SISTEM FILTER KATA UMPATAN DENGAN MENGEMBANGKAN SENTIWORD BAHASA INDONESIA MENGGUNAKAN ALGORITMA NAÏVE BAYES CLASSIFIER

Yoga Pramana Putra[1], Ridwan Rismanto[2], Istaghna Faza Kamil[3]

[1,2,3] Program Studi Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Malang
[1]pramanay@polinema.ac.id, [2]rismanto@polinema.ac.id, [3]istaghnafaza@gmail.com

***Abstrak*— Pendidikan etika sangatlah penting sebagai pondasi dalam perkembangan seorang manusia. Pendidikan tentang komunikasi dapat diajarkan dalam setiap aspek kehidupan, di dalam bersosial media nilai-nilai etika dapat di terapkan dalam setiap yang kita ucapkan, seperti pemilihan bahasa ataupun kata dalam berkomentar di social media. Saat ini banyak sekali *tweet* bermunculan baik yang bersifat positif maupun negative, *tweet* yang sifatnya negatif dapat menimbulkan efek buruk bagi publik, perlu mendapat perhatian serius agar tidak mempengaruhi generasi penerus bangsa dengan cara memberi solusi mengembangkan sistem filter kata umpatan menggunakan *sentiword* bahasa Indonesia dengan Algoritma *Naive Bayes Classifier*. *Naive Bayes Classifier* merupakan salah satu metode machine learning yang memanfaatkan perhitungan probabilitas dan statistic, *Naive Bayes Classifier* menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap klasifikasi, Klasifikasi *Naive Bayes* juga memperlihatkan tingginya akurasi dan cepat ketika digunakan untuk dataset dengan jumlah besar. Penelitian ini menggunakan 200 data *training* serta 20 data *testing* dan mendapatkan akurasi sebesar 90% dalam menggunakan algoritma *Naïve Bayes Classifier*.**

Kata kunci*— *Text Minning, Naïve Bayes, Umpatan.

I. PENDAHULUAN

Memasuki zaman informasi, kita menyaksikan bagaimana media sosial memiliki kekuatan dominan dalam mempengaruhi setiap kehidupan manusia. Dengan perkembangan teknologi informasi dan komunikasi, Internet (website) atau media sosial menjadi komunikasi interaktif sekaligus komunikasi massa.

Berdasarkan dampak dari perkembangan teknologi tersebut, maka masyarakat di Indonesia masuk ke dalam era masyarakat informasi. Dengan arus informasi yang pesat dan besar secara kapasitas, dalam hal ini tentunya perlu memerhatikan pengendalian yang tepat sebagai upaya dari perlindungan terhadap informasi yang kurang baik tentunya [2].

Saat ini banyak sekali tweet bermunculan baik yang bersifat positif maupun negatif. Untuk tweet yang sifatnya positif dan bermanfaat, tidak menjadi masalah apabila publik menerimanya. Namun apabila ada tweet yang sifatnya negatif dan menimbulkan efek buruk bagi publik, perlu mendapat perhatian serius agar tidak mempengaruhi generasi penerus bangsa. Berkomentar memang tidak dapat dikontrol dalam komunikasi di sosial media, namun hal ini dapat diminimalisir dengan memberikan pendidikan sejak dini mengenai etika dalam menggunakan media sosial yang baik dan benar.

Pendidikan etika sangatlah penting sebagai pondasi dalam perkembangan seorang manusia. Pendidikan tentang komunikasi dapat diajarkan dalam setiap aspek kehidupan, di dalam bersosial media, nilai-nilai etika dapat diajarkan dalam setiap yang kita ucapkan, seperti pemilihan bahasa ataupun kata dalam berkomentar. Mengapa dalam berkomentar? Karena saat kita berkomentar, semua yang kita ucapkan itulah yang akan menggambarkan tingkat intelektualitas pribadi seseorang.

Dengan adanya system ini bertujuan untuk memberi solusi dengan cara mengembangkan sistem filter kata umpatan menggunakan sentiword bahasa Indonesia dengan Algoritma Naive Bayes Classifier.

II. TINJAUAN PUSTAKA

2.1. *Text Minning*

Text Mining adalah proses yang mencoba mengekstrak informasi berguna dari teks natural language. Hal itu bisa diartikan sebagai proses menganalisa teks untuk mengekstrak informasi yang berguna untuk tujuan tertentu. *Text mining* memiliki tujuan dan menggunakan proses yang sama dengan data mining, namun memiliki input yang berbeda (Imam, dkk., 2017).

