TEXT MINING DALAM ANALISIS SENTIMEN ASURANSI MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER*

Luthfia Oktasari[*], Yulison Herry Chrisnanto, Rezki Yuniarti
Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Jenderal Achmad Yani
Jalan Terusan Jenderal Sudirman PO BOX 148 Cimahi - Bandung, Telp./Fax: (022) 6656190
[*]e-mail: luthfiaoktas@gmail.com

Abstrak

Text mining adalah salah satu teknik penambangan data yang berupa teks. Analisis sentimen didefinisikan sebagai ilmu untuk melakukan analisa dari pendapat, sikap, emosi seseorang ke dalam bahasa tertulis. Salah satu media yang dapat digunakan dalam melakukan analisis sentimen yaitu melalui media sosial yang menjadi sarana dalam menunjang perusahaan atau organisasi untuk kegiatan penyampaian informasi kepada masyarakat seperti yang dapat dilihat pada perusahaan penyedia layanan asuransi. Pendapat masyarakat terhadap jasa penyedia asuransi banyak disampaikan di media sosial salah satunya pada akun jejaring sosial facebook. Analisis sentimen dibutuhkan untuk mendapatkan informasi yang dapat digunakan sebagai manajemen reputasi dan sarana evaluasi bagi perusahaan. Pada penelitian ini dibuat sistem dengan tujuan untuk menghasilkan informasi sentimen masyarakat yang mengarah ke sentimen positif dan negatif mengenai asuransi dengan menggunakan metode Naïve Bayes Classifier. Dari Pengujian yang telah dilakukan pada penelitian dengan pre-proses,pendekatan rule based dan klasifikasi menggunakan metode Naïve Bayes Classifier diperoleh hasil akurasi sebesar 95%.

Kata kunci: *Analisis sentimen, Asuransi, Naïve Bayes Classifier.*

1. PENDAHULUAN

Asuransi telah menjadi bisnis yang berkembang pesat saat ini. Sebagian masyarakat telah menganggap penting asuransi sebagai bentuk pengurangan risiko untuk meminimalisir kerugian yang timbul dari suatu kejadian yang tidak diinginkan, banyak produk asuransi yang ditawarkan oleh penyedia layanan jasa asuransi diantaranya asuransi kesehatan, pendidikan, kecelakaan dan lainnya. Berdasarkan data Asosiasi Asuransi Jiwa Indonesia (AAJI) pada tahun 2014 di Indonesia terdapat sekitar 62 juta masyarakat yang telah menggunakan produk asuransi.

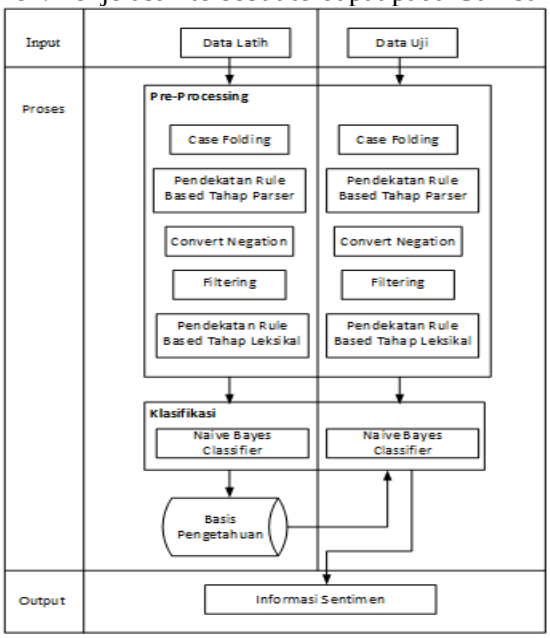
Media sosial merupakan media penyampaian informasi yang banyak menjadi pilihan masyarakat, dengan adanya media sosial pengguna dapat memanfaatkan akun yang dimiliki untuk mengungkapkan perasaan baik atau buruk terhadap suatu produk maupun pelayanan dibandingkan harus bicara kepada *customer service*, objek atau orang yang mempunyai keterkaitan dengan masalahnya secara langsung. Media sosial juga menjadi sarana yang menunjang perusahaan atau organisasi dalam kegiatan penyampaian informasi, seperti yang dapat dilihat pada perusahaan penyedia layanan jasa asuransi. Pendapat masyarakat terhadap jasa penyedia asuransi banyak disampaikan di media sosial salah satunya pada akun jejaring sosial *facebook*. Dari ulasan masyarakat pihak penyedia layanan jasa asuransi dapat mengetahui respon sentimen masyarakat terhadap asuransinya, namun pihak asuransi masih mengalami kesulitan dalam menangkap aspirasi yang di berikan masyarakat terutama pendapat yang disampaikan melalui media sosial.

