Analisis Sentimen Menggunakan Algoritma K-Means untuk Mengetahui Kalimat Positif maupun Negatif pada Buletin APTIKOM

Anton Romadoni Junior
Universitas Buana Perjuangan
Karawang, Indonesia
if17.antonjunior@mhs.ubpkarawang.ac.id

Hanny Hikmayanti Handayani
Universitas Buana Perjuangan
Karawang, Indonesia
hanny.hikmayanti@ubpkarawang.ac.id

Anis Fitri Nur Masruriyah
Universitas Buana Perjuangan
Karawang, Indonesia
anis.masruriyah@ubpkarawang.ac.id

Abstract—
Kurangnya pemahaman literasi membuat masyarakat mengalami kesulitan dalam mengeksplorasi pengetahuan. Sehingga, Asosiasi Pendidikan Tinggi Ilmu Komputer (APTIKOM) membuat media informasi secara daring yang dapat menjelaskan secara detail terhadap pemahaman masyarakat yang berbeda - beda. Media tersebut diberi nama Buletin APTIKOM yang terbit setiap bulan sejak tahun 2020. Melalui buletin ini diharapkan dapat memberikan pengetahuan terhadap masyarakat yang awam dalam teknologi terkini. Namun, dari penerbitan buletin pada aplikasi buletin APTIKOM, tidak adanya pengecekan mengenai sentimen penulisan berupa kalimat positif maupun negatif. Maka dari itu, dibuatnya penelitian ini adalah untuk mencari kalimat positif maupun negatif dalam buletin APTIKOM. Data yang digunakan merupakan data yang diambil dari Buletin APTIKOM. Cara untuk mengetahui sentimen analisis pada buletin tersebut terdapat beberapa proses yaitu dengan menggunakan *Text Processing*, *Term Frequency – Inverse Document Frequency* (TF-IDF), algoritma K-Means dan *Sum of Square Error* (SSE) sebagai evaluasi. Implementasi proses tersebut yaitu menggunakan Bahasa pemrograman R atau bisa disebut R Studio. Hasil dari penelitian menunjukan bahwa dari akurasi K-means dengan dataset berupa 115 kalimat terdapat 1 kalimat bernilai negatif dan 12 kalimat bernilai positif. Dari hasil evaluasi SSE semakin banyak cluster yang digunakan semakin besar nilai SSE dan nilai akurasi k-means semakin kecil. Dapat disimpulkan SSE terbaik bernilai 75.0% dari 2 cluster dan akan bernilai besar jika ditambahkannya cluster secara terus menerus dengan nilai maksimal sampai 100%.

Kata kunci — Analisis Sentimen, Buletin APTIKOM, K-Means, R, *Sum of Square Error, Term – Inverse Document Frequency* dan *Text Mining.*

I. PENDAHULUAN

Kurangnya pemahaman literasi membuat masyarakat mengalami kesulitan dalam mengeksplorasi pengetahuan. Asosiasi Pendidikan Tinggi Ilmu Komputer (APTIKOM) membuat media informasi secara daring yang dapat menjelaskan secara detail terhadap pemahaman masyarakat yang berbeda - beda. Media tersebut diberi nama Buletin APTIKOM yang terbit setiap bulan sejak tahun 2020. Melalui buletin ini diharapkan dapat memberikan pengetahuan terhadap masyarakat yang awam dalam teknologi terkini. Namun, dari penerbitan buletin pada aplikasi buletin APTIKOM, tidak adanya pengecekan mengenai sentimen penulisan berupa kalimat positif maupun negatif. Maka dari itu, dibuatnya penelitian ini adalah untuk mencari kalimat positif maupun negatif dalam buletin APTIKOM. Sentimen penulisan merupakan tulisan yang menggambarkan perasaan yang berlebih terhadap pendapat atau pandangan orang lain. Analisis sentimen merupakan *clustering* paradoks atau konflik dari teks yang ada dalam kalimat maupun dokumen. Memperoleh tingkat dalam aspek pendapat yang dikemukakan dalam kalimat maupun dokumen atau aspek bersifat positif, negatif atau netral [1].

Pada penelitian [2] adalah untuk memahami penelitian skala besar dan kegunaan *text mining* dan analisis tren untuk mengolah data atau menyelidiki tema penelitian. Sehingga tren mereka dalam kerangka mendapatkan waktu yang efisien. Pada penelitian [3] tentang memproses klasifikasi teks berlabel dan klasifikasi teks tak berlabel. Dalam kasus klasifikasi teks berlabel, kami menggunakan algoritma pembelajaran mesin yang diawasi untuk melatih pengklasifikasi kami. Sementara dalam kasus klasifikasi teks tanpa label, kami mengolah data menggunakan algoritma pembelajaran mesin tanpa pengawasan untuk melatih pengklasifikasi kami. Sehingga dikategorikan ke dalam berbagai kelas untuk klasifikasi teks dan berbagai algoritma bahasa mesin digunakan di kelas untuk menemukan hasil. Pada penelitian [4] tentang menyajikan gambaran umum tentang konsep. Dari mengolah data tujuan analisis sentimen multimodal, meninjau keadaan seni, dan membahas tantangan dan perspektif yang terkait dengan lapangan. Sehingga secara otomatis mengungkap sikap mendasar yang kita pegang terhadap suatu entitas. Pada penelitian [5] tentang memberikan jawaban atas ketiga pertanyaan penelitian tersebut, dan mengolah data untuk itu telah dilakukan analisis literatur kualitatif. Dengan menggunakan kajian pustaka sistematis, sitasi, dan investigasi kutipan. Penelitian ini, *text mining* dapat membantu dalam analisis pelanggan, peluang pemasaran, pencegahan penipuan, peningkatan kegiatan operasional, dan pengembangan model bisnis baru. Sehingga hasil penelitian menunjukkan bahwa perbankan kredit menjadi trend utama dengan topik risiko, deteksi penipuan, persetujuan kredit, dan kebangkrutan. Pada Penelitian [6] tentang mempresentasikan studi empiris dengan memanfaatkan pemodelan LDA untuk menemukan topik penelitian utama UC. Penelitian ini mengolah data dan menganalisis dinamika dan struktur intelektual topik. Penelitian ini yang berkembang tentang pengobatan tetapi tidak menentukan stadium kanker. Sehingga studi ini memberikan pemahaman yang lebih baik tentang tren penelitian UC dan arah penelitian potensial di masa depan.

Dari masalah di atas dapat dibuktikan bahwa *text mining* ini cocok untuk masalah buletin dalam pengembangan buletin yang dibuat untuk mengetahui sentimen pada kalimat positif maupun negatif. Sehingga penggunaan *text mining* untuk penelitian ini yaitu agar buletin yang dilakukan mampu diperiksa dan memenuhi persyaratan dengan hasil yang terbaik untuk diunggah ke aplikasi buletin.

II. METODE PENELITIAN

A. Bahan dan Peralatan Penelitian

Bahan penelitian ini diambil dari Buletin APTIKOM berupa buletin sesuai tema yang telah diunggah ke aplikasi Buletin APTIKOM yang terbit setiap satu bulan sekali. Peralatan Penelitian ini menggunakan *Hardware* dan *Software. Berikut merupakan* Hardware dan Software yang digunakan untuk penelitian :

1) Hardware

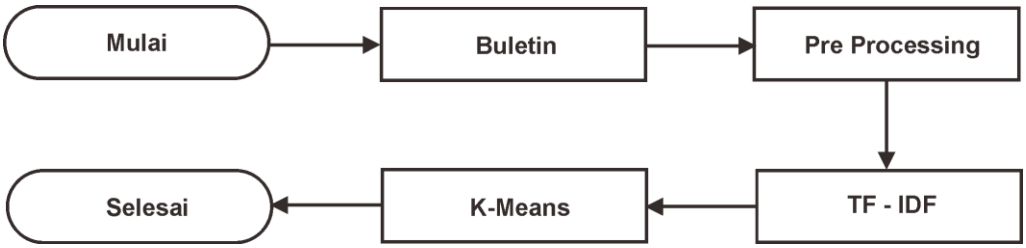
- Laptop ASUS ROG , Processor Intel(R) Core(TM) i5-9300H CPU @ 2.40 GHz, RAM 8,00, dengan sistem operasi Windows 10

2) Software

- Microsoft Word 2016.
- R Studio

B. Prosedur Penelitian

Prosedur penelitian mempunyai tahapan proses seperti pada Gambar 1.