2.2. *Text Preprocessing*

Text Preprocessing berfungsi untuk mengubah data teks yang tidak terstruktur atau sembarang menjadi data yang terstruktur [3]. Serta menormalkan teks menjadi bentuk yang tepat, sehingga data yang sebelumnya masih mentah dapat diproses menggunakan sentiment analysis menjadi data yang berkualitas [4]. *Preprocessing* terdiri dari tokenisasi dan *filtering (Stopword Removal)*. Tokenisasi untuk memisahkan dokumen menurut tokennya. Filtering yaitu membuang kata-kata yang tidak berguna dalam proses klasifikasi [8]. Tahapan *preprocessing* yang dilakukan adalah sebagai berikut :

1. *Tokenizing*

   Tokenizing adalah proses pemotongan sebuah dokumen menjadi bagian-bagian, yang disebut dengan token. Pada saat bersamaan *tokenizing* juga berfungsi untuk membuang karakter tertentu yang dianggap sebagai tanda baca [3].
2. *Stopword Removal*

   Stopword removal adalah proses penghilangan kata-kata yang tidak berkontribusi banyak pada isi dokumen. Kata-kata yang termasuk ke dalam *stopword* dihilangkan karena memberikan pengaruh yang tidak baik dalam

proses *text minning* seperti kata-kata “saya”, “kamu”, “dia”, dan lain-lain [3].

2.3. Naïve Bayes Classifier

Naive Bayes Classifier merupakan salah satu metode machine learning yang memanfaatkan perhitungan probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. Metode *Naive Bayes Classifier* menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap klasifikasi. Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan *vocabulary*, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin dapat menjadi representasi dokumen. Selanjutnya adalah penentuan probabilitas prior bagi tiap kategori berdasarkan sampel dokumen [5].

Naive Bayes Classifier merupakan salah satu metode yang populer untuk keperluan data mining karena penggunaannya yang mudah dan dalam pemroresan memiliki waktu yang cepat, mudah diimplementasikan dengan strukturnya yang cukup sederhana dan untuk tingkat efektifitasnya memiliki efektifitas yang tinggi. Klasifikasi *Naive Bayes* juga memperlihatkan tingginya akurasi dan cepat ketika digunakan untuk dataset dengan jumlah besar [7]. *Naive Bayes* merupakan algoritma yang sering digunakan dalam pengkategorian teks, dimana konsep dasarnya adalah menggabungkan probabilitas kata-kata dan kategori sebuah dokumen

Klasifikasi yang harus dilakukan yaitu membuat tabel khusus untuk mengetahui kemunculan kata disetiap kalimat yang sudah memiliki label positif dan negatifi, Setelah itu pada tahap perhitungan pertama dapat dituliskan seperti persamaan dibawah ini:

$$P(C_i) = \frac{f_d(C_i)}{|D|}$$

Keterangan:
$f_d(C_i)$: Jumlah dokumen yang memiliki kategori C_i
$|D|$: Jumlah seluruh training dokumen

Proses selanjutnya yaitu dengan menerapkan rumus perhitungan Naive Bayes yang bisa dituliskan dengan persamaan:

$$P(W_k|C_i) = \frac{n_k + 1}{n + |Vocabulary|}$$

Keterangan:
n_k : Nilai kemunculan kata pada kategori C_i
n : Jumlah keseluruhan kata pada kategori C_i
$|Vocabulary|$: Jumlah keseluruhan kata

Kemudian dilakukan proses pemilihan kelas yang optimal maka dipilih nilai peluang terbesar dari setiap probabilitas kelas yang ada. Maka didapatkan rumus untuk memilih nilai terbesar seperti pada persamaan berikut:

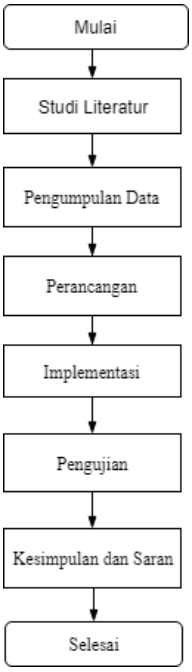
$$V_{NB} = argmax\ P(V_j) \prod_{w \in words} P(W|V_j)$$

Keterangan :
V_{NB} : semua kategori yang diujikan V
j : kategori, dengan :
P($W \mid V_j$) : probabilitias W pada kategori V_j

III. METODOLOGI

Metodologi penelitian menjelaskan bagaimana langkah-langkah atau tahapantahapan yang akan dilakukan dalam penelitian untuk dapat menjawab perumusan masalah penelitian.

Tahapan penelitian dalam implementasi *Single Pass Clustering* pada *pre-processing* temu kembali koleksi teks dilakukan seperti pada gambar 1 sebagai berikut:



Gambar 1. Tahapan Penelitian

3.1 Studi Literatur

Studi literatur digunakan untuk mengumpulkan informasi mengenai data-data yang diperlukan untuk analisa sistem dan *dataset*. Serta perhitungan dalam metode algoritma *Naïve Bayes Classifier* yang digunakan pada aplikasi. Studi literatur ini dilakukan dengan cara menggunakan internet, jurnal serta buku untuk mendapatkan referensi yang terkait dengan apa yang

dibutuhkan. Sumber studi literatur diperoleh baik dari dalam maupun luar negeri.

3.2 Pengumpulan Data

Data yang digunakan dalam penelitian ini berupa kalimat yang berasal dari *crawling*. *Crawling* data ini di ambil melalui sosial media *Twitter*, dimana hasil pemngambilan data tersebut berupa sekumpulan kalimat atau *tweet*

3.3 Perancangan

Dalam tahap ini, ditentukan arsitektur dari sistem dan perancangan antarmuka dari sistem yang akan dibuat..

3.4 Implemetasi

Pada tahapan implementasi ini akan dilakukan pembuatan modul-modul yang telah dirancang dalam tahap perancangan kedalam bahasa pemrograman.

3.5 Pengujian

Pengujian merupakan tahapan dimana sistem akan dijalankan. Tahap pengujian diperlukan untuk menjadi ukuran bahwa sistem dapat dijalankan sesuai dengan tujuan.

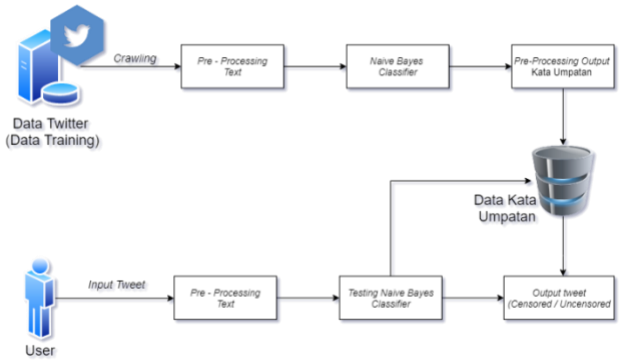
3.6 Kesimpulan

Tahap ini dilakukan untuk mendapatkan kesimpulan yang bisa diambil dari penelitian yang dilakukan.

IV. PERANCANGAN

4.1. Analisa Data

Proses yang akan dibangun diterapkan dalam sistem filter kalimat umpatan. Nantinya proses yang akan berjalan dalam sistem akan berkerja seperti pada gambar 2.



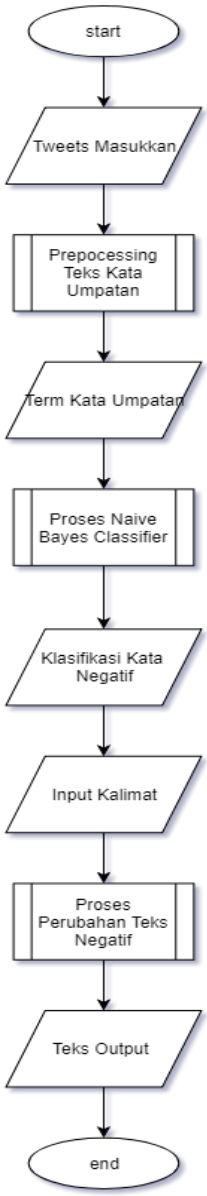
Gambar 2. Arsitektur Sistem

Pada sistem filter kata umpatan, data *twitter* hasil *crawling* di gunakan sebagai data *training*. Setelah itu dilakukan *preprocessing* untuk perhitungan *Naïve Bayes* dalam penentuan kata positif atau kata umpatan, serta dilakukan *preprocessing* kembali untuk mendapatkan *output* sebuah *database (corpus)* kata umpatan.

Pada aplikasi sistem filter kata umpatan, data *testing* di peroleh dari *input tweet user*. Setelah itu dilakukan *preprocessing tweet* untuk perhitungan *Naïve Bayes* dalam penentuan kalimat positif atau kalimat umpatan, jika data tes menyatakan kalimat umpatan maka dilakukan pencocokan kata dalam *database* kata umpatan untuk penyensoran kata tersebut. Jika kalimat dinyatakan kalimat positif maka tidak ada kata yang di sensor.

4.2. **Perancangan Proses**

Pada perancangan proses dilakukan untuk menjelaskan bagaimana proses yang akan bekerja dalam sistem seperti pada gambar 3.



Gambar 3. Flowchart cara kerja sistem

V. IMPLEMENTASI

Setelah dilakukan analisa dan perancangan terhadap sistem yang akan dibangun, maka tahap selanjutnya adalah tahap implementasi. Tahap implementasi ini dilakukan terhadap dua bagian dari sistem, yakni untuk *database* dan sistem itu sendiri.

VI. PENGUJIAN

Untuk mengetahui performa algoritma yang digunakan yaitu *Naïve Bayes Classifier* maka perlu dilakukan pengujian tingkat akurasi. Tahap yang dilakukan dalam analisis tingkat akurasi sebagai berikut :

- Merekap klasifikasi kalimat yang dilakukan melalui testing menggunakan aplikasi.
- Merekap klasifikasi kalimat yang dilakukan melalui testing secara manual.
- Melakukan perbandingan antara hasil klasifikasi aplikasi dengan hasil klasifikasi secara manual.
- Melakukan perhitungan tingkat akurasi algoritma hasil yang sama dan hasil berbeda dari kedua testing yang dilakukan dengan rumus sebagai berikut

$$Accuracy = \frac{\Sigma data\ benar}{\Sigma semua\ data} \text{ x } 100$$

VII. KESIMPULAN DAN SARAN

Dengan proses sistem yang sedemikian rupa seperti pada perancangan bisa diambil kesimpulan Sistem Klasifikasi filter kata umpatan berjalan dengan baik secara fungsional dan menghasilkan *output* yang diharapkan.

Metode *Naïve Bayes Classifier* secara efektif dan cepat mampu memberikan hasil yang diharapkan dengan tingkat akurasi sebesar 90% dari 20 data training yang digunakan.

Saran untuk penyempurnaan sistem filter kata umpatan ini ialah data training sebaiknya lebih banyak karena semakin banyak data training maka akurasi semakin baik

DAFTAR PUSTAKA

[1] Akhmad Pandhu Wijaya, Heru Agus Santoso. 2016." *Naive Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government*" Jurusan Teknik Informatika, Universitas Dian Nuswantoro Semarang.

[2] Andi Faisal Bakti, & Venny Eka Meidasari. 2014. "*Tantangan dan Peluang Pendidikan Komunikasi dan Penyiaran Islam*". Jurusan Komunikasi dan Penyiaran Islam, UIN Syarif Hidayatullah, Jakarta.

[3] Chaerul Sutami. 2015. "*Perbandingan Metode Klasifikasi Naive Bayes Classifier Dan Lexicon Based Dalam Analisis Sentimen* ". Jurusan Teknik Informatika, Fakultas Teknik, Universitas Widyatama Bandung. https://repository.widyatama.ac.id/xmlui/handle/123456789/5864

[4] Iin Kusumawati. 2017 "*Analisa Sentimen Menggunakan Lexicon Based Untuk Melihat Persepsi Masyarakat Terhadap Kenaikan Harga Rokok Pada Media Sosial Twitter*" Program Studi Informatika Fakultas Komunikasi Dan Informatika Universitas Muhammadiyah Surakarta.

[5] Imam Fahrur Rozi , Ridwan Rismanto , Danis Karmanita, 2017. "*Analisis Sentimen Tentang Respon Masyarakat Terhadap Transportasi Umum Online Dan Konvensional Menggunakan Naïve Bayes Classifier Dan Lexicon Based*". Skripsi Program Studi Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Malang, Malang.

[6] Jaka Eka Sembodo , Erwin Budi Setiawan , ZK Abdurahman Baizal , 2016. "*Data Crawling Otomatis pada Twitter*". Computational Science, School of Computing, Telkom University.

[7] Prananda Antinasari , Rizal Setya Perdana , M. Ali Fauzi. 2017. "*Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku*". Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Fakultas Ilmu Komputer Universitas Brawijaya

[8] Shima Fanissa, M. Ali Fauzi, Sigit Adinugroho. 2018. "*Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naïve Bayes dan Seleksi Fitur Query Expansion Ranking*". Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Fakultas Ilmu Komputer Universitas Brawijaya