KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 2, No 5, April 2022
Hal 178-185
https://djournals.com/klik

Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier

Tengku Ridwansyah

Program Studi Teknik Informatika, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Budi Darma,
Jalan Sisingamanganraja No. 338,Medan, Sumatera Utara, Indonesia
Email: umberella.zone@gmail.com

Abstrak−Kota Medan adalah salah satu kota terbesar di Indonesia yang memiliki pesona akan wisata, kuliner dan juga penunjang sarana dan prasarana yang cukup memadai didalam kota, dan tentunya akan dilirik pariwisata luar negeri atau mancanegara. Sehingga perlu dilakukan penelitian guna mengetahui bagaimana sentimen masyarakat dunia terhadap pariwisata kota Medan tersebut, apakah mayoritas masyarakat dunia menilai positif atau negatif.Tanggapan masyarakat dunia mengenai pariwisata kota Medan didapat dari Application Programming Interface (API) pada Twitter karena media sosial tersebut memiliki pengguna yang sangat banyak di Indonesia bahkan hingga mencapai 19,5 juta pengguna dari total 300 juta pengguna global. Pada penelitian ini, salah satu bagian dari algoritma Text Mining adalah praproses teks, praproses teks yang digunakan adalah case folding, tokenizing, stopwords, dan stemming. Untuk praproses stemming digunakan algoritma snowball stemmer Sedangkan pada analisis klasifikasi data teks tersebut menggunakan algoritma Naïve Bayes Classifier dan untuk mempartisi data akan di gunakan Metode K-Fold Cross Validation.Data yang digunakan dalam penelitian ini adalah kumpulan tweet mengenai pariwisata kota Medan pada tanggal 1 Desember 2019 hingga 8 Desember 2019. Data didapat dari Twitter API (Application Programming Interface) sebanyak 2000 tweet. Penggalian data dari Twitter melalui Twitter API (Application Programming Interface) menggunakan RStudio sebagai console proses penggalian data.

Kata Kunci : Kota Medan; Text Mining; Twitter; Naïve Bayes Classifier; K-Fold Cross Validation.

Abstract−The city of Medan is one of the largest cities in Indonesia which has tourist charm, culinary and also supporting facilities and infrastructure that are quite adequate in the city, and of course will be ogled by foreign or foreign tourism. So it is necessary to do research in order to find out how the public sentiment towards tourism in the city of Medan, whether the majority of the world community considers it positive or negative. The response of the world community regarding Medan tourism is obtained from the Application Programming Interface (API) on Twitter because social media has a lot of users in the world. Indonesia has even reached 19.5 million users out of a total of 300 million users worldwide. In this study, one part of the Text Mining algorithm is text preprocessing, the text preprocessing used is case folding, tokenizing, stopwords, and stemming. For the preprocessing stemming, the snowball stemmer algorithm is used, while the analysis of the text data classification uses the Naïve Bayes Classifier algorithm and to partition the data the K-Fold Cross Validation method will be used. The data used in this study is a collection of tweets about tourism in the city of Medan on December 1, 2019 until December 8, 2019. Data is obtained from the Twitter API (Application Programming Interface) as many as 2000 tweets. Extracting data from Twitter via Twitter API (Application Programming Interface) using RStudio as a console for the data retrieval process.

Keywords : Medan City; Text Mining; Twitter; Naïve Bayes Classifier; K-Fold Cross Validation

1. PENDAHULUAN

Dengan adanya kemajuan ilmu pengetahuan dan Teknologi (IPTEK) yang terus berkembang pesat hingga saat ini, hampir semua aktivitas dan kegiatan manusia sangat bergantung kepada teknologi. Teknologi Informasi sangat memegang peranan penting di berbagai aspek kehidupan kemanusiaan karena dapat menghubungkan dan menyediakan berbagai informasi melalui web. Web atau website adalah halaman informasi sudah disediakan melalui jalur internet sehingga bisa dapat diakses oleh khalayak dunia selama terhubung dengan jaringan internet dan web yang menyajikan dua tipe tekstual yaitu fakta begitupun opini [1]. Fakta adalah suatu pernyataan aktual mengenai entitas dan yang terjadi di dunia, sedangkan opini adalah pernyataan subyektif yang digambarkan sentimen atau persepsi seseorang mengenai entitas atau terjadi di dunia. Web sudah menyajikan berbagai fakta dan opini dari hal apapun melalui blog individu, media sosial, dan micro-blog lainnya, oleh karena itu jika ada seseorang atau organisasi maupun perusahaan/instansi yang dapat memperoleh opini publik tentang suatu produk atau layanan tertentu, maka pengguna data yang didapati dalam web dapat menjadi alternatif yang maksimal daripada menggunakan survei konvensional.

Kota Medan merupakan ibu kota provinsi Sumatera Utara, Indonesia. Kota medan adalah kota terbesar posisi ketiga di Indonesia setelah kota Jakarta dan kota Surabaya, serta kota terbesar di luar pulau Jawa. Kota Medan memiliki jalan masuk dari inti kota menuju ke pelabuhan kapal dan bandar udara yang sudah difasilitasi oleh jalan besar tol dan stasiun kereta api, kota ini juga memiliki beraneka ragam pariwisata begitu juga ragam budaya dan adat sosial nya. Selain salah satu dari kota terbesar di Negara Indonesia, Medan merupakan kota memiliki berbagai entertaiment dan hiburan, begitu juga ragam budaya untuk dijadikan sebagai destinasi dari pariwisata nya dalam kota tersebut. Oleh karena itu maka kota Medan layak dijadikan bahan perbincangan bagi masyarakat Indonesia sendiri maupun Internasional. [2]

Opini masyakarat dunia yang ada di twitter dapat digunakan sebagai objek analisis sentimen untuk dapat menyaksikan pendapat masyarakat dunia mengenai tentang kesan mereka terhadap pariwisata di kota medan apakah positif atau negatif. Data yang di gunakan didapati dari hasil pencarian di twitter terdiri dari 2000 tweet yang akan dipilah menjadi dua bagian yaitu untuk data sentimen negatif dan data sentimen positif. Melalui data tersebut dapat disimpulkan tentang opini masyarakat tentang kota Medan terhadap pariwisatanya.

