilakukan untuk mengetahui data survey apakah sesuai dengan impresi pada genre buku. Pada tahap ini hasil survey yang dilakukan kepada komunitas pecinta buku fiksi yang ada di Indonesia. Hasil data survey menunjukkan nilai jumlah impresi dari sampel 100 buku yang telah disebar untuk mendapatkan penilaian impresi. Data tersebut telah sesuai dengan impresi yang ada pada genre buku itu sendiri.

Pengujian text mining dilakukan untuk mengetahui apakah system pada text mining yang meliputi tokenizing, filtering dan stemming dapat berjalan serta menghasilkan output yang sesuai. Pengujian pada tahap Data Processing terdiri dari beberapa tahap, seperti yang ditunjukkan pada skema berikut.

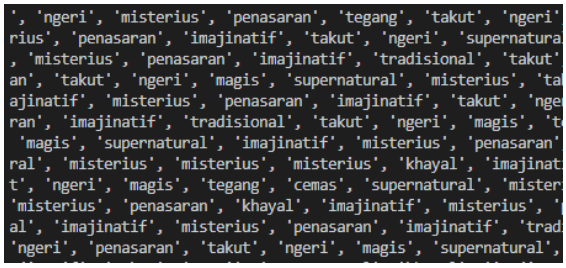
1. Tahap tokenizing, data teks artikel dibersihkan dari karakter yang mengandung selain huruf abjad (A-Z) kemudian dipecah menjadi array kata.
2. Tahap token filtering, array kata disaring dengan menghilangkan stopword.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

3. Kemudian pada tahap stemming & lemma, array kata impresi ditransformasikan menjadi kata dasar dengan menghilangkan imbuhan dan menyempurnakan kata agar sesuai dengan kata dasar.
4. Kemudian pada tahap term frequency, data kata dasar ditransformasi menjadi tabel yang memuat kata dan jumlah (frekuensi) kata dalam satu buku.
5. Setelah itu, pada tahap data filtering, tabel term frequency disaring menjadi keyword. Penentuan keyword dipilih dari kata yang memiliki jumlah (frekuensi) yang melebihi *threshold*. Penentuan *threshold* (*t*) didapatkan dari jumlah TF maksimal (*max_TF*) dibagi dua.

Tokenizing

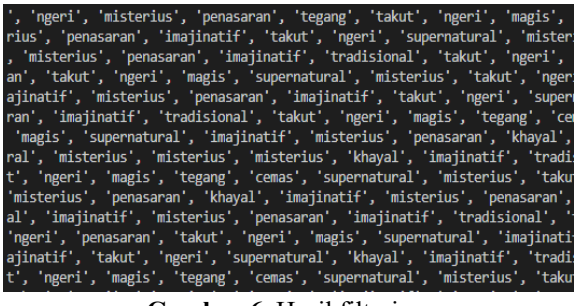
Pengujian pertama dari Data Processing adalah tahap tokenizing. Tahap tokenizing menghilangkan karakter yang mengandung selain huruf abjad (A-Z) dan kemudian dipecah menjadi array kata seperti yang ditunjukkan pada gambar 5. Tahap ini teks dipecah menjadi array kata berdasarkan spasi. Selain itu, tahap ini juga menghilangkan teks yang berupa angka dan tanda baca.



Gambar 5. Hasil tokenizing

Filtering

Pada tahap ini, hasil output dari tokenizing akan dihilangkan kata-kata yang tidak penting (kata depan, kata ganti, dll) dengan menggunakan algoritma stopwords. Untuk hasil dari filtering dapat dilihat pada gambar 6. *Stopword* perlu dihilangkan karena kata kata yang terkandung dalam *stopword* merupakan kata yang umum muncul dalam bahasa Indonesia.



Gambar 6. Hasil filtering

Stemming

Pada tahap ini, hasil output dari filteirng akan dilanjutkan dengan proses untuk menghilangkan imbuhan dan akhiran. Dan hasil dari proses stemming dapat dilihat pada gambar 7.

