JURNAL KHATULISTIWA INFORMATIKA VOL. 11 NO. 2 Desember 2023
p-ISSN: 2339-1928 e-ISSN: 2579-633X Halaman : 128 - 134

ANALISIS TEXT CLUSTERING PADA DATA MINING MENGGUNAKAN METODE K-NEAREST NEIGHBOR (KNN) DAN DECISION TREE

Owen Baihaqie Islamuddin [1]; Imam Yuadi [2]

Master of Human Resource Development [1]; Department of Information and Library Science [2]
Airlangga University Surabaya
owen.baihaqie.islamuddin-2022@pasca.unair.ac.id

INFO ARTIKEL

Diajukan :
14 September 2023

Diterima :
12 November 2023

Diterbitkan:
31 Desember 2023

Kata Kunci :
Orange Data Mining, Sentimen Analisis, Sepak Bola, Twitter, Visualisasi Data

INTISARI

Twitter saat ini menjadi sebuah platform media sosial yang sangat menarik untuk diamati. Topik yang sedang trend di Twitter seringkali memunculkan komentar dan opini dari masyarakat Indonesia. Salah satu topik yang menarik untuk dianalisis adalah sepak bola Indonesia yang saat ini banyak menyita perhatian sepak bola dunia. Penelitian ini bertujuan untuk menganalisis tanggapan masyarakat terhadap topik tersebut melalui komentar di Twitter, dengan menggunakan metode analisis vader, tweet profiler, dan visualisasi distribution. Proses analisis dilakukan dengan menggunakan aplikasi Orange Data Mining, yang melibatkan tahapan preprocess text, seperti transformation, filtering, normalization, dan tokenization untuk memastikan teks bisa dianalisis. Melalui analisis sentimen menggunakan text clustering di Twitter kita dapat memahami bagaimana masyarakan mengungkapkan dan menggambarkan suatu persepsi terhadap kasus yang sedang dibicarakan baik secara positif, negatif maupun netral. Hasil analisis ini bahwa respon dari masyarakat terhadap isu yang sedang hangat dibicarakan dengan respon masyarakat adalah rasa sukacita, terkejut dan takut.

I. PENDAHULUAN

Jejaring sosial kini menjadi salah satu kebutuhan utama yang tidak bisa dipisahkan dari kehidupan sehari-hari. Saat ini banyak sekali media sosial yang diminati masyarakat, salah satunya adalah Twitter. Twitter atau yang sekarang menjadi aplikasi X adalah salah satu jenis media sosial mikroblog yang memungkinkan penggunanya menulis dan mempublikasikan aktivitas dan/atau opininya. (Husnusyifa, 2019). Melalui Twitter, pengguna dapat berbagi aktivitas sehari-hari seperti memposting foto atau mengutarakan pendapat tentang suatu hal. Trending topik di Twitter selalu menjadi topik perbincangan hangat di masyarakat saat ini.

Berekspresi di Twitter dapat menjadi salah satu media yang dapat dijadikan sebagai suatu objek penelitian. Platform Twitter saat ini sedang ramaikan dengan adanya fenomena kerusuhan dan keresahan di kalangan pecinta sepak bola tanah air sehingga menarik perhatian dunia. Persatuan Sepak Bola Indonesia (PSSI) terancam mendapatkan sanksi dari Federasi Sepak Bola Dunia (FIFA) atas terjadinya tragedi yang memilukan yang menelan banyak korban (Alfandis, 2022). Terkini Timas Indoneisa juga sedang berjuang di Kualifikasi Piala Asia U-17 dan berlaga tanpa adanya dukungan supporter yang datang ke stadion.

Pada intinya, penelitian ini mencoba menganalisis komentar dengan menggunakan metode Vader, khususnya metode analisis lexicon-based berbasis rule–based sentiment analysis. Vader akan memproses suatu kalimat atau teks yang menghasilkan class sentiment berupa positif, negatif, netral dan kompleks. Dalam penelitian ini menggunakan preprocess text dalam Orange Data Mining untuk mengidentifikasi tipe konten dari beberapa informasi tweet dengan menerapkan teknik clustering.