Sentimen analisis atau opinion mining adalah suatu proses analisis berlandaskan komputasi tentang pendapat, sentimen, maupun perasaan. Sentimen analisis berguna untuk memeriksa kecenderungan beberapa sentimen atau pendapat, apakah sentimen itu cenderung ber-opini positif ataupun negatif. Sebelum melakukan analisis sentimen, diperlukan praproses teks dengan metode text mining untuk mengolah data teks agar siap untuk dianalisis. Praproses teks tersebut meliputi case folding, tokenizing, stopwords, dan stemming. Case folding merupakan pra-proses untuk merubah semua teks menjadi huruf kecil. Tokenizing adalah proses memecah teks yang berasal dari kalimat menjadi kata per kata. Stopwords merupakan kosa kata yang tidak termasuk kata unik atau ciri dari sebuah dokumen sehingga perlu dihilangkan. Stemming adalah proses untuk mendapatkan kata dasar dengan menghilangkan imbuhan pada kata. Pada penelitian ini, proses stemmer menggunakan algoritma Snowball Stemmer. Algoritma tersebut merupakan algoritma yang dikembangkan untuk stemming berdasarkan grammer Bahasa Internasional (English Language) antara lain untuk mendapati kata dasar dengan menghapus imbuhan kata antara lain awalan kata (prefix), sisipan kata (infix), akhiran kata (suffixes), serta kombinasi kata antara awalan kata dan akhiran kata (confixes). Adapun pada klasifikasi sentimennya akan menggunakan algoritma mengklasifikasi teks. Penelitian ini bermaksud agar mengklasifikasi tweet menjadi 2 (dua) sentimen antara lain sentimen positif serta sentimen negatif. Didalam penelitian ini memakai teks bahasa internasional didapati dari twitter berupa tweet. Opini/tanggapan masyakarat dunia didapati dari tweet tersebut bisa dioptimalkan sebagai bahan sentimen analisis supaya pendapatkan tanggapan masyarakat dunia mengenai tentang kesan mereka terhadap pariwisata kota medan apakah positif atau negatif. Data yang di gunakan didapati dari hasil pencarian di twitter terdiri dari 2000 tweet yang akan dibagi menjadi dua yaitu untuk data negatif dan positif, dan untuk metode akurasi ketepatan klasifikasi untuk sentimen analisis nya adalah NBC (Naïve Bayes Classifier) dan K-fold cross validation berfungsi untuk mempartisi data menjadi 2 (dua) data kategori yaitu train data dan test data melalui ratio 90%:10% yang guna nya untuk mengurasi bias yang terjadi dalam pengambilan sampel data dari tweet.

Penelitian terdahulu mengenai sentiment analysis adalah Analisis sentimen twitter menggunakan text mining dengan algoritma Naïve Bayes Classifier, penelitian tersebut membahas hasil analisis sentimen terkait pilkada Jawa Barat yang akan dilangsungkan pada tahun 2018. Akurasi yang didapat dari penelitian tersebut sebesar 84% [3]. Selain itu, penelitian dengan metoda yang sama pernah dilakukan Moh. Yasid mengenai analisis sentimen maskapai Citilink pada twitter dengan menggunakan metode naïve bayes. Penelitian tersebut menyimpulkan bahwa akurasi dengan NBC(Naïve Bayes Classifier) sebesar 0.778 atau 77% [4]. Didapati banyak metoda klasifikasi pada ilmu statistika yang difungsikan untuk sentimen analisis namun metoda yang sering dipergunakan pada klasifikasi teks adalah Algoritma Naïve Bayes Classifier. Metoda NBC (Naïve Bayes Classifier) telah banyak dipergunakan terhadap penelitian tentang Text Mining sebab memiliki kelebihan antara lain metode sederhana namun memiliki akurasi yang tinggi[5]

Melalui penelitian ini maka penulis menyimpulkan untuk dapat menghitung data sentimen positif atau sentimen negatif dalam opini masyarakat dunia tentang pariwisata kota medan, begitu juga akurasi klasifikasi dari hasil sentimen nya menggunakan NBC (Naïve Bayes Classifier) dan juga diharapkan dapat memberikan saran kepada pemerintahan untuk pariwisata kota Medan dalam menyajikan sarana dan prasarana untuk menunjang turis luar negeri terhadap kota Medan sebagai pertimbangan dalam menentukan baik atau buruk nya pariwisata kota Medan yang akan dilihat..

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Didalam penyusunan penelitian ini, penulis menggabungkan data yang di perlukan untuk penelitian ini. Adapun langkah-langkah yang penulis lakukan dalam penyusunan penelitian ini yaitu :

1. Penelitian Pustaka (Library research)
   Dalam melakukan penelitian pustaka, penulis membaca buku dan jurnal yang terkait dengan judul skripsi yang akan dikembangkan.
2. Sumber Data
   Pemakaian data pada penelitian ini adalah kumpulan tweet masyarakat dunia mengenai kota Medan.
3. Analisa
   Upaya mendapatkan dan menata dengan sistematis sumber data berdasarkan hasil pada sumber data yaitu kumpulan tweet masyarakat dunia mengenai kota Medan.
4. Penyusunan Laporan
   Setelah mendapati hasil pada tahapan ini dikerjakan pada sistem yang telah selesai diterapkan dan diuji serta hasil analisa disusun secara lengkap.

2.2 Text Mining

Algoritma Text Mining merupakan algoritma digunakan agar penggalian data supaya dapat melengkapi masalah kebutuhan informasi dengan menerapkan metode machine learning, data mining, natural language processing, manajemen pengetahuan serta pencarian informasi. Metode text mining melibatkan pra-proses dokumen seperti pengkategorian teks, ekstraksi informasi, serta ekstraksi kata. Teknik ini dapat berguna untuk mengekstraksi informasi terhadap sumber data melewati identifikasi serta eksplorasi bentuk yang menarik [6]. Text Mining adalah metode yang digunakan untuk menangani permasalahan klasifikasi, information retrival, information extraction serta clustering [7]. Pada umumnya proses kerja dari Text Mining banyak mengapdopsi dari penelitian data mining namun yang menjadi perbedaan adalah

pola yang digunakan oleh Text Mining diambil dari sekumpulan bahasa alami yang tidak terstruktur sedangkan dalam data mining pola yang diambil dari database yang terstruktur. Tahapan algoritma Text Mining secara umum adalah praproses teks.

2.3 Praproses Teks

Pra-proses teks adalah salah satu bagian pada algoritma Text Mining, proses mengolah teks yang difungsukan untuk pengubahan bentuk dokumen menjadi data yang terstruktur sesuai dengan kepentingannya supaya dapat diolah lebih lanjut pada proses text mining. Langkah pra-proses teks dalam klasifikasi berfungsi untuk memaksimalkan akurasi klasifikasi data. Pra-proses tulisan kalimat maupun pembentukan kata yang berimbuhan. Adapun 4 (empat) aturan pembentukan kata yang berimbuhan (affix) untuk mengubah arti kata dasar yaitu antara lain :

1. Kata awalan (prefix)
   Imbuhan yang dapat ditambahkan pada awal kata dasar. Imbuhan ini terbagi dalam dua jenis yaitu :
   a. Standar, yang mencakup imbuhan “in-”, “on-”, dan “verb –ing”.
   b. Kompleks, yang mencakup imbuhan “verb-1”, “verb-2”, “verb-3”.
2. Akhiran (suffix)
   Imbuhan yang ditambahkan di belakang kata dasar. Suffix yang sering digunakan didalam reguler verbs yaitu “present”, “past”, dan “past participle”. Selain itu, imbuhan kata yang menunjukkan keterangan kepemilikan (‘my’, ‘your’, dan ‘our’, dan their) dan kepemilikan dia/benda (“his”, “her”, dan “its”) juga dapat dikategorikan sebagai suffix.
3. Kata awalan dan kata akhiran (confix)
   Imbuhan ini yang ditambahkan pada depan kata dan belakang kata dasar (prefix serta suffix) dengan bersama-sama.
4. Kata sisipan (inffix)
   imbuhan ini yang disisipkan pada tengah kata dasar. Rule terbentuknya kata pada bahasa Inggris berkaitan dengan pra-proses teks karena hasil akhir pra-proses teks dinantikan mendapati kata dasar yang sesuai pada Grammer English Language. Adapun tahapan pada pra-proses teks antara lain :
   a. Case Folding
      Adalah proses pengubahan seluruh karakter pada teks menjadi huruf kecil serta menghilangkan tanda baca serta angka. Prosedur case folding yaitu memproses huruf alphabet dari “a” sampai dengan “z” saja agar karakter selain huruf tersebut akan dihapus [8].
   b. Tokenizing
      Adalah proses pemecahan pada awalnya kalimat menghasilkan kata-kata atau memutus deretan pada string menghasilkan potongan-potongan, seperti kalimat berdasarkan setiap kata yang menyusunnya.
   c. Stopwords
      Adalah kosa-kata yang tidak termasuk kata unik ataupun ciri khas terhadap suatu dokumen/data atau tidak memiliki pesan apapun secara signifikan pada teks ataupun kalimat [9]. Kosa-kata yang dimaksud yaitu seperti kata penghubung serta kata keterangan yang tidak termasuk kata unik, misalnya “dari”, “akan”, “seorang”, dan lain sebagainya.
   d. Stemming
      Yaitu tahapan untuk mendapati kata dasar dengan proses menghilangkan awalan kata, akhiran kata, sisipan kata, dan confixes (kombinasi dari awalan kata dan akhiran kata). Di penelitian ini metode stemming yang digunakan yaitu algoritma Snowball Stemmer.
   e. Removal URL
      adalah sebuah tahapan menghapus URL atau alamat di website yang ada di tweet [10].

2.4 Algoritma *Naïve Bayes Classifier*

Algoritma *Naïve Bayes Classifier* adalah metode yang menerapkan konsep probabilitas bersyarat [18]. Pada dasar nya Algoritma NBC dapat dinotasikan pada persamaan sebagai berikut.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (1)$$