Gambar 7. Hasil stemming

Perhitungan Term Frequency

Pada tahap selanjutnya, output dari proses text mining dilakukan perhitungan Tf untuk mendapatkan pembobotan nilai dari setiap impresi dari masing-masing dokumen. Percobaan yang dilakukan adalah membentuk tabel Term Frequency (TF), seperti yang ditunjukkan pada Tabel 1. Tabel TF memuat seluruh kata yang ada pada suatu buku dan frekuensinya dalam suatu buku. Berikut merupakan contoh hasil perhitungan Tf untuk buku berjudul "Kisah Tanah Jawa".

Tabel 1. Nilai Term Frequency

No	Impresi	Nilai Tf
1	Aneh	0
2	Berani	0
3	Canggih	0
4	Cemas	16
5	Estetik	0
6	Fantastis	0
7	Futuristis	0
8	Historis	47
9	horror	8
10	ilmiah	0
11	imajinatif	16
12	irasional	8
13	kasar	8
14	kejam	8
15	khayal	16
16	kontemporer	16
17	lampau	46
18	logis	0
19	lucu	0

20	magis	18
21	misterius	45
22	modern	0
23	ngeri	53
24	patriotik	0
25	penasaran	23
26	politis	0
27	psikologis	0
28	romantis	0
29	sedih	0
30	spekulatif	0
31	supernatural	57
32	takut	66
33	tegang	31
34	tradisional	38

Searching merupakan fitur utama yang dimiliki oleh aplikasi ini, searching digunakan dalam pencarian dan identifikasi informasi pada database yang sesuai dengan kata kunci yang diinputkan oleh pengguna. Pengujian Searching dilakukan dengan beberapa tahap, tahap yang pertama user harus memasukkan kata kunci(query), tahap kedua adalah melakukan proses text mining pada kata kunci, tahap ketiga adalah melakukan perhitungan nilai kemunculan / frekuensi tiap kata pada query, dan tahap terakhir, query akan dilakukan perhitungan kemiripan dengan metode *cosine similarity* dan akan didapatkan artikel beserta nilai kemiripan sepuluh teratas. Berikut percobaan yang telah dilakukan

Tahap pertama user harus memasukkan kata kunci(query), pada percobaan kali ini user memasukkan query yaitu "takut ngeri". Tahap kedua dilakukan text mining pada kanca kunci diatas, dengan beberapa tahapan yaitu tokenizing, filtering, dan stemming. Tahap tokenizing menghilangkan karakter yang mengandung selain huruf abjad (A-Z) dan kemudian dipecah menjadi array kata. Proses stemming menghilangkan kata imbuhan dan kemudian mengubahnya menjadi kata dasar. Proses filtering yaitu menghilangkan kata kata yang terkandung dalam *stopword*. Pada proses tokenizing query "takut ngeri" dipecah per kata menjadi 'takut, 'ngeri' dan tidak ada karakter angka atau tanda baca.

Proses filtering pada query, tidak ada kata yang dihapus, karena tidak ada kata yang sering muncul pada KBBI. Proses stemming pada query diatas hasilnya adalah tetap karena tidak terdapat kata yang berimbuhan. Tahap ketiga adalah membentuk tabel Term Frequency (TF), perhitungan seluruh kata yang ada pada *keyword* tersebut dan frekuensinya tiap kata.

Proses selanjutnya merupakan pengujian kedekatan dilakukan dengan menggunakan kedekatan cosine similarity untuk mendapatkan nilai paling besar atau berarti paling dekat dengan query impresi yang di inputkan oleh user. Hasil perhitungan cosine similarity selanjutnya akan disorting berdasarkan nilai tertinggi, untuk mendapatkan buku mana saja yang memiliki nilai kedekatan paling tinggi dengan queri impresi yang di inputkan oleh user. Hasil output dari pencarian queri "takut ngeri" terlihat pada tabel 6 berikut.