II. BAHAN DAN METODE

Metode penelitian ini menggunakan metode eksperimen dengan melakukan observasi terhadap variabel–variabel sebagai objek yang diteliti. Metode eksperimen ini merupakan ilmu yang mempelajari evaluasi terhadap suatu kondisi tertentu yang perlu dikendalikan agar satu atau beberapa variabel yang dapat dikendalikan.

1. Metode Analisis Data

   a. KNN (K-Nearest Neighbor)

      KNN (K-Nearest Neighbor) untuk mengklasifikasikan objek yang jaraknya sangat dekat dan memerlukan informasi training yang disebut dengan prosedur supervised. Prinsip kerja KNN adalah mengevaluasi informasi dengan neighbor data latih dengan mencari jarak terdekat. Algoritma KNN untuk menghitung jarak yaitu

$$di = \sqrt{\sum_{i=1}^{n} (x_{ij} - p_j)^2}$$

      Keterangan :

      di = Jarak sampel

      x_ij= Data sampel pengetahuan

      p_j = Data input var ke-j

      n = Jumlah sampel

   b. Decision Tree

      Decision Tree atau algoritma C4.5 adalah salah satu teknik klasifikasi yang digunakan untuk mengekstrak hubungan yang relevan dalam data. Algoritma C4.5 adalah jenis klasifikasi yang dapat menghasilkan model pohon keputusan yang mudah dipahami program yang membuat pohon keputusan berdasarkan pada set data input berlabel. Pada saat menghitung menghitung information gain dengan rumus:

$$Gain(S, A) = entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} entropy(Si)$$

      Keterangan :

      S: Himpunan kasus

      A: Atribut

      n: Jumlah patisi S

      Pi: Proporsi dari Si terhadap S

      A|Si|: Jumlah kasus pada partisi ke-i

      |S|: Jumlah kasus dalam S

   c. Studi Literatur Penelitian sebelumnya serta studi literatur mengenai definisi dan permasalahan terkait text mining.

   d. Business Understanding menganalisis isu-isu dan realitas yang berkembang di masyarakat saat ini.

   e. Data Understanding tweet yang dikumpulkan (crawling tweet) menggunakan API Twitter pada tanggal 10 Oktober 2022. Penelitian opini di Twitter menggunakan metode pengumpulan tweets secara manual yaitu memasukkan berbagai macam kata kunci terkait topik Sepak Bola Indonesia di kolom pencarian Twitter dengan menggunakan kata pencarian: timnas, timansindonesia dan sepakbola.

   f. Data preprocess text (Transformation, Tokenization, Normalization dan Filtering). Penentuan load dictonary dan class attribute dengan mencocokkan kata dasar dengan kamus kata sentimen untuk menentukan isi kalimat sentimen (positif, netral, negatif). Semua data tweet diberi label berdasarkan kelas, ada 3 kelas yang akan digunakan dalam penelitian ini yaitu kelas positif, kelas negatif dan kelas netral. Proses labelling tweet dilakukan secara manual.

   g. Clustering data text mining dengan aplikasi orange data mining menggunakan visualisasi distribution yang membantu memvisualisasikan data text mining

dengan sentimen pengguna dari tweet yang diproses.

2. Pengolahan Data

a. Web Scrapping

Metode pengumpulan data yang digunakan untuk mengumpulkan data pendukung yang diperlukan untuk pembuatan aplikasi. Dataset yang digunakan diambil dari Twitter API menggunakan cara proses crawling dengan aplikasi orange data mining. Total data yang digunakan pada aplikasi analisis sentimen terhadap kasus Sepak Bola Indonesia di Twitter sebanyak 300 tweet dengan kata kunci timnas, timansindonesia dan sepak bola pada periode pengumpulan 10 Oktober 2022.

b. Orange Data Mining

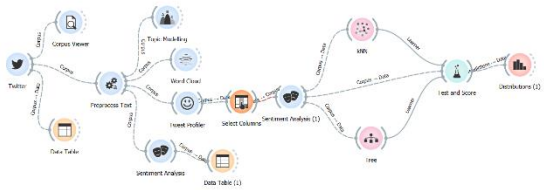
Dalam permasalahan riset ini Orange Data Mining menampilkan sejumlah menu widget untuk mencari data informatif, terutama yang berasal dari konten status dan opini/komentar dari akun Twitter yang akan membuat crowd screen dan Cloud perlahan mengeksploitasi utilitas data Orange. Dengan memanfaatkan teknik data mining, dimungkinkan untuk menemukan informasi tersembunyi dalam kumpulan data dan menggunakannya untuk menganalisis dan memprediksi perilaku di masa depan. Klasifikasi merupakan metode penambangan catatan yang memberikan label kelas pada kumpulan suatu kasus yang tidak diklasifikasikan.