Analisis sentimen banyak digunakan sebagai acuan manajemen reputasi yang dapat membantu memperbaiki kinerja perusahaan. Terdapat beberapa penelitian yang telah membahas mengenai analisis sentimen diantaranya analisis sentimen terhadap acara televisi berdasarkan opini publik (Aditia Rakhmat, 2014). Penerapan character N-gram untuk sentiment analysis review hotel menggunakan algoritma naive bayes (Elly Indrayuni, 2015). Penelitian lainnya menggabungkan metode naive bayes dengan metode lain yaitu penelitian analisis sentimen data presiden Jokowi dengan pre-processing normalisasi dan stemming menggunakan metode naive bayes dan SVM (Nurirwan Saputra, 2015**).** Penelitian lainnya menggunakan metode selain naive bayes yaitu pada penelitian analisis sentimen twitter untuk teks berbahasa Indonesia dengan maximum entropy dan support vector machine (Noviah Dwi Putranti, 2014). Pada penelitian ini dilakukan penelitian mengenai analisis sentimen terhadap asuransi berdasarkan data ulasan di *facebook* menggunakan

metode *naïve bayes classifier* untuk membangun sistem yang dapat memberi informasi terhadap penilaian sentimen yang mengarah ke sentimen positif dan sentimen negatif.

2. METODOLOGI

Metodologi pada penelitian ini terdiri dari beberapa tahapan, diantaranya tahap identifikasi data masukan, pra-proses, klasifikasi dengan metode *Naïve Bayes Classifier(NBC)*, dan keluaran. Data masukan berupa data latih dan data uji yang digunakan bersumber dari *fanpage* Prudential Indonesia di *facebook*. Pada tahap pra-proses terdiri dari case folding, convert negation, filtering serta pendekatan rule based yaitu parser dan leksikal. Data yang telah diproses pada pra-proses diproses kembali dengan menggunakan metode Naïve Bayes Classifier. Keluaran dari proses tersebut yaitu informasi sentimen. Penjelasan tersebut terdapat pada Gambar 1.



Gambar 1. Rancangan sistem

a. *Case Folding, Convert Negation. Filtering*

Tahap *case folding* merupakan tahapan untuk merubah seluruh huruf kapital yang terdapat pada dokumen menjadi huruf kecil. Tahap *convert negation* merupakan tahapan untuk mengkonversi teks yang mengandung kata negasi yang terdiri dari kata "kurang" dan "tidak". Tahap *filtering* merupakan tahapan mengambil kata-kata yang penting (*wordlist*) dari hasil proses sebelumnya.

b. *Pendekatan Rule Based(Tahap parser dan leksikal)*

Tahap *parser* merupakan tahapan untuk pemotongan *string input* berdasarkan kata yang menyusunnya. Tahap leksikal merupakan tahapan membuat kamus kata berbahasa Inggris yang akan dialihbahasakan kedalam bahasa Indonesia (Ema Utami, 2007).

c. Metode Naïve Bayes Classifier

Salah satu metode klasifikasi yang dapat digunakan adalah metode *Naïve Bayes* yang sering disebut dengan *Naive Bayes Classifier (NBC)*. Kelebihan metode *NBC* adalah sederhana tetapi memiliki akurasi yang tinggi. Berdasarkan penelitian yang berjