Journal of Information Technology and Computer Science (INTECOMS)
Volume 3 Nomor 2, Desember 2020
e-ISSN : 2614-1574
p-ISSN : 2621-3249

TEXT MINING MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING UNTUK MEMPREDIKSI KEINGINAN PASAR TERKAIT PERJALANAN WISATA

TEXT MINING USES K-MEANS CLUSTERING ALGORITHM TO PREDICT MARKET DESIRES FOR TOURISM TRAVEL

Eka Sabna[1], Budi Mustika[2], Hendry Fonda[3], Dedy Irfan[4], Ambiyar[5]
[123]STMIK Hang Tuah Pekanbaru, [45]Universitas Negri Padang
es3jelita@yahoo.com

ABSTRACT

This research will discuss about how to group (clustering) data in the form of text on social media twitter using the K-means algorithm. This method is used to obtain a description of a data set by revealing the tendency of each individual data group to group with other data individuals. Data obtained from social media which is the object of research is Twitter. The research sample was at most 250 (two hundred and fifty) travel-related tweets. Tools or applications used in the implementation of text mining in this study are Rapidminer Studio v.9.7. This research was conducted at PT. Hika Raya Berkah is a company that operates in the business of travel and travel agency services, both domestic and foreign. Based on the data analysis that has been carried out on tweets with the search word "beach tourism" and continued with clustering using Rapid Miner Studio, the results are the words "beach" and "kidul" occupy the highest position. Meanwhile, when viewed based on the name of the beach, the word "lombok" is obtained. From these results, the company can conclude that the current popular beach attractions are Gunung Kidul in Yogyakarta and Lombok Beach in West Nusa Tenggara..

***Keywords**: Text Mining, K-Means, Tour, Twitter.*

ABSTRAK

Penelitian ini akan dibahas mengenai bagaimana cara pengelompokan (*clustering*) data berupa teks di media sosial *twitter* dengan menggunakan algoritma *K-means*. Metode ini digunakan untuk mendapatkan deskripsi dari sekumpulan data dengan cara mengungkapkan kecenderungan setiap individu data untuk berkelompok dengan individu-individu data lainnya. Data diperoleh dari media sosial yang dijadikan objek penelitian adalah *Twitter*. Sampel penelitian paling banyak 250 (dua ratus limapuluh) buah *tweet tweet* terkait wisata. *Tools* atau aplikasi yang digunakan dalam implementasi *text mining* pada penelitian ini adalah *Rapidminer Studio v.9.7*. Penelitian ini dilaksanakan di PT. Hika Raya Berkah sebuah perusahaan yang bergerak dalam usaha jasa biro perjalanan wisata, baik dalam negeri maupun mancanegara. Berdasarkan analisis data yang telah dilakukan terhadap *tweets* dengan kata pencarian "wisata pantai" dan dilanjutkan dengan clustering menggunakan *RapidMiner Studio*, menghasilkan kata "pantai" dan "kidul" menempati posisi paling tinggi. Sedangkan jika dilihat berdasarkan nama pantai, diperoleh hasil kata "lombok". Dari hasil inilah perusahaan dapat menyimpulkan bahwa objek wisata pantai yang populer pada saat ini adalah gunung kidul di Yogyakarta dan pantai lombok di Nusa Tenggara Barat.

Kata Kunci: Text Mining, K-Means, Wisata, Twitter.

PENDAHULUAN

Pada era *modern* seperti saat ini, sosialisasi antar individu dapat dilakukan dengan komunikasi tidak langsung yaitu melalui media sosial(Cahyono, 2016). Media sosial atau sering disebut situs jejaring sosial (*social network sites*) adalah suatu alat (situs media *online*) yang dapat digunakan untuk melakukan komunikasi tanpa adanya interaksi langsung antar individu(Sosiawan, 2020).

Knowledge adalah *power* dalam dunia bisnis saat ini, dan *knowledge* diturunkan dari data dan informasi, organisasi bisnis yang bisa secara efektif dan efisien masuk ke beragam sumber data teks mereka akan memiliki

knowledge yang diperlukan untuk membuat keputusan yang lebih baik, yang membawa ke keuntungan kompetitif atas berbagai bisnis yang sedang ketinggalan di belakang. Inilah yang mengakibatkan kebutuhan terhadap *text mining* cocok dengan gambaran besar bisnis hari ini(Öztürk, N., & Ayvaz, S., 2018).