III. HASIL DAN PEMBAHASAN

Aplikasi Orange Data Mining adalah alat data mining yang dapat menghitung secara otomatis sesuai dengan widget yang dipilih.

1. Skenario Penelitian

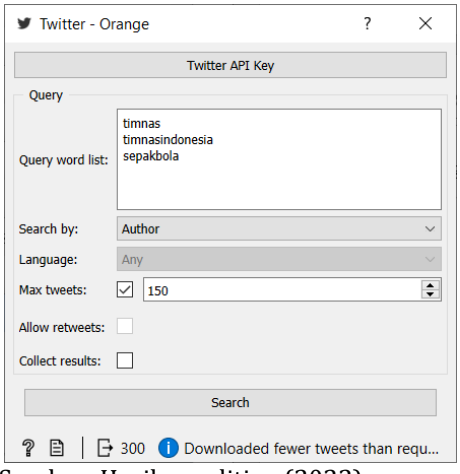
Menampilkan desain widget text clustering dengan menggunakan aplikasi Orange seperti yang ditunjukkan pada gambar di bawah ini.



Sumber: Hasil penelitian (2022)
Gambar1. Design Widget Text Clustering

2. Crawling Data (Twitter)

Data yang dikumpulkan dalam penelitian ini adalah trending topik pada Twitter masyarakat di Indonesia. Jumlah tweet yang diambil dari 300 tweet mengenai Sepak Bola Indonesia dari API dan token Twitter. Dalam Orange Data mining data inputan bisa disebut dengan corpus. Widget corpus yaitu kumpulan dokumen yang bisa menyajikan jumlah baris kalimat, serta memastikan mana fitur yang akan dan tidak akan diinput untuk keperluan analisis.



Sumber: Hasil penelitian (2022)
Gambar2. Data Text Mining (Twitter)

3. Preprocess Text

Dalam hal ini, penerapan text mining untuk pemrosesan teks sebelum analisis teks disebut text preprocessing. Teks

dipecah menjadi unit-unit yang lebih kecil (token), transformation, tokenization, normalization, serta filtering. Langkah-langkah yang diterapkan secara berurutan dalam analisis dapat diaktifkan atau dinonaktifkan di Orange Data Mining di widget Preprocessing Text. Di bawah ini adalah langkah-langkah untuk memproses teks sebelum menganalisisnya dengan Orange Data Mining.

a. Transformation

Langkah pertama dalam preprocessing text adalah transformation, yaitu proses mengubah data input untuk di transformasikan ke huruf kecil secara default.

b. Tokenization

Tokenisasi adalah metode membagi teks menjadi komponen-komponen yang lebih kecil (kata, kalimat, bigram). Termasuk: Word dan Punctuation akan membagi teks berdasarkan kata dan tetap meninggalkan simbol tanda baca (tidak menghapusnya); misalnya: This Sample. (This), (sample), (.). Tweet, akan terbagi teks dengan model Twitter pra-trained, yang berisi hashtag, emoji dan simbol khusus lainnya. Misalnya: This words. :-) #simple → (This), (Words), (.), (:-)), (#simple). Pada dasarnya Word, Punctuation, dan tweet mempunyai karakteristik proses yang sama, namun word dan punctuation merupakan proses yang utama dalam tokenization. Word dan punctuation sering digunakan untuk menganalisis tren.

c. Normalization

Proses selanjutnya adalah normalization yang penerapan kosakata pada sumber dan teks. Memecah teks menjadi kata-kata akan menghasilkan teks yang terpisah dalam sebuah kalimat. Isi konten dan opini seringkali mempunyai posisi yang tidak sempurna (typo). Melalui proses normalisasi ini, makna teks diidentifikasi menggunakan pengubah WordNet, yang menggunakan pengubah WordNet Lemmatizer untuk membuat jaringan persamaan kata (sinonim) kognitif buat token (kata) berdasarkan basis informasi lexicon (kamus) bahasa Indonesia yang besar dari NLTK (Alami Language Toolkit).