Metode NBC adalah algoritma yang berfungsi supaya mencari nilai untuk probabilitas tertinggi dan mengklasifikasikan data uji terhadap kategori yang paling tepat. Algoritma *Naive Bayes Classifier* juga adalah salah satu metoda yang digunakan untuk mengklasifikasi teks. Kelebihan NBC merupakan algoritma-nya sederhana akan tetapi mempunyai akurasi yang tinggi. Terdapat 2 (dua) tahapan dalam mengklasifikasikan *tweet*. Cara pertama yaitu melakukan pelatihan kepada *tweet* yang sudah dikenal kategorinya. Selain itu cara kedua merupakan tahapan klasifikasi *tweet* yang belum dikenal kategorinya [11]. Didalam metode NBC semua dokumen akan direpresentasikan terhadap pasangan atribut “a_1, a_3,..., a_n” dimana a_1 yaitu kata pertama, sedangkan a_2 yaitu kata kedua dan seterusnya. Kemudian V merupakan himpunan pengelompokan *tweet*. disaat tahap klasifikasi algoritma akan mencari probabilitas tertinggi dari semua pengelompakan dokumen yang diujikan (VMAP). Adapun persamaan VMAP adalah pada persamaan (2.2) sebagai berikut :

$$V_{MAP} = \arg\max P\left(V_j\right) \prod_i P\ (a_i|Vj) \quad (2)$$

Nilai $P(Vj)$ dihitung pada proses *train*, didapat dengan rumus yaitu :

$$P(V_j) = \frac{|\ doc\ j\ |}{|\ train\ |} \quad (3)$$

Dimana |*doc j*| adalah jumlah *tweet* yang memiliki pengelompokan *j* didalam *train*. Kemudian |*train*| adalah jumlah *tweet* didalam sampel yang digunakan untuk *train*. Untuk setiap probabilitas kata ai pada setiap kategori $P(a_i|V_j)$, dihitung pada saat *train*.

$$P(a_i|V_j) = \frac{n_i+1}{|\ n+kosakata\ |} \quad (4)$$

Dimana n_i merupakan jumlah kemunculan kata a_i dalam tweet yang berkategori V_j, sedangkan n merupakan banyaknya seluruh kata dalam *tweet* dengan kategori V_j dan |kosakata| merupakan banyaknya kata dalam data *train*.

2.5 Metode K-Folds Cross Validation

K-fold cross validation merupakan salah satu dari teknik yang difungsikan untuk memilah data menjadi *train* data serta *test* data. Teknik ini banyak diterapkan peneliti karena didapati mengurangi bias yang didapatkan didalam pengambilan sebuah sampel. *K-fold cross validation* berlaku continyu membagi data-data menjadi *train* data dan *test* data, sehingga setiap data akan mendapat kesempatan menjadi *test* data [12]. *K* yaitu besar angka pemilahan data yang digunakan untuk pembagian *train* dan *test*. Tabel berikut adalah ilustrasi pembagian data dengan metode *K-Fold cross validation*.

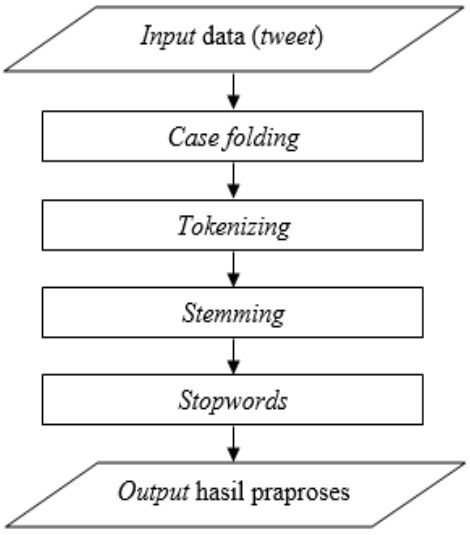
Tabel 1. Ilustrasi Dalam Pembagian Data

Fold 1	*Fold 2*	Fold 3	…	*Fold K*
Test	Train	Train	…	Train
Train	Test	Train	…	Train
Train	Train	Test	…	Train
…	…	…		…
Train	Train	Train	…	Test

3. HASIL DAN PEMBAHASAN

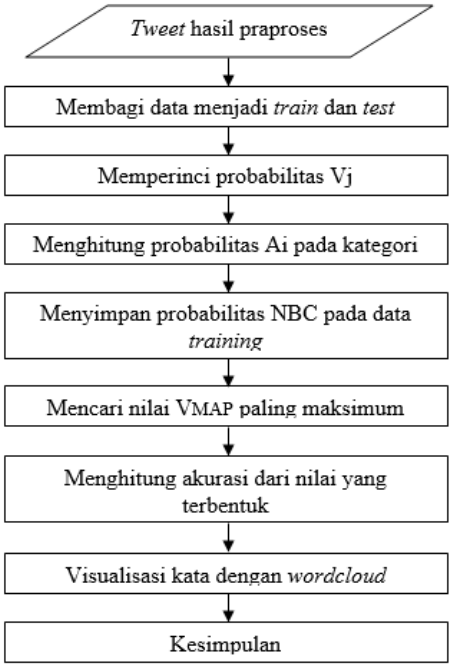
3.1 Analisa Data

Analisa adalah proses yang sangat utama agar mengetahui tahapan yang berlaku pada permasalahan yang nantinya dikaji. Didalam analisa terdapat beberapa bagian yaitu sumber data, struktur data dan langkah analisis dari data. Untuk mengetahui lebih lanjut dalam analisa maka terlebih dahulu mengetahui diagram alir algoritma text mining dan untuk mengetahui lebih lanjut tentang penjabaran dari diagram alir ini akan di bahas di sub bab langkah analisis data, diagram alir algoritma text mining disajikan pada Gambar 1 sebagai berikut :



Gambar 1. Diagram Alir Praproses Teks

Dan setelah melalukan praproses teks didalam algoritma text mining dan mendapati hasil dari algoritma tersebut, maka akan dilakukan klasifikasi menggunakan algoritma NBC untuk mengetahui sentimen. berikut adalah Diagram Alir Klasifikasi menggunakan algoritma Naïve Bayes Classifier (NBC) terdapat pada gambar 2 :

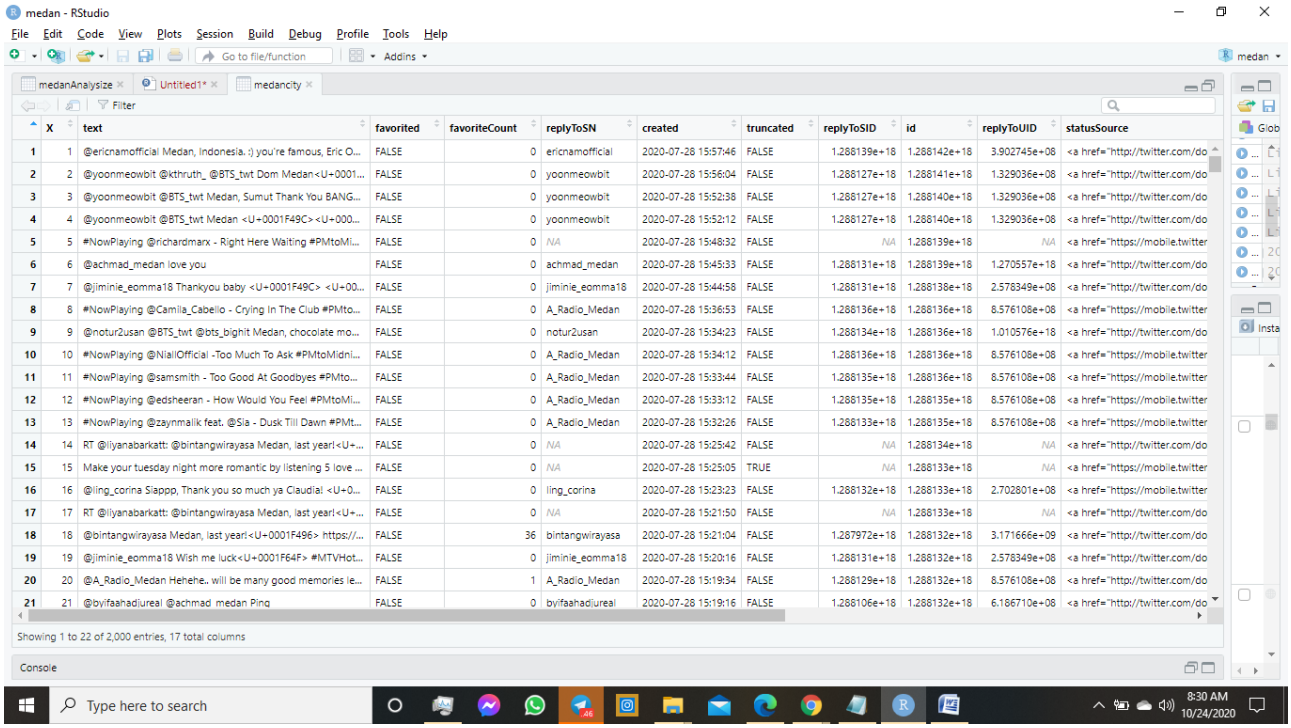


Gambar 2. Diagram Alir Klasifikasi NBC

Data yang diperoleh pada penelitian ini yaitu kumpulan tweet mengenai pariwisata kota Medan pada tanggal 1 Desember 2019 hingga 8 Desember 2019. Data tersebut didapati dari Twitter API (Application Programming Interface) sebanyak 2000 tweet. Penggalian data dari Twitter melalui Twitter API (Application Programming Interface) menggunakan RStudio sebagai console proses penggalian data.

Data yang total 2000 terhadap pariwisata kota Medan dipilah menjadi train data dan test data dengan pembagian 90%:10% menggunakan K-fold cross validation. K-Fold Cross Validation berfungsi untuk menentukan data twitter yang layak dijadikan data testing dari setiap data training, dan data testing tersebut dijadikan patokan untuk diklasifikasikan didalam algoritma Naïve Bayes Classifier (NBC), dan bertujuan untuk data dari twitter tersebut dapat mengurangi bias yang terjadi dalam pengambilan sampel. Struktur data yang digunakan dalam penelitian ini setelah dilakukan pra-proses pada data teks tweet terdiri dari variabel prediktor yaitu kata dasar setiap tweet dan variabel respon yaitu klasifikasi sentimen tweet (sentimen positif maupun sentimen negatif).

Adapun contoh struktur pada data penelitian sebelum pra-proses yaitu:



	X	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID	statusSource
1	1	@ericnamofficial Medan, Indonesia. :) you're famous, Eric O...	FALSE	0	ericnamofficial	2020-07-28 15:57:46	FALSE	1.288139e+18	1.288142e+18	3.902745e+08	<a href="http://twitter.com/do
2	2	@yoonmeowbit @kthruth_ @BTS_twt Dom Medan<U+0001...	FALSE	0	yoonmeowbit	2020-07-28 15:56:04	FALSE	1.288127e+18	1.288141e+18	1.329036e+08	<a href="http://twitter.com/do
3	3	@yoonmeowbit @BTS_twt Medan, Sumut Thank You BANG...	FALSE	0	yoonmeowbit	2020-07-28 15:52:38	FALSE	1.288127e+18	1.288140e+18	1.329036e+08	<a href="http://twitter.com/do
4	4	@yoonmeowbit @BTS_twt Medan <U+0001F49C><U+000...	FALSE	0	yoonmeowbit	2020-07-28 15:52:12	FALSE	1.288127e+18	1.288140e+18	1.329036e+08	<a href="http://twitter.com/do
5	5	#NowPlaying @richardmarx - Right Here Waiting #PMtoMi...	FALSE	0	NA	2020-07-28 15:48:32	FALSE	NA	1.288139e+18	NA	<a href="https://mobile.twitter
6	6	@achmad_medan love you	FALSE	0	achmad_medan	2020-07-28 15:45:33	FALSE	1.288131e+18	1.288139e+18	1.270557e+18	<a href="http://twitter.com/do
7	7	@jiminie_eomma18 Thankyou baby <U+0001F49C> <U+00...	FALSE	0	jiminie_eomma18	2020-07-28 15:44:58	FALSE	1.288131e+18	1.288138e+18	2.578349e+08	<a href="http://twitter.com/do
8	8	#NowPlaying @Camila_Cabello - Crying In The Club #PMto...	FALSE	0	A_Radio_Medan	2020-07-28 15:36:53	FALSE	1.288136e+18	1.288136e+18	8.576108e+08	<a href="https://mobile.twitter
9	9	@notur2usan @BTS_twt @bts_bighit Medan, chocolate mo...	FALSE	0	notur2usan	2020-07-28 15:34:23	FALSE	1.288134e+18	1.288136e+18	1.010576e+18	<a href="http://twitter.com/do
10	10	#NowPlaying @NiallOfficial -Too Much To Ask #PMtoMidni...	FALSE	0	A_Radio_Medan	2020-07-28 15:34:12	FALSE	1.288136e+18	1.288136e+18	8.576108e+08	<a href="https://mobile.twitter
11	11	#NowPlaying @samsmith - Too Good At Goodbyes #PMto...	FALSE	0	A_Radio_Medan	2020-07-28 15:33:44	FALSE	1.288135e+18	1.288136e+18	8.576108e+08	<a href="https://mobile.twitter
12	12	#NowPlaying @edsheeran - How Would You Feel #PMtoMi...	FALSE	0	A_Radio_Medan	2020-07-28 15:33:12	FALSE	1.288135e+18	1.288135e+18	8.576108e+08	<a href="https://mobile.twitter
13	13	#NowPlaying @zaynmalik feat. @Sia - Dusk Till Dawn #PMt...	FALSE	0	A_Radio_Medan	2020-07-28 15:32:26	FALSE	1.288133e+18	1.288135e+18	8.576108e+08	<a href="https://mobile.twitter
14	14	RT @iiyanabarkatt: @bintangwirayasa Medan, last year!<U+...	FALSE	0	NA	2020-07-28 15:25:42	FALSE	NA	1.288134e+18	NA	<a href="http://twitter.com/do
15	15	Make your tuesday night more romantic by listening 5 love ...	FALSE	0	NA	2020-07-28 15:25:05	TRUE	NA	1.288133e+18	NA	<a href="https://mobile.twitter
16	16	@ling_corina Siappp, Thank you so much ya Claudia! <U+0...	FALSE	0	ling_corina	2020-07-28 15:23:23	FALSE	1.288132e+18	1.288133e+18	2.702801e+08	<a href="https://mobile.twitter
17	17	RT @iiyanabarkatt: @bintangwirayasa Medan, last year!<U+...	FALSE	0	NA	2020-07-28 15:21:50	FALSE	NA	1.288133e+18	NA	<a href="http://twitter.com/do
18	18	@bintangwirayasa Medan, last year!<U+0001F496> https://...	FALSE	36	bintangwirayasa	2020-07-28 15:21:04	FALSE	1.287972e+18	1.288132e+18	3.171666e+09	<a href="http://twitter.com/do
19	19	@jiminie_eomma18 Wish me luck<U+0001F64F> #MTVHot...	FALSE	0	jiminie_eomma18	2020-07-28 15:20:16	FALSE	1.288131e+18	1.288132e+18	2.578349e+08	<a href="http://twitter.com/do
20	20	@A_Radio_Medan Hehehe.. will be many good memories le...	FALSE	1	A_Radio_Medan	2020-07-28 15:19:34	FALSE	1.288129e+18	1.288132e+18	8.576108e+08	<a href="http://twitter.com/do
21	21	@byifaahadiureal @achmad_medan Ping	FALSE	0	byifaahadiureal	2020-07-28 15:19:16	FALSE	1.288106e+18	1.288132e+18	6.186710e+08	<a href="http://twitter.com/do

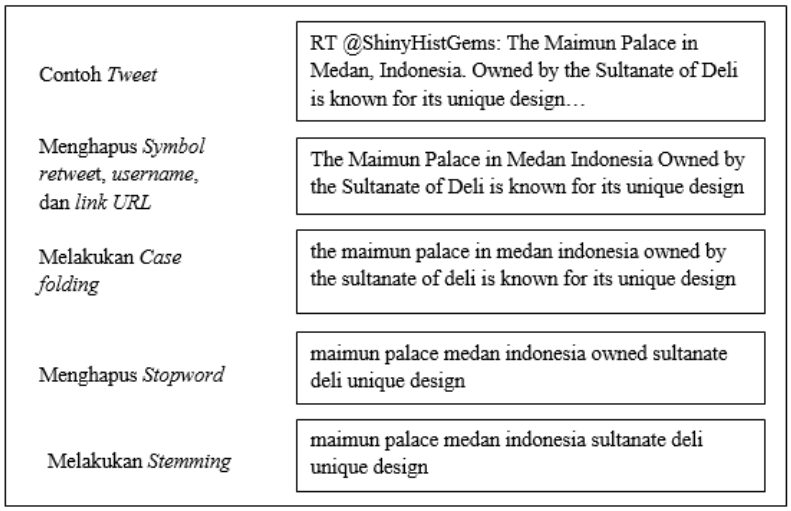
Gambar 3. Contoh Struktur Data Sebelum Pra-proses teks melalui RStudio

3.2 Pembahasan

Didalam pembahasan akan di uraikan tentang bagaimana membuat simulasi tahap proses algoritma *text mining* dan hasil dari algoritma tersebut beserta proses nya. Dan juga akan dibahas mengenai tentang perhitungan klasifikasi yang menggunakan *naïve bayes classifier* (NBC) beserta hasil dari klasifikasi tersebut untuk mengetahui sentimen dari hasil dari algoritma *text mining*.