Penelitian ini akan dibahas mengenai bagaimana cara pengelompokan (*clustering*) data berupa teks di media sosial *twitter* dengan menggunakan algoritma *K-means*(Salloum, dkk., 2017). Metode ini merupakan metode yang populer digunakan untuk mendapatkan deskripsi dari sekumpulan data dengan cara mengungkapkan kecenderungan setiap individu data untuk berkelompok dengan individu-individu data lainnya(Vishwakarma, dkk., 2017).

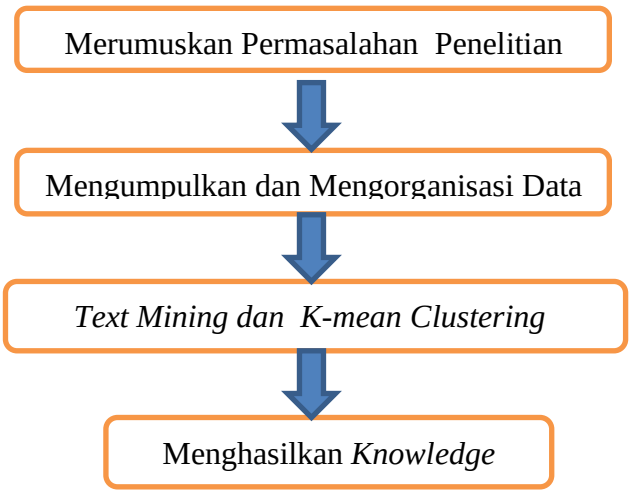
PT. Hika Raya Berkah adalah perusahaan yang bergerak dalam bidang jasa biro perjalanan wisata yang melayani perjalanan wisata dengan tujuan domestik maupun internasional. Paket perjalanan yang ditawarkan pun beragam, mulai dari keberangkatan dengan kelompok (bersama-sama) maupun secara pribadi maupun keluarga (*private*). Dilihat dari banyaknya calon konsumen yang datang dan konsultasi tentang rencana perjalanan wisatanya, maka perlu untuk memprediksi penjualan paket perjalanan wisata yang sesuai dengan selera konsumen.

Adapun tujuan dari penelitian ini antara lain agar perusahaan dapat memprediksi dan merencanakan penjualan produk paket perjalanan wisata, dapat mengetahui destinasi wisata apa yang tengah populer dan sedang diminati oleh masyarakat secara umum. Dapat memberikan informasi tambahan bagi perusahaan dalam mempelajari keinginan pasar dan dapat menjadi dasar bagi Perusahaan dalam pengambilan keputusan.

Penelitian ini diharapkan dapat mendatangkan manfaat buat perusahaan yakni PT. Hika Raya Berkah sebagai Biro Perjalanan Wisata untuk memprediksi serta menentukan arah kebijakan maupun pengambilan keputusan mengenai pemasaran produk-produk perjalanan wisatanya..

METODE

Secara garis besar, langkah-langkah yang perlu dilakukan pada pendekatan deskriptif analitis dapat digambarkan sebagai berikut :



Gambar 1. Langkah-langkah Pendekatan Deskriptif Analitis

Langkah-langkah penelitian deskriptif analitis sebagaimana sebagaimana gambar 1 diatas dapat diuraikan sebagai berikut :

1. Merumuskan permasalahan penelitian.
   Dilakukan langkah-langkah sebagai berikut :
   a. Proses identifikasi dan merumuskan masalah yang hendak diteliti dan dapat dilaksanakan dengan sumber yang ada.
   b. Menentukan tujuan dari penelitian yang akan dikerjakan. Tujuan dari

penelitian harus konsisten dengan rumusan dan definisi dari masalah.

c. Menentukan batas masalah, sejauh mana penelitian deskriptif analitis akan dilaksanakan.

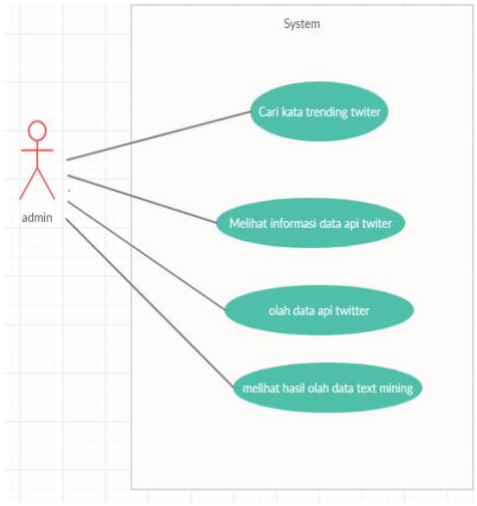
d. Menulusuri sumber-sumber kepustakaan maupun jurnal yang ada hubungannya dengan masalah yang ingin dipecahkan.

e. Merumuskan hipotesis.