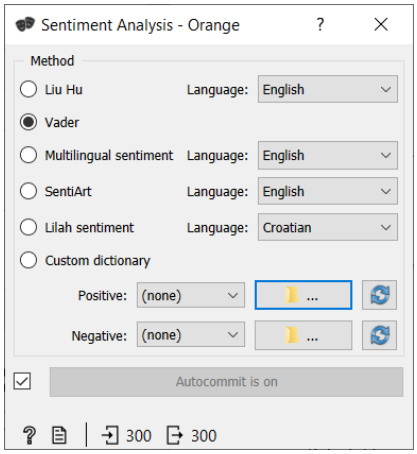
d. Filtering

Pemfilteran menghapus atau mempertahankan pilihan kata. Termasuk: Stopwords menghapus kata-kata penutup dari teks (misalnya menghapus 'and', 'or', 'in', ... dll). Dapat juga berisi daftar kata sandi terpisah yang disediakan dalam format file .txt dengan satu stopword per baris. Dengan opsi filter bahasa dapat digunakan untuk berbagai bahasa dengan bahasa Indonesia ditetapkan sebagai default. Regexp menghilangkan kata-kata yang cocok dengan ekspresi reguler ini \.|,|:|;|!|\?|\(|\)|\||\+|'|"|‘|’|“|”|'|\'|…|\-|–|—\$|&|\|>|<|\/| Dan secara default diatur untuk menghapus tanda baca. Most Frekuensi Token (kata yang sering muncul) dengan wordcloud di Orange Data pada dasarnya untuk melihat frasa mana yang sering muncul di dokumen dan menentukan berapa banyak frasa yang paling sering ditampilkan menggunakan Most Frekuensi Token. Langkah prepocess text telah dilakukan, setelah itu data disajikan sebagai teks terpisah dan ditampilkan dalam bentuk word cloud di Orange Data Mining.

Sumber: Hasil penelitian (2022)
Gambar3. World Cloud Sepak Bola Indonesia

4. Analisis Sentimen

Analisis menggunakan algoritma Vader untuk mengklasifikasikan popularitas atau class sentiment yaitu positif, negatif, dan netral.

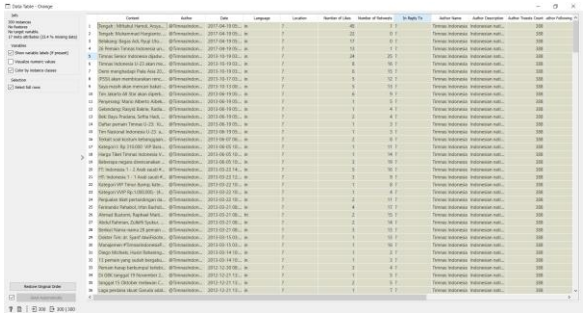


Sumber: Hasil penelitian (2022)
Gambar4. Widgest Sentiment Analysis

Vader akan mengklasifikasikan dan menetapkan poin teks berdasarkan nilai setiap kata yang tercantum dalam lexicon vader. Compound merupakan hasil akhir dari evaluasi skor total. Skor total inilah yang hendak direkaptulasi serta dibanding hasilnya. Pada saat analisis terdapat beberapa atribut variabel yang difokuskan dari fitur yang menjadi atribut yang akan digunakan untuk dianalisis (used features) dari widget corpus untuk teks hasil terjemahan yang terdapat pada setiap informasi, dengan tujuan untuk memperoleh hasil berupa atribut positif negatif, netral, serta skor total (compound).