3.2.1 Simulasi Pra-proses teks

Penjelasan mengenai hasil dari setiap langkah pra-proses teks akan digambarkan di simulasi pra-proses teks terhadap sebuah data *tweet*. *Tweet* yang akan dipakai sebagai contoh yaitu *tweet* "*RT @ShinyHistGems: The Maimun Palace in Medan, Indonesia. Owned by the Sultanate of Deli is known for its unique design…*". Berikut adalah gambar dari Simulasi Praproses Teks pada gambar 4.



Gambar 4. Simulasi Praproses Teks

Maka hasil dari proses algoritma *text mining* untuk *tweet* pada gambar 3.4 dari simulasi tersebut adalah "maimun palace medan indonesia sultanate deli unique design". Untuk *tweet* berikutnya, seperti tweet "*Guess Coming soon at Delipark Medan https://t.co/sdAviqvjxr*" akan diberlakukan pra-proses teks dengan tahapan yang sama sehingga memperoleh hasil pra-proses terakhir sebagai berikut.

guess	delipark	medan

Gambar 5. Contoh Hasil Pra-proses Teks

Didapati dari kedua contoh hasil pra-proses teks pada *tweet* diatas, maka didapati struktur data setelah pra-proses teks sebagai berikut ini :

Tabel 2. Contoh Struktur Data Setelah Pra-proses Teks

Tweet	Variabel Prediktor									
ke -	maimun	palace	medan	indonesia	sultanate	deli	unique	design	guess	delipark
1	1	1	1	1	1	1	1	1		
2			1						1	1

Pembentukan struktur data setelah dilakukan praproses teks seperti pada Tabel 2, yaitu menjadikan setiap kata menjadi variabel prediktor dan meletakkannya pada satu baris. Jika terdapat tambahan kata (variabel prediktor) dari *tweet* baru, maka kata tersebut diletakkan pada baris yang sama dan di kolom berikutnya. Namun jika terdapat kata yang sama atau kata yang telah ada pada struktur data, maka kata tersebut tidak dimasukkan lagi pada struktur data. Sehingga tidak terdapat kata atau variabel prediktor yang sama dalam struktur data. Nilai dari setiap kata tersebut merupakan jumlah kemunculan kata dalam *tweet* ke-i seperti yang terdapat pada Tabel 2 mengenai contoh struktur data setelah praproses teks.

3.2.2 Pembagian Data menggunakan Metode *K-Fold Cross Validation*

Seperti yang di ketahui di landasan teori bahwa *K-Fold Cross Validation* adalah metoda yang mempilah data menjadi *train* data dan *test* data, maka data yang menjadi data *train* merupakan data yang sudah dilakukan pra-proses teks yaitu *twitter* pertama pada gambar 4 adalah "maimun palace medan indonesia sultanate deli unique design" kemudian data yang akan menjadi data *test* yaitu pada gambar 5. yaitu "guess delipark medan" yang sudah dilakukan pra-proses teks,

dan berikutnya pembagian data tersebut dengan menggunakan metode *K-Fold Cross Validation* akan dilakukan perhitungan klasifikasi guna untuk menentukan sentimen dari penggabungan *train* data maupun *test* data tersebut. Metode ini bertujuan untuk mengurangi bias yang terjadi didalam pengambilan sampel.

3.2.3 Perhitungan Klasifikasi

Adapun yang menjadi ilustrasi perhitungan untuk dilakukannya klasifikasi sentimen dengan Algoritma NBC dari contoh *tweet* pada tabel 2 sebagai *train* data serta contoh *tweet* "*Guess Coming soon at Delipark Medan*" sebagai *test* data. Perhitungan tersebut dilakukan supaya dapat dilakukan pemilahan apakah contoh *tweet* tersebut sebagai *test* data memiliki sentimen negatif ataupun positif. Langkah awal yang dilakukan yaitu menghitung probabilitas pada setiap kelas sentimen dengan persamaan (3) yaitu :

$$P(V_1) = \frac{|\,doc1\,|}{|\,training\,|} = \frac{1}{2} = 0{,}5$$
$$P(V_2) = \frac{|\,doc2\,|}{|\,training\,|} = \frac{1}{2} = 0{,}5$$

Dimana $P(V_1)$ merupakan probabilitas bersentimen positif serta $P(V_2)$ merupakan probabilitas bersentimen negatif. Kemudian dilakukan perhitungan probabilitas kemunculan setiap kata pada masing-masing kategori dengan persamaan (4).

P(maimun | v1) = (1+1)/(8/10) = 0,40
P(maimun | v2) = (0+1)/(3/10) = 0,30
P(palace | v1) = (1+1)/(8/10) = 0,40
P(palace | v2) = (0+1)/(3/10) =0,30
P(medan | v1) = (1+1)/(8/10) = 0,40
P(medan | v2) = (1+1)/(3/10) = 0,150
P(indonesia | v1) = (1+1)/(8/10) = 0,40
P(indonesia | v2) = (0+1)/(3/10) = 0,30
P(sultanate | v1) = (1+1)/(8/10) = 0,40
P(sultanate | v2) = (0+1)/(3/10) = 0,30
P(deli | v1) = (1+1)/(8/10) = 0,40
P(deli | v2) = (0+1)/(3/10) = 0,30
P(unique | v1) = (1+1)/(8/10) = 0,40
P(unique | v2) = (0+1)/(3/10) = 0,30
P(design | v1) = (1+1)/(8/10) = 0,40
P(design | v2) = (0+1)/(3/10) = 0,30
P(guess | v1) = (0+1)/(8/10) = 0,80
P(guess | v2) = (1+1)/(3/10) = 0,150
P(delipark| v1) = (0+1)/(8/10) = 0,80
P(delipark | v2) = (1+1)/(3/10) = 0,150