Gambar 1 Prosedur Penelitian

Buletin, buletin merupakan data yang diambil dari buletin aptikom untuk diproses menggunakan preprocessing. Pre processing merupakan proses menghilangkan permasalahan permasalahan yang dapat mengganggu hasil dari pada proses data, pre processing mempunyai tahapan yang terdiri dari to lower, remove numbers. Remove punctuation, strip white space, remove word atau stopwords dan stem document. Term Frequency – Inverse Document Frequency (TF-IDF) yang merupakan langkah menghitung nilai atau bobot sebuah kata terhadap suatu dokumen dan menentukan sentiment berupa kalimat positif maupun negatif pada dokumen. kmeans tahapan yang digunakan untuk mengetahui nilai cluster pada kalimat positif maupun negatif dan proses evaluasi dengan menggunakan metode Sums Square Error yang merupakan cara dalam melakukan validasi cluster melalui jumlah kuadrat setiap anggota cluster menuju pusatnya.

III. HASIL DAN PEMBAHASAN

A. *Pre Processing*

Nampak pada Tabel 1 terdapat perubahan dari kolom sebelum diproses dengan kolom sesudah diproses. To lower merupakan perubahan semua karakter huruf menjadi huruf kecil, Remove numbers merupakan penghilangan karakter angka yang ada pada kata, Remove punctuation merupakan penghilangan pembatas seperti tanda simbol, Strip white space merupakan penghilangan spasi yang berlebihan pada dokumen, Remove words merupakan penghilangan kata-kata pada dokumen. *Stop words* merupakan kosakata yang bukan merupakan ciri (kata unik) pada dokumen dan *Stem document* merupakan proses pemetaan dan penguraian berbagai bentuk dari suatu kata menjadi bentuk kata dasar atau bisa disebut stem.

Tabel 1 Hasil Pre Processing

No	Proses	Sebelum	Sesudah
1	*To lower*	Di era saat ini Natural Language Processing sangat dibutuhkan terutama di sektor perusahaan karena dalam perusahaan didalamnya akan ada informasi data	di era saat ini natural language processing sangat dibutuhkan terutama di sektor perusahaan karena dalam perusahaan didalamnya akan ada informasi data
2	*Remove Numbers*	Sekarang ini Ñ 2016Ð2020 Ñ beliau menjadi Kaprodi S3 Teknik Elektro dan Informatika ITB.	sekarang ini ñ ð ñ beliau menjadi kaprodi s teknik elektro dan informatika itb.

3	*Remove Punctuation*	Biasanya data yang tersedia sebagian besar berbentuk teks, itulah mengapa Natural Language Processing sangat penting dan berpeluang besar diberbagai sektor lainnya.	biasanya data yang tersedia sebagian besar berbentuk teks itulah mengapa natural language processing sangat penting dan berpeluang besar diberbagai sektor lainnya
4	*Strip White Space*	Di era saat ini Natural Language Processing sangat dibutuhkan terutama di sektor perusahaan karena dalam perusahaan didalamnya akan ada informasi data	di era saat ini natural language processing sangat dibutuhkan terutama di sektor perusahaan karena dalam perusahaan didalamnya akan ada informasi data
5	*Remove Words dan Stop Words*	Di era saat ini Natural Language Processing sangat dibutuhkan terutama di sektor perusahaan karena dalam perusahaan didalamnya akan ada informasi data	di era natural language processing butuh utama sektor usaha usaha informasi data
6	*Stem Document*	Di era saat ini Natural Language Processing sangat dibutuhkan terutama di sektor perusahaan karena dalam perusahaan didalamnya akan ada informasi data	di era saat ini natural language processing sangat butuh utama di sektor usaha karena dalam usaha dalam akan ada informasi data