5. Tabel Data

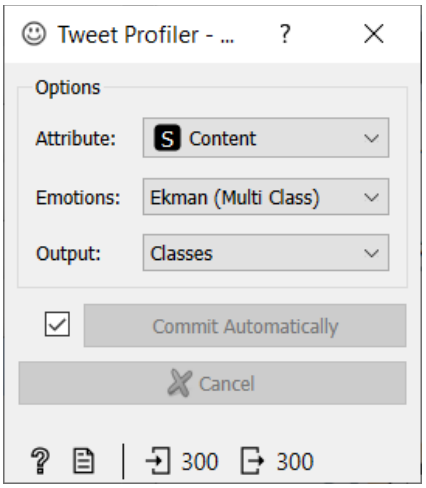
Tabel data juga merupakan repon yang dihasilkan untuk menampilkan semua atribut yang dipilih untuk ditampilkan sebagai output di widget select column pada langkah sebelumnya. Dengan menggunakan data crawling Twitter menggunakan API hasil analisis dalam tabel data hasil analisis sentimen akan menghitung seberapa positif, negatif, dan netral dengan melihat skor total (Compound) dengan formula perhitungan dalam format data numerik.



Sumber: Hasil penelitian (2022)
Gambar5. Tabel Data

6. Tweet Propfiler

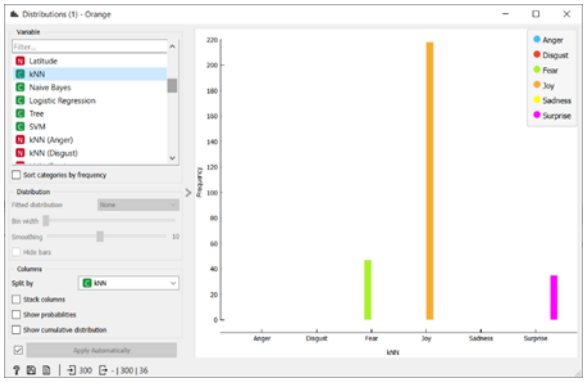
Tweet Profiler mengumpulkan data opini dari setiap tweet atau dokumen yang disajikan server. Widget mengirimkan data ke server tempat model menghitung skor probabilitas dan/atau skor sentimen. Setelah mengirim data ke server, model menghitung skor probabilitas dan/atau sentimen yang sesuai dari widget. Widget ini mendukung 3 klasifikasi emosi, yaitu Ekmans, Plutchiks dan Profile of Mood States (POMS). Klasifikasi emosi terdapat 3 kategori yaitu Ekmans, Plutchiks ataupun Profile of Mood States. Klasifikasi kelas jamak akan menghasilkan satu emosi yang sangat bisa jadi per dokumen, sedangkan banyak label akan menciptakan nilai dalam kolom untuk tiap emosi. Penelitian ini akan menggunakan atribut Konten untuk analisis, klasifikasi emosi Ekman dengan opsi multi-kelas dan memilih untuk mengamati variabel Emosi yang telah dikelompokkan dengan orange data mining. Pada penelitian ini menggunakan data 300 tweets mengenai Sepak Bola Indonesia. Data yang telah di crawling menggunakan widget dari orange data mining dengan Corpus dan dihubungkan ke Tweet Profiler.



Sumber: Hasil penelitian (2022)
Gambar6. Tweet Profilier

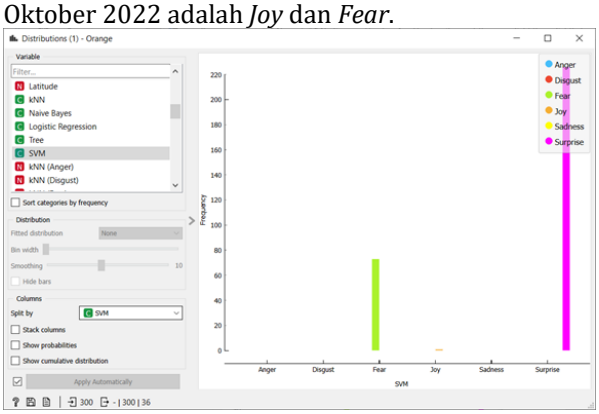
7. Distribution

Widget Distribusi menampilkan distribusi nilai atribut diskrit atau kontinu. Jika data berisi variabel kelas, distribusi dapat dikondisikan di kelas. Setelah melakukan tweet profiler pada widget, langkah selanjutnya menghubungkan corpus pada distribution. Hasil akan terlihat 6 bentuk emosi dari data Twitter yang telah diinput. Pada penelitian ini widget menampilkan emosi para pengguna Twitter dengan kata timnas, timansindonesia dan sepakbola. Dari hasil 6 emosi ini data menunjukan bahwa respon dari Twitter per tanggal 10 Oktober 2022.