Tabel 2. Hasil cosine similarity impresi takut ngeri

No	Judul Buku	Nilai Cosine Similarity
1	Misteri Rumah Bu Sri	0.86829623149961
2	Misteri Rumah Kosong	0.78944992120459
3	Danur Gerbang Dialog	0.76623830396647
4	Rahasia Hujan	0.75584611840887
5	Diambang Kematian	0.75065202465483
6	Kutukan Hantu Opera	0.70863003565807
7	Perburuan Dalam Kegelapan	0.69288896282924
8	Malaikat Berhati Gelap	0.68378384462997
9	Rahasia Gelap	0.66709007654534
10	Misteri Taman Berhantu	0.66020153919085

Dari hasil seperti pada tabel 2 diatas dapat disimpulkan bahwa untuk percobaan inputan dengan dua impresi "takut ngeri" berjalan dengan baik, serta memberikan output yang sesuai dengan urutan nilai cosine secara descanding.

Disamping itu juga dilakukan percobaan atas 33 impresi. 33 impresi didapatkan dari proses survey data yang sebelumnya telah kami lakukan. Dari hasil percobaan pencarian 33 impresi, pada setiap impresi memiliki error rate masing-masing. Dari nilai error rate tersebut dapat dilihat pada tabel 3 di bawah ini:

Tabel 3. Nilai error rate

No	Impresi	Error Rate
1	Aneh	0%
2	Berani	0%
3	Canggih	50%
4	Cemas	0%
5	Estetik	0%
6	Fantastis	0%
7	Futuristis	0%
8	Historis	0%
9	Ilmiah	0%
10	Imajinatif	0%
11	Irasional	0%
12	Kasar	0%
13	Kejam	0%
14	Khayal	0%
15	Kontemporer	0%
16	Lampau	0%
17	Logis	0%
18	Lucu	0%
19	Magis	0%
20	Misterius	0%
21	Modern	0%
22	Ngeri	0%
23	Patriotik	0%
24	Penasaran	0%
25	Politis	0%
26	Psikologis	0%
27	Romantis	0%
28	Sedih	80%
29	Spekulatif	0%
30	Supernatural	0%
31	Takut	0%
32	Tegang	0%
33	Tradisional	0%

Berdasarkan pada tabel 3. diatas hampir tidak ada nilai error rate kecuali pada dua impresi yakni sedih dan canggih. Hal tersebut terjadi dikarenakan memang dari data survei hampir tidak ada atau sangat sedikit yang memberikan penilaian impresi sedih dan canggih pada buku. Dari data pada tabel 3 dapat disimpulkan untuk nilai rata –rata error rate dari keseluruhan percobaan pada masing – masing impresi adalah sebesar 3,93%.

KESIMPULAN

Berdasarkan tahapan-tahapan yang telah dilakukan untuk mencari buku fiksi berdasarkan semantik impresi menghasilkan hasil yang cukup baik dengan nilai rata-rata error rate sebesar 3,39%. Nilai ini didapatkan dari percobaan semua pencarian buku fiksi terhadap impresi data yang kami peroleh melalui survei, yakni terdapat 33 impresi.

Saran untuk penelitian selanjutnya adalah pengembang dapat lebih lanjut untuk dapat membuat sistem pencarian berdasarkan semantik ini dengan data impresi yang lebih banyak atau dengan data impresi yang realtime.

DAFTAR PUSTAKA

Adams, T. (2018). Judging a Book By Its Cover: Are First Impressions Accurate? *Journal of Architectural Education*, *72*(1), 180–181. https://doi.org/10.1080/10464883.2018.1412204

Informatikalogi. (2017). Pembobotan Kata atau Term Weighting TF-IDF.

Iswara, A. J. (2019). Jenis Buku Apa yang Paling Laris di Indonesia.

Khoctitah, S., & Rahmawati, N. (2015). *Aplikasi Pencarian Buku Berbasis Web Semantik Untuk Perpustakaan Universitas Muhammadiyah Yogyakarta*. (December), 2–4.

Sezgin, A. (2014). *The First Impression of "Best Selling" Agenda Books: Semiotic Analysis of Book Covers*. Osmaniye: Osmaniye Korkut Ata University.

Zebua, J., & Mustikasari, M. (2017). *Aplikasi Pencarian Buku Berbasis Web Semantik Untuk Perpustakaan SMK Yadika 7 Bogor*. Universitas Gunadarma.