Kemudian yaitu mencari probabilitas tertinggi pada *tweet* yang diujikan. *Tweet test* setelah dibuat pra-proses teks, kemudian kata tersebut terdiri dari kata "unique", "medan", dan "delipark". Agar dicari probabilitas tertinggi dari setiap kata pada *tweet* tersebut menggunakan persamaan (2).

$$P(V_1)\prod_i P(a_i|V_1) = (0{,}5)\big(P(unique|V_1) \times P(medan|V_1) \times P(delipark|V_1)\big) = (0{,}5)(0{,}40 \times 0{,}40 \times 0{,}80)$$
$$= 0{,}064$$
$$P(V_2)\prod_i P(a_i|V_2) = (0{,}5)\big(P(unique|V_2) \times P(medan|V_2) \times P(delipark|V_2)\big) = (0{,}5)(0{,}30 \times 0{,}30 \times 0{,}150)$$
$$= 0{,}00675$$
$$V_{MAP} = \arg\max P(V_j)\prod_i P(a_i|V_j) = V_1$$

Nilai pada probabilitas teks setiap *tweet test* yang terbesar yaitu probabilitas setiap teks terhadap sentimen positif maka *tweet test* tersebut dapat diklasifikasikan menjadi *tweet* " sentimen positif ".

4. KESIMPULAN

Adapun hal yang didasari dari hasil implementasi didapati kesimpulan antara lain Total kata pada data 2000 tweet pariwisata kota medan adalah 18.965 kata Hasil count kata merupakan hasil dari analisa sentimen menggunakan software RStudio pada pariwisata kota Medan adalah 1803 tweet positif atau 80.2% sedangkan tweet sentimen negatif yaitu 137 atau sebesar 19.8%, Secara umum penggunaan software RStudio untuk penetapan hasil analisa sentimen menunjukkan bahwa software tersebut sangat cocok dalam mengimplementasikan text mining, analysis sentiment, machine learning, natural language program lainnya, Kosa kata sering muncul di pariwisata kota Medan kelompok keseluruhan adalah kebanyakan menyangkut isu politik. Sedangkan pada kelompok kata yang sering muncul lainnya adalah kata "medan",

"family", dan "running" dan Pada penelitian berikutnya, penelitian yang sama dapat dikembangkan menggunakan API Stream dan dapat dibuat program untuk otomatisasi klasifikasi. Sehingga hasil analisis sentimen dapat diakses secara realtime. Selain itu, daftar kata pada stopwords dapat dilengkapi dengan daftar kata singkatan dan daftar kata dalam berbagai bahasa Negara lain nya..

REFERENCES

[1] G. A. Adji, T. B. Buntoro, and M. M. Agustin, "Identification of Ambiguous Sentence Pattern in Indonesian Using Shift-Reduce Parsing," Proceeding 1st Int. Con ference Comput. Sci. Eng., pp. 61–63, 2014.
[2] Eri Zuliarso. Adhi Viky Sudiantoro "Artificial Intelligence konsep dan penerapannya," in Artificial Intelligence konsep dan penerapannya, Seno, Ed. Yogyakarta: Andi, 2014, pp. 2–3.
[3] B. Widodo and S. Derwin, "Artificial Intelligence konsep dan penerapannya," in Artificial Intelligence konsep dan penerapannya, Seno, Ed. Yogyakarta: Andi, 2014, p. 7.
[4] Van Meter and Van Horn, "The Policy Implementation Process," in Artificial Intelligence konsep dan penerapannya, Seno, Ed. Yogyakarta: Andi, 2014, p. 209.
[5] C. J. S. Lelywiay, S. Widowati, and K. M. L, "Deteksi Pola Ambiguitas Struktural pada Spesifikasi Kebutuhan Perangkat Lunak menggunakan Pemrosesan Bahasa Alami," vol. 4, no. 3, pp. 51–64, 2019, doi: 10.21108/indojc.2019.4.3.355.
[6] R Feldman and J Sanger, The Text Mining Handbook:." [Online]. Available: https://github.com/famrashel/idn-tagged-corpus/blob/master/Indonesian_Manually_Tagged_Corpus_ID.tsv.
[7] Awalludin M.pd, Pengantar Bahasa Indonesia untuk perguruan tinggi. Yogyakarta: Deepublish, 2017.
[8] O. Anton, W. Prihartono, and S. Sos, "Surat Kabar & Konvergensi Media (Studi Deskriptif Kualitatif Model Konvergensi Media Pada Solopos)," vol. 4, no. 1, pp. 105–116, 2016, doi: 10.12928/channel.v4i1.4210.
[9] D. RAMAIAH K, "GATE AND PGECET FOR COMPUTER SCIENCE AND INFORMATION TECHNOLOGY," in GATE AND PGECET FOR COMPUTER SCIENCE AND INFORMATION TECHNOLOGY, 2nd ed., Delhi: asoke k gosh, 2019, pp. 2–5.
[10] Suendri, "Implementasi Diagram UML (Unified Modelling Language) Pada Perancangan Sistem Informasi Remunerasi Dosen Dengan Database Oracle (Studi Kasus: UIN Sumatera Utara Medan)," J. Ilmu Komput. dan Inform., vol. 3, no. 1, pp. 1–9, 2018.
[11] R. AS and M. Shalahudin, Rekayasa Perangkat Lunak : Terstruktur dan berorientasi objek. Bandung:.