2. Mengumpulkan dan mengorganisasi data.
   Pada tahap ini diawali dengan autentikasi pada *twitter* (*API*), dilanjutkan dengan pengumpulan data berupa teks dengan cara diunduh dari *twitter* terkait dengan Destinasi Wisata paling banyak 250 (dua ratus lima puluh) *tweet*(Nababan, dkk., 2020). Setelah teks diunduh, dilanjutkan dengan tahap *Preprocessing* data teks (*Case folding*, *Tokenizing*, *Filtering* dan *Stemming*)(Rosid, dkk., 2020).
3. *Text Mining dan K-means Clustering* menggunakan *RapidMiner Studio*.
   Pada tahap ini, data yang telah diunduh dan diolah tersebut, akan dilanjutkan dengan proses *clustering* menggunakan aplikasi *RapidMiner Studio*(Mustika, dkk., 2020).
4. Menghasilkan *Knowledge*
   Setelah semua tahap diatas selesai dan telah diperoleh hasil, maka dilanjutkan dengan manarik kesimpulan dan membuat laporan akhir(Sabna, 2020).

HASIL DAN PEMBAHASAN

Dalam mengembangkan sistem yang sedang dirancang data yang akan dijadikan masukan adalah berupa kata kunci yang menjadi dasar oleh sistem dalam melakukan pencarian yaitumenggunakan kata kunci "wisata pantai", maka sistem akan melakukan pencarian di *twitter* terkait kata kunci tersebut yang sedang ramai dibicarakan. Hasilnya bisa beragram mulai dari tujuan yang paling banyak dibicarakan hingga perkiraan biaya yang dibutuhkan dalam sebuah paket perjalanan wisata. Perancangan dimulai dengan mengidentifikasi *actor* dan *use case. Use case* digunakan untuk memberikan gambaran secara umum dari sistem yang akan dirancang. Berapa *use case* yang teridentifikasi seperti dijelaskan dibawah ini.



Gambar 2. *Usecase Diagram*

Perancangan dimulai dengan mengidentifikasi *actor* dan *use case. Use case* digunakan untuk memberikan gambaran secara umum dari sistem yang akan dirancang(Melyanti & Iqbal, 2020). Berapa *use case* yang teridentifikasi seperti dijelaskan dibawah ini.

1. Melakukan pemprosesan data masukan yang didapat dari sumber *twitter* untuk pencarian wisata yang lagi popular.
2. Mendata dari data masukan, mendata bertujuan untuk memberi informasi mengenai karakteristik dari data yang

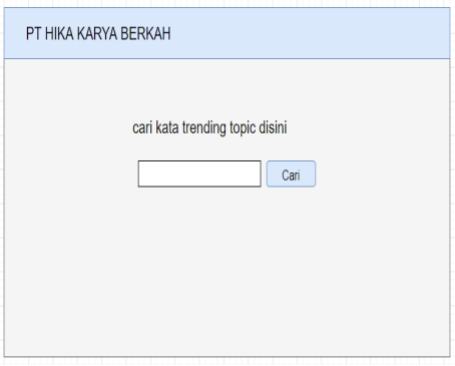
diperoleh yang ditampilkan berupa nama atribut nilai, beserta jumlah dari nilai atribut, jumlah data, tipe data.

