IMPLEMENTASI ALGORITMA NAIVE BAYES PADA KLASIFIKASI TWEET UNTUK MENGETAHUI TINGKAT KEMALASAN SISWA

Dena Nurani Katresna[1)], Faisal Muhammad Dzikry[2)]
Program Studi Informatika Fakultas Teknik Universitas Siliwangi
e-mail: 167006009@student.unsil.ac.id1, 167006058@student.unsil.ac.id2

Abstrak

Twitter merupakan salah satu media sosial yang banyak digunakan oleh masyarakat saat ini. Salah satu pengguna twitter yang aktif adalah masyarakat usia sekolah. Twitter menjadi tempat siswa untuk meluapkan isi hatinya. Kondisi malas dalam belajar yang dialami siswa juga turut diungkapkan melalui tweets. Tujuan penelitian ini untuk menerapkan algoritma naïve bayes dalam klasifikasi tweet terkait kemalasan siswa. Masing-masing faktor yang mempengaruhi kemalasan tersebut dihitung. Data tweet yang digunakan dalam perhitungan sebanyak 773.225. Parameter yang digunakan dalam perhitungan: ulangan, guru, kurikulum, tugas, fullday. Hasil penelitian diketahui bahwa "full day" merupakan parameter dengan nilai tertinggi yang berhubungan dengan kemalasan.

Kata Kunci : malas, Twitter, Naive Bayes

Abstract

Twitter is one of the social media that is widely used by people today. One of the active Twitter users is the school age community. Twitter is a place for students to vent their hearts out. The lazy conditions experienced by students in learning were also expressed through tweets. The purpose of this study is to apply the naïve Bayes algorithm in the classification of tweets related to student laziness. Each of the factors affecting laziness is counted. The tweet data used in the calculation were 773,225. Parameters used in the calculation: tests, teacher, curriculum, assignments, full day. The results showed that "full day" is the parameter with the highest value associated with laziness.

Keywords*: Laziness, Twitter, Naive Bayes*

I. PENDAHULUAN

Perkembangan akses teknologi informasi yang semakin pesat pada era globalisasi saat ini dapat memberikan kemudahan untuk berkomunikasi lebih efektif dan efisien. Pengguna internet di Indonesia saat ini mencapai 63 juta orang dan 95 persennya menggunakan internet untuk mengakses media sosial. Salah satu media sosial yang diakses masyarakat adalah Twitter.

Media sosial twitter merupakan salah satu situs *micro blogging* yang memungkinkan penggunanya untuk menulis berbagai topik dan membahas isu-isu yang sedang terjadi. Twitter memberikan layanan kepada penggunanya untuk mengirim atau membaca *tweets* yang telah dibagikan dengan membatasi karakter maksimal 140. Adanya layanan tersebut menyebabkan masyarakat lebih memilih menuangkan opininya melalui media sosial daripada menyampaikannya secara langsung. Opini yang tertuang dalam bentuk tweets tersebut tersimpan dalam bentuk data yang dapat dimanfaatkan untuk mencari sebuah informasi. Namun, dalam pemanfaatannya membutuhkan analisis yang tepat sehingga informasi yang dihasilkan dapat membantu banyak pihak untuk mendukung suatu keputusan atau pilihan.

Text mining merupakan sebuah proses pengambilan informasi dari data tekstual yang memiliki kualitas tinggi serta dapat mengetahui permasalahan dalam teks dari sebuah topik tertentu. Seperti pada penelitian [5] melakukan analisa sentimen terhadap feedback yang diberikan oleh orang tua murid dengan membagi sentimen menjadi 3 kategori yaitu sentimen positif, negatif dan netral. Salah satu teknik pembelajaran dari text mining adalah Naïve Bayes Classifier. Metode Naïve Bayes Classifier dianggap sebagai metode yang berpotensi baik untuk melakukan klasifikasi data daripada metode klasifikasi lainnya dalam hal akurasi dan komputasi [1], [6]. Algoritma Naïve Bayes Classifier dapat digunakan untuk memprediksi suatu nilai dari variabel dalam data testing [4]. Oleh karena itu, penelitian ini mencoba melakukan analisa menggunakan Naïve Bayes untuk melihat tingkat

kemalasan siswa dilihat dari beberapa faktor.