Sumber: Hasil penelitian (2022)
Gambar7. Distribution KNN

Pada widget distribution KNN menampilkan emosi para pengguna Twitter dengan hashtag timnas, timansindonesia dan sepakbola. Dari hasil 6 emosi ini data menunjukan bahwa respon dari Twitter per tanggal tanggal 10 Oktober 2022 adalah *Joy* dan *Fear*.



Sumber: Hasil penelitian (2022)
Gambar8. Distribution Tree

Pada widget distribution Tree menampilkan emosi para pengguna Twitter dengan hashtag timnas, timansindonesia dan sepakbola. Dari hasil 6 emosi ini data menunjukan bahwa respon dari Twitter per tanggal tanggal 10 Oktober 2022 adalah *Suprise* dan *Fear*.

IV. KESIMPULAN

Beberapa hasil uraian analisis yang dilakukan sehingga dapat disimpulkan yaitu beberapa trending topik yang menjadi fokus pembicaraan mengenai menjadi isu yang sedang ramai dibicarakan dan untuk melihat respon masyarakat terhadap hal tersebut. Metode analisis menggunakan tweet profiler dapat mengetahui mood atau emosi para pengguna Twitter dengan trending topik yang sedang terjadi di suatu negara khususnya mengenai isu-isu tersebut. Dengan melakukan analisis clustering distribution kita dapat mengetahui klasifikasi emosi para pengguna Twitter menggunakan visualisasi yang telah diinput ke dalam setiap Corpus Orange Data Mining.

V. REFERENSI

Alfandis. 4 Oktober, 2002. Ancaman Sanksi FIFA Imbas Tragedi Kanjuruhan, PSSI Diminta Berbenah detiksulsel.

Husnusyifa, Annisa. 2019. Pengaruh Penggunaan Media Sosial Twitter Terhadap Sikap Fanatisme Penggemar (Studi Pada Media Sosial Twitter @BTOBIndonesia Terhadap Sikap Fanatisme Penggemar). IDEA: Jurnal Humaniora.

Hozair., Anwari., & Alim Syariful. 2021. Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa dengan Model K-Nearest Neighbor, Decision Tree. Jurnal Nero, 6(2).

Irmayanti Windi. 2021. Visualisasi Data Pada Data Mining Menggunakan Metode Klasifikasi Naive Bayes. Jurnal Khatulistiwa Informatika, 9(1), 68-72.

Istiqomah Santi., & Marsaroh Siti. 2022. Sistem Prediksi Penjualan Hijab Menggunakan Algoritma Prediksi Di Aplikasi Orange (Studi Kasus : Kota Tasikmalaya). Jurnal Saintesa, 2(1).

Nur Akbar, M., Annisa Safitri, S., Nasrullah., & Mubarak. 2022. Analisis Sentimen Pengguna Indihome dengan Metode Klasifikasi Suppoet Vektor Machine (SVM). Jurnal Shift, 2(1).

Oktaria Sihombing, L., Hannie., & Arif Dermawan, B. 2021. Sentimen Analisis Customer Review Produk Shopee Indonesia Menggunakan Algoritma Naive Bayes Classifier. Jurnal Pendidikan Informatika, 5(2).

Sentiya, A., & Suroyo, H. (2019). Analisis Text Clustering Akun Fanpage Shopee Indonesia Dengan Komentar Followers Menggunakan Tools Orange Data Mining. Bina Darma Conference, 1055–1067. http://conference.binadarma.ac.id/index.php/BDCCS/article/view/660

Wiguna, R. A. Raffaidy, & Rifai, A. I. 2021. Analisis Text Clustering Masyarakat Di Twitter Mengenai Omnibus Law Menggunakan Orange Data Mining. Journal of Information Systems and Informatics, 3(1), 1–12. https://doi.org/10.33557/journalisi.v3i1.78