B. *Term Frequency – Inverse Document Frequency (TF-IDF)*

TF-IDF merupakan suatu cara untuk memberikan bobot hubungan suatu kata (*term*) terhadap dokumen [7]. TF – IDF ini adalah langkah sesudah *pre processing* yang sudah diproses menjadi data bersih dari sebuah dokumen. Cara kerja dari TF – IDF ini merupakan teknik untuk menghitung bobot atau nilai sebuah kata terhadap suatu dokumen. Dalam TF – IDF pun termasuk proses menemukannya sentimen analisis atau bisa disebut kalimat positif maupun negatif terhadap suatu dokumen. Berikut merupakan hasil proses TF – IDF dan analisis sentimen seperti pada gambar 2 dan 3. Berdasarkan hasil proses TF – IDF yang terdapat pada dokumen dengan data yang sudah diproses *pre processing* terlihat bahwa sebuah kata berubah menjadi bobot nilai yang terlihat pada gambar 2. Berdasarkan hasil proses analisis sentimen terlihat bahwa terdapat pada nomor baris kalimat terdapat positif dan negatif yang terlihat pada gambar 3 :

word	n
bahasa	35
nlp	27
komputer	24
data	21
manusia	20
teks	17
sepert	16
aptikom	14
ilmu	13
informasi	11

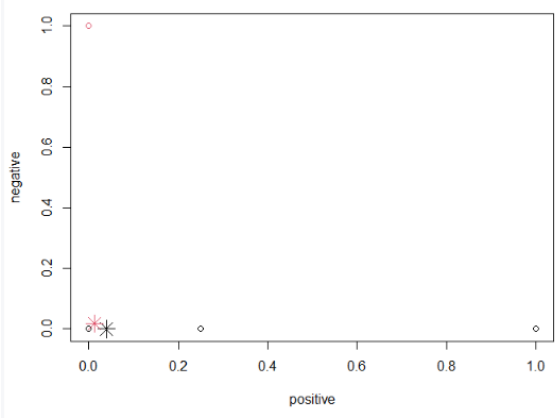
Gambar 2 Hasil TF-IDF

nomor	negative	positive	sentiment
100	0	1	1
14	0	1	1
30	0	1	1
36	1	0	-1
4	0	1	1
83	0	1	1
93	0	1	1
94	0	4	4
96	0	1	1
97	0	1	1

Gambar 3 Hasil Analisis Sentimen

C. K-Means

Algoritma K-means digunakan pada penelitian ini adalah untuk mengetahui nilai *cluster* atau kelompok dalam kalimat positif maupun negatif. Berikut merupakan hasil nilai k-means seperti pada gambar 4. Berdasarkan hasil nilai k-means terdapat dilihat bahwa banyaknya kalimat yang bernilai netral karena memiliki *centroid* pada nilai 0 dan sedikit yang bernilai positif negatif pada nilai 1 dan 0,25 yang terlihat pada gambar 4 dan Sums Square Error (SSE) merupakan pengujian pada penelitian ini digunakan untuk mengevaluasi performa algoritma K-Means. SSE adalah jumlah dari selisih kuadrat antara setiap observasi dan rata-rata kelompoknya. Cara untuk mendapatkan hasil dari SSE adalah dengan menghitung dari masing-masing nilai *cluster*. Berikut merupakan hasil nilai SSE seperti pada gambar 5. Berdasarkan hasil nilai SSE terdapat dilihat bahwa nilai SSE bernilai 75.0% yang terlihat pada gambar 5.