II. LITERARUER REVIEW

2.1 Media Sosial

Media sosial adalah media *online* (daring) yang dimanfaatkan sebagai sarana pergaulan sosial secara online di internet. Di media sosial, para penggunanya dapat saling berkomunikasi, berinteraksi, berbagi, *networking*, dan berbagai kegiatan lainnya. Media sosial menggunakan teknologi berbasis web atau aplikasi yang dapat mengubah suatu komunikasi ke dalam bentuk dialog interaktif. Beberapa contoh media sosial yang banyak digunakan adalah YouTube, Facebook, Blog, Twitter, dan lain-lain. Twitter merupakan salah satu media sosial yang dapat menjadi sumber yang sangat bermanfaat untuk mengumpulkan data yang digunakan dalam menelusuri hal-hal yang sedang berkembang.

2.2 Teknik Penggalian Web

Perkembangan informasi berbasis web pada masa kini telah berkembang dengan sangat pesat dan dengan jumlah data yang besar. Data yang mengandung berbagai macam informasi tercampur dengan data lainnya dalam bentuk dan volume yang bervariasi, sehingga diperlukan cara yang efktif dalam menggali informasi. Cara menggali informasi tersebut adalah *Data Mining* dan *Text Mining.*

2.3 Text Mining

Text mining dapat didefinisikan sebagai suatu proses menggali informasi dimana user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen-komponen dalam data mining yang salah satunya adalah kategorisasi. Masalah yang sering muncul dalam text mining diantaranya: jumlah data yang besar, adanya data yang tidak diinginkan(*noise*), dan data yang tidak terstruktur. Tujuan dari penggunaan text mining adalah untuk menentukan informasi yang diinginkan pengguna dari dokumen-dokumen. Selain itu dapat digunakan untuk mengkategorikan dan mengelompokkan teks. Tahapan pada *text mining* :

1. Tahap tokenizing atau parsing adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya.
2. Tahap filtering adalah tahap mengambil kata-kata penting dari hasil token.
3. Tahap stemming adalah tahap mencari *root* kata dari tiap kata hasil filtering.
4. Tahap tagging adalah tahap mencari bentuk awal dari tiap kata lampau atau kata hasil stemming.
5. Tahap analyzing merupakan tahap penentuan seberapa jauh keterhubungan antar kata-kata antar dokumen yang ada.

2.4 Pengambilan Data

Tahap pertama dari text mining yaitu pengambilan data yang salah satunya dapat dilakukan dengan menggunakan library twitterscraper. Twitterscraper mampu mengambil data tanpa batas yang ditentukan, tidak menggunakan API Twitter karena hanya mampu mengambil 72.000 tweet/jam.

III. METODOLOGI

Terdapat beberapa tahap yang dilakukan dalam penelitian ini, seperti ditampilkan pada gambar 1.

Gambar 1. Metode Penelitian

Gambar 1 menampilkan tahapan utama yang dilakukan dalam penelitian ini. Tahap pertama merupakan proses input data berupa kalimat opini yang diperoleh dari tweet seputar malas sekolah melalui media sosial twitter. Tahap kedua, melakukan data preprocessing yang terdiri dari: *tokenizing, cleansing, normalization* dan *case folding*. Kalimat opini yang telah dilakukan proses preprocessing berupa kalimat dalam bentuk teks Bahasa Indonesia yang mudah dipahami. Tahap ketiga, kalimat diproses dengan melakukan seleksi fitur Chi Square. Tahap keempat, dilakukan proses klasifikasi menggunakan algoritma Naïve Bayes Classifier. Tahap terakhir menghasilkan data yang terklasifikasi.

IV. HASIL DAN PEMBAHASAN

Hasil pengambilan tweet dengan keyword "sekolah" diperoleh 773.225 data dalam format json dengan ukuran 562 MB. Data yang diperoleh merupakan seluruh data pada tahun 2000 – 2019 dengan limit 1.000.000 tweet.

```
twitterscraper sekolah -l 1000000 –o
sekolah.json
```

4.1 Cleaning Data

Pada penelitian ini data yang diambil hanya data *full name* dan *text (tweet)*. Ukuran data seteleh cleaning menjadi 89 MB. Tahapan pada Cleaning data :

1. Mengambil data .json dan diubah kedalam bentuk array

```
$filenya =
```

```
file_get_contents("sekolah.json");
$data = json_decode($filenya,true);
```

2. Data yang sudah diubah dalam bentuk array hanya diambil tweet dan nama saja, dengan menggunakan pengulangan. Data hasil cleaning dimasukkan ke dalam database noSQL (MongoDB).

```
$m = new MongoClient();
$db = $m->sekolah;
$collection = $db->sekolah;
for($a=0;$a<count($data);$a++){
 $b=$a+1;
 $databuatinput[$a]['fullname'] =
 $data[$a]['fullname'];
 $databuatinput[$a]['text'] =
 $data[$a]['text'];
 $collection->insert($databuatinput[$a]);
}
```

4.2 Filtering

Data yang sebelumnya dimasukkan ke dalam database MongoDB dipanggil kembali dan dilakukan proses filtering. Filtering merupakan proses identifikasi data yang relevan dengan data yang dibutuhkan. Cara melakukan proses filtering:

1. Mencari kata yang mengandung faktor tertentu, seperti: ulangan, suru, kurikulum, tugas, dan fullday.

```
for ($x=0;$x<count($data);$x++){
 for($y=0;$y<count($kondisi);$y++){
  if(strstr($data[$x],$kondisi[$y]){
       ...} } }
```

2. Mencari kata yang berkaitan dengan sifat malas terhadap faktor yang telah ditentukan. Seperti kata positif : tetap semangat, tetap sekolah, dan lainnya. Kata negatif: malas, tidak masuk, tidak sekolah, dan lainnya. Kata-kata tersebut dibuat dalam bentuk array.

```
for ($x=0;$x<count($data);$x++){
 for($y=0;$y<count($kondisi);$y++){
  if(strstr($data[$x],$kondisi[$y]){
     for($z=0;$z<count($negatif);$z++){
          if(strstr($data[$a]['text']),$ne
          gatif[$z])) {
          $ket ="Malas"; $malas++;
       }
     }
     for($z=0;$z<count($positif);$z++){
        if(strstr($data[$a]['text']),$po
        sitif [$z])) {$ket ="Tidak Malas";
       $tidakmalas++;
       }
     }
  }
 }
}
```

Jumlah tweet menjadi 150.178 Contoh hasil filtering ditampilkan pada tabel 1.

Tabel 1. Data Hasil Filtering

Ulangan	Guru	Kurikulum	Tugas	Full Day	Malas
Tidak	Ya	Tidak	Tidak	Tidak	**Tidak**
Tidak	Ya	Tidak	Tidak	Tidak	**Tidak**
Tidak	Ya	Tidak	Tidak	Tidak	**Tidak**
Ya	Tidak	Tidak	Tidak	Ya	**Ya**

4.3 Perhitungan dengan Naive Bayes

Algoritma naive bayes merupakan sebuah metode klasifikasi menggunakan metode probabilitas dan statistik. Ciri utama dari Naive Bayes Classifier adalah asumsi yang kuat akan independensi dari masing-masing kondisi atau kejadian. Adapun Tahapan dari proses algoritma Naive Bayes adalah :

1. Mengitung jumlah kelas/tabel
2. Menghitung jumlah kasus per kelas
3. Kalikan semua variabel kelas
4. Bandingkan hasil per kelas

Proses klasifikasi Naive Bayes menggunakan rumus:

$$P(H|X) = \frac{P(X|H)}{P(X)}.P(H)$$

Keterangan :

- **X** : Data dengan *class* yang belum diketahui
- **H** : Hipotesis data merupakan suatu *class* spesifik
- **P(H|X)** : Probabilitas hipotesis H berdasar kondisi X (*posteriori probabilitas*)
- **P(H)** : Probabilitas hipotesis H (*prior probabilitas*)
- **P(X|H)** : Probabilitas X berdasarkan kondisi pada hipotesis H
- **P(X)** : Probabilitas X

Tabel 2. Hasil perhitungan dengan Naïve Bayes

Rasio Tabel			
Faktor	Malas		Total
	Ya	Tidak	
Ulangan	33616/ 115291	9053/ 34887	**42669/ 150178**
Guru	8180/ 115291	16383/ 34887	**24563/ 150178**
Kurikulum	23397/ 115291	1166/ 34887	**24563/ 150178**
Tugas	17549/ 115291	7014/ 34887	**24563/ 150178**
Fullday	32549/ 115291	1271/ 34887	**33820/ 150178**
	115291 / 150178	**34887 / 150178**	

P-ISSN 2477-3891
E-ISSN 2615-4765

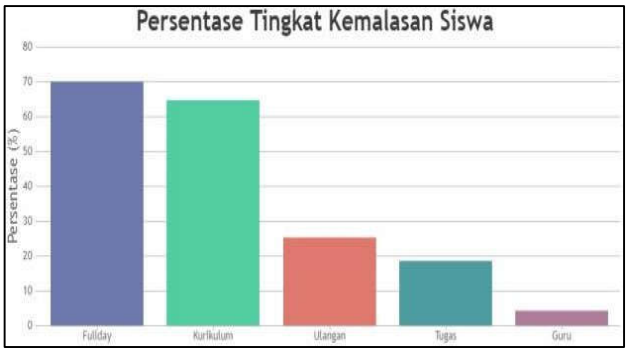
Tabel 3. Frequensi

Frekuensi Tabel		
Faktor	Malas	
	Ya	Tidak
Ulangan	33616	9053
Guru	8180	16383
Kurikulum	23397	1166
Tugas	17549	7014
Fullday	32549	1271

Peluang(Malas | Ulangan)
= (33616/115291 x 42669/150178) : 115291/150178
= 0.10791148959911
Peluang(Tidak Malas | Ulangan)
= (9053/34887 x 42669/150178) : 34887/150178
= 0.31737866911385
Peluang(Malas | Guru)
= (8180/115291 x 24563/150178) : 115291/150178
= 0.015116244212942
Peluang(Tidak Malas | Guru)
= (16383/34887 x 24563/150178) : 34887/150178
= 0.33063405921805
Peluang(Malas | Kurikulum)
= (23397/115291 x 24563/150178) :115291/150178
= 0.043236523942566
Peluang(Tidak Malas | Kurikulum)
= (1166/34887 x 24563/150178) : 34887/150178
= 0.023531667768312
Peluang(Malas | Tugas)
= (17549/115291 x 24563/150178) : 115291/150178
= 0.032429702896444
Peluang(Tidak Malas | Tugas)
= (7014/34887 x 24563/150178) : 34887/150178
= 0.14155327420835
Peluang(Malas | Fullday)
= (32549/115291 x 33820/150178) : 115291/150178
= 0.082817180571846
Peluang(Tidak Malas | Fullday)
= (1271/34887 x 33820/150178) : 34887/150178
= 0.035317658537017

Tabel 4. Hasil Perhitungan Akhir

Faktor	Malas	Tidak Malas	Tingkat Kemalasan
Ulangan	0.107911 48959911	0.317378669 11385	25.37361549 2463%
Guru	0.015116 244212942	0.330634059 21805	4.372011842 9219%
Kurikulum	0.043236 523942566	0.023531667 768312	64.75617031 8032%
Tugas	0.032429 702896444	0.141553274 20835	18.63958384 6706%
Fullday	0.082817 180571846	0.035317658 537017	70.10394325 38%



Gambar 2.Visualisasi Presentase Tingkat Kemalasan

Gambar 2 menampilkan visualisasi presentasi tingkat kemalasan siswa berdasarkan hasil akhir perhitungan. Dari gambar tersebut diketahui bahwa tingkat kemalasan tertinggi berada pada parameter "full day".

V. KESIMPULAN DAN SARAN

Penggunaan naive bayes pada kasus analisis data twitter faktor kemalasan siswa terbilang cocok, dikarenakan hasil perhitungan sesuai. Kasus yang diambil mengenai kemalasan siswapun ternyata layak diangkat, karena banyaknya keluhan siswa dalam menghadapi pendidikannya. Hasil ini dapat menjadi acuan untuk perkembangan di sektor pendidikan.

DAFTAR PUSTAKA

[1] Joshi, M., & Vala, H. (2014). Opinion Mining For Sentimental Data Classification. International Journal of Research in Information Technology, 3(1), 1-13.
[2] Ling, J., N Kencana, I. P. E., & Oka, T.B. (2014). Analisis Sentimen Menggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Chi Square. E-Jurnal Matematika, 3(3), 92- 99

[3] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures On Human Language Technologies, 5(1), 1-167.
[4] Nurrohmat, M. A. (2015). Aplikasi Pemrediksi Masa Studi dan Predikat Kelulusan Mahasiswa Informatika Universitas Muhammadiyah Surakarta Menggunakan Metode Naïve Bayes. Khazanah Informatika, (1).
[5] Patel, T., Undavia, J., & Patel, A. (2015), Sentiment Analysis of Parents Feedback for Educational Institutes, International Journal of Innovative and Emerging Research in Engineering, 2(3).
[6] Ting, S. L., Ip, W. H., & Tsang, A. H. (2011). Is Naïve Bayes a Good Classifier for Document Classification?. International Journal of Software Engineering and Its Applications, 5(3), 37-46.