3. Melakukan proses *Case Folding* yang bertujuan untuk melakukan penyaringan kata pada data yang tidak konsisten dalam penggunaan huruf kapital. *Case Folding* berperan dalam mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (biasanya huruf kecil atau *lowercase*).
4. Melakukan proses *Tokenizing* yang merupakan tahap pemotongan *string input* berdasarkan tiap kata atau pengelompokan. *Tokenizing* secara garis besar memecah sekumpulan karakter dalam suatu teks ke dalam satuan kata, bagaimana membedakan karakter-karakter tertentu yang dapat diperlakukan sebagai pemisah kata atau bukan. Sebagai contoh karakter *whitespace*, seperti *enter*, tabulasi, spasi dianggap sebagai pemisah kata. Namun untuk karakter petik tunggal **(‘)**, titik **(.)**, semikolon **(;)**, titik dua **(:)** atau lainnya, dapat memiliki peran yang cukup banyak sebagai pemisah kata.
5. Melakukan proses *Filtering* adalah tahap mengambil kata-kata penting dari hasil *token*. Bisa menggunakan **algoritma *stoplist* (membuang kata kurang penting)** atau ***wordlist* (menyimpan kata penting)**. *Stoplist*/ *stopword* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words. Contoh stopwords adalah *“yang”*, *“dan”*, *“di”*, *“dari”* dan lain sebagainya.
6. Melakukan proses *Stemming*. Teknik ***Stemming*** diperlukan selain untuk memperkecil jumlah *indeks* yang berbeda dari suatu dokumen, juga untuk melakukan pengelompokan kata-kata lain yang memiliki kata dasar dan arti yang serupa namun memiliki bentuk atau *form* yang berbeda karena mendapatkan imbuhan yang berbeda, Sebagai contoh kata bersama, kebersamaan, menyamai, akan distem ke *root word*-nya yaitu *“sama”*. Namun, seperti halnya *stopping*, kinerja *stemming* juga bervariasi dan sering tergantung pada *domain* bahasa yang digunakan.

Aktor yang terdapat pada sistem ini hanya satu yaitu pengguna dari aplikasi. Gambaran keseluruhan interaksi antara *actor* dengan *use case* terdapat pada Gambar 2 diatas.

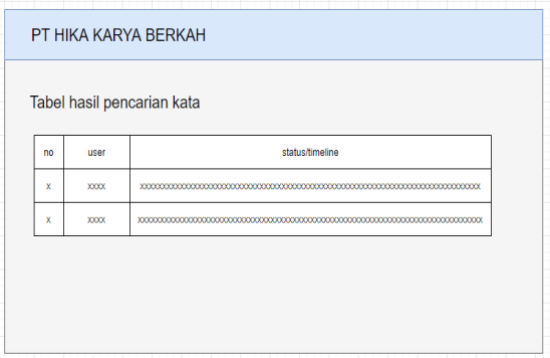
Adapun beberapa rancangan *user interface* implementasi *text mining* pada *twitter* dengan algoritma *K-means* yang tertera sebagai berikut :

1. *User Interface* Halaman Pencarian Kata



Gambar 3. Halaman *Web*

Setelah *admin* melakukan pencarian, maka sistem akan menampilkan halaman hasil pencarian dalam bentuk *table* seperti pada gambar 4. dibawah ini.

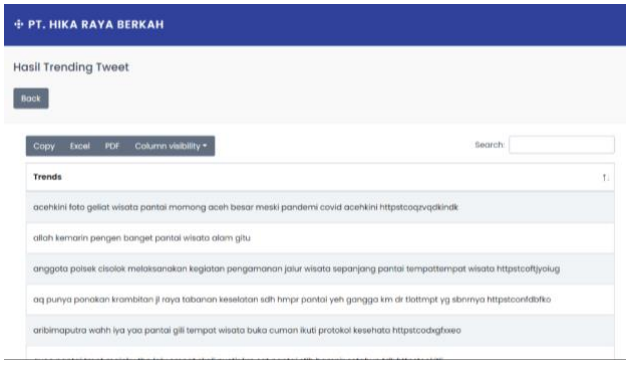


Gambar 4. Halaman Utama

Prototipe pada gambar 4. merupakan rancangan untuk halaman hasil pencarian data, data yang dihasilkan didapat dari data *API twitter* seperti nama pengguna dan status atau *timeline* pengguna.

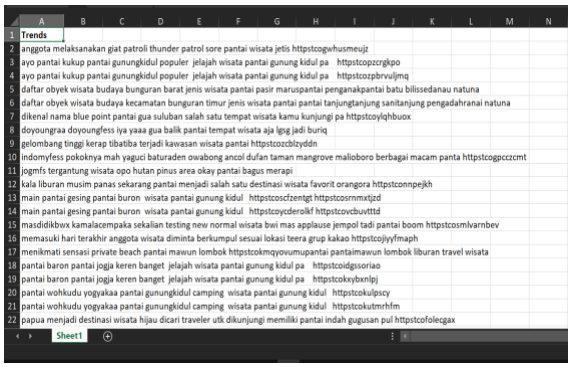
2. Halaman hasil pencarian

Pada halaman ini akan ditampilkan data *tweet* dari pengguna *twitter* yang berkaitan dengan kata kunci "wisata pantai" tersebut. data ini dapat di unduh dengan klik tombol "*excel*" yang terdapat di bagian atas sebelah kiri halaman Halaman ini dapat dilihat pada Gambar 5 dibawah ini.