Gambar 4 Hasil Algoritma K-Means

```
Within cluster sum of squares by cluster:
[1] 15448.80 16259.59
 (between_SS / total_SS =  75.0 %)
```

Gambar 5 Hasil Sums Square Error

IV. KESIMPULAN DAN SARAN

A. Kesimpulan

- Hasil penelitian ini merupakan pengimplementasian *text mining* terhadap dokumen buletin APTIKOM untuk menentukan sentimen analisis. Proses pertama menggunakan proses *text processing* yang menghasilkan perubahan dalam kalimat buletin menjadi kata dasar. Proses kedua TF – IDF yang menghasilkan berapa banyak kata yang sering muncul dan menentukan sentimen analisis berupa kalimat positif dan negatif yang terlihat pada gambar 3. Proses terakhir dengan algoritma k-means yang menghasilkan *centroid* pada banyaknya kalimat netral dan sedikitnya kalimat positif dan negatif yang terlihat pada gambar 4 dan SSE yang bernilai 75.0%.
- Hasil penelitian ini terhadap sentimen penulisan buletin APTIKOM yang bertema NLP terdapat 1 kalimat bernilai negatif dan 9 kalimat bernilai positif terlihat pada gambar 3 dari 115 kalimat. Terlihat baik karena banyaknya kalimat yang menghasilkan kalimat netral berupa 105 kalimat yang diproses menggunakan *text processing*, TF – IDF dan algoritma k-means.

B. Saran

Dalam penelitian ini dataset yang digunakan adalah menggunakan dataset buletin APTIKOM. Agar di ketahui lebih lanjut kinerja algoritma K-Means untuk *text mining* maka untuk penelitian selanjutnya akan menambah jumlah dataset yang digunakan. Selain itu karena dataset yang digunakan didalam penelitian ini berbahasa indonesia, maka untuk penelitian selanjutnya dataset yang digunakan adalah buletin berbahasa Inggris.

PENGAKUAN

Naskah ilmiah ini adalah sebagian dari penelitian Tugas Akhir milik Anton Romadoni Junior dengan judul Implementasi Algoritma K-Means untuk Analisis Sentimen pada Buletin APTIKOM, yang dibimbing oleh Hanny Hikmayanti Handayani dan Anis Fitri Nur Masruriyah.

DAFTAR PUSTAKA

[1] D. S. Pamungkas, N. A. Setiyanto, and E. Dolphina, "Analisis Sentiment Pada Sosial Media Twitter Menggunakan Naive Bayes Classifier Terhadap Kata Kunci 'Kurikulum 2013'','" vol. 14, no. 4, pp. 299–314, 2015.

[2] A. Karami, M. Lundy, F. Webb, and Y. K. Dwivedi, "Twitter and Research: A Systematic Literature Review through Text Mining," *IEEE Access*, vol. 8, pp. 67698–67717, 2020, doi: 10.1109/ACCESS.2020.2983656.

[3] S. Sharma and S. Kr., "Review on Text Mining Algorithms," *Int. J. Comput. Appl.*, vol. 134, no. 8, pp. 39–43, 2016, doi: 10.5120/ijca2016907972.

[4] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S. F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, pp. 3–14, 2017, doi: 10.1016/j.imavis.2017.08.003.

[5] M. P. Bach, Ž. Krstič, S. Seljan, and L. Turulja, "Text mining for big data analysis in financial sector: A literature review," *Sustain.*, vol. 11, no. 5, 2019, doi: 10.3390/su11051277.

[6] H. J. Lin, P. C. Y. Sheu, J. J. P. Tsai, C. C. N. Wang, and C. Y. Chou, "Text mining in a literature review of urothelial cancer using topic model," *BMC Cancer*, vol. 20, no. 1, pp. 1–7, 2020, doi: 10.1186/s12885-020-06931-0.

[7] M. Nurjannah and I. Fitri Astuti, "PENERAPAN ALGORITMA TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) UNTUK TEXT MINING Mahasiswa S1 Program Studi Ilmu Komputer FMIPA Universitas Mulawarman Dosen Program Studi Ilmu Komputer FMIPA Universitas Mulawarman," *J. Inform. Mulawarman*, vol. 8, no. 3, pp. 110–113, 2013.