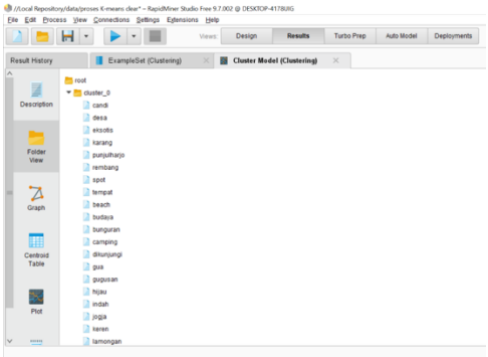
Gambar 5. Halaman awal hasil pencarian

Hasil unduhan berupa data dengan format "*xlsx*" inilah yang dijadikan data awal hasil pencarian terpopuler pada *twitter*. Halaman *excel* dapat dilihat pada Gambar 6 dibawah ini.



Gambar 6. Halaman awal *Microsoft Excel*

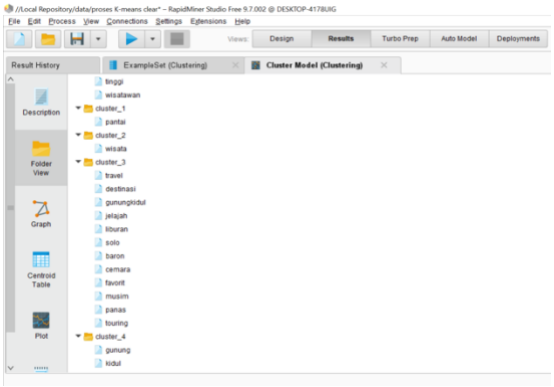
Pada halaman ini menampilkan hasil *clustering K-means* dalam bentuk sebuah tabel dimana didalamnya berisi data yang sudah di *cluster*. Dapat dilihat pada gambar 7. di bawah ini,



Gambar 7. Halaman Hasil *K-means Example Set*

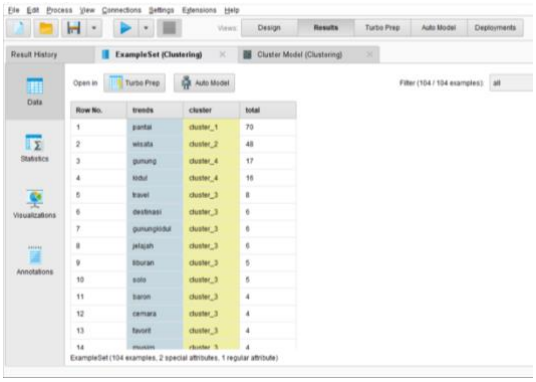
- Halaman hasil dari *K-means Cluster Model*

  Halaman ini berisikan informasi tetang jumlah *cluster* yang ada beserta dengan nilai perkelompok *cluster*. Dapat dilihat pada gambar 8 dibawah ini.

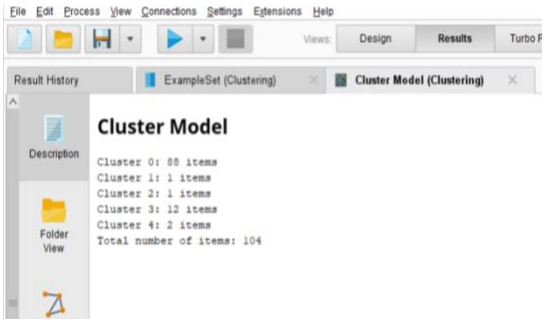


Gambar 8. Halaman Hasil *K-means Cluster Model*

- Halaman isi dari setiap *cluster*
  Contoh pada tahapan ini dapat dilihat pada gambar 9 dan 10 dibawah ini.

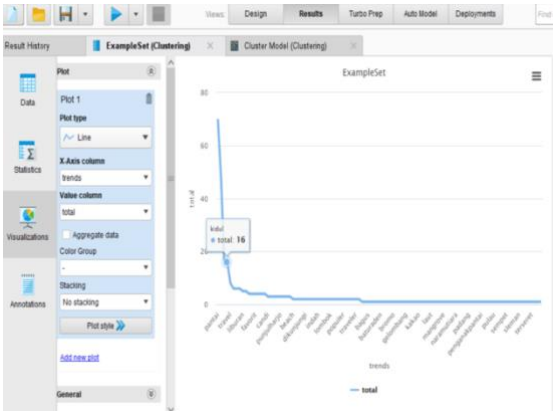


Gambar 9. Halaman isi dari setiap *Cluster*



Gambar 10. Halaman isi dari setiap *Cluster*

Pada halaman ini kita bisa mengetahui topik apa yang paling sering dibicarakan di *twitter* pada *line* yang paling tinggi, pada proses ini orang di *twitter* banyak membicarakan tentang pantai, dapat dilihat pada gambar 11 dibawah ini.



Gambar 11. Halaman Hasil *K-means* dalam bentuk *line*

SIMPULAN

Setelah melakukan analisis dan pengujian, dapat disimpulkan sebagai berikut :

1. *Text Mining* dengan algoritma *K-means Clustering* yang diimplementasikan pada media sosial *twitter*, dapat dijadikan salah satu cara oleh perusahaan dalam menentukan arah kebijakan pemasarannya. Hasil yang diperoleh dari sistem ini menunjukkan informasi wisata yang populer dan sedang ramai dibicarakan ditengah masyarakat.
2. Hasil analisis data yang dilakukan terhadap *tweet* dengan kata pencarian "wisata pantai" menghasilkan kata "gunung kidul" dan "lombok" sebagai tren.
3. Dengan sistem *text mining* yang memanfaatkan pencarian terpopuler pada *twitter* dapat dijadikan pertimbangan oleh perusahaan dalam menentukan arah kebijakan pemasaran produk wisatanya.

DAFTAR PUSTAKA

Cahyono, A. S. (2016). Pengaruh media sosial terhadap perubahan sosial masyarakat di Indonesia. *Jurnal Publiciana*, *9*(1), 140-157.

Öztürk, N., & Ayvaz, S. (2018). Sentiment analysis on Twitter: A

text mining approach to the Syrian refugee crisis. *Telematics and Informatics*, *35*(1), 136-147.

Melyanti, R., & Iqbal, M. (2020). Sistem Informasi Manajemen Penelitian Dan Pengabdian Masyarakat Di Bagian P3m (Studi Kasus: Stmik Hang Tuah Pekanbaru). *Jurnal Ilmu Komputer*, *9*(2), 165-176.

Mustika, B., Sabna, E., & Irawan, Y. (2020). Implementasi Text Mining Pada Twitter Dengan Algoritma K-Means Clustering Sebagai Dasar Kebijakan Marketing Biro Perjalanan Wisata. *Jurnal Ilmu Komputer*, *9*(2), 134-147.

Nababan, A. P. R., Lumenta, A. S. M., Rindengan, Y. D., Pontoh, F. J., & Akay, Y. V. (2020). Analisis Sentimen Twitter Pasca Pengumuman Hasil Pilpres 2019 Menggunakan Metode Lexicon Analysis. *Jurnal Teknik Informatika*, *15*(1), 33-44.

Rosid, M. A., Fitrani, A. S., Astutik, I. R. I., Mulloh, N. I., & Gozali, H. A. (2020, June). Improving text preprocessing for student complaint document classification using sastrawi. In *IOP Conference Series: Materials Science and Engineering* (Vol. 874, No. 1, p. 012017). IOP Publishing.

Sabna, E. (2020). Analisis Text Mining Dari Hasil Wawancara. *Jurnal Ilmu Komputer*, *9*(1), 46-48.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J*, *2*(1), 127-133.

Sosiawan, E. A. (2020). Penggunaan situs jejaring sosial sebagai media interaksi dan komunikasi di kalangan mahasiswa. *Jurnal Ilmu Komunikasi*, *9*(1), 60-75.

Vishwakarma, S., Nair, P. S., & Rao, D. S. (2017). A Comparative Study of K-means and K-medoid Clustering for Social Media Text Mining. *INTERNATIONAL JOURNAL*, *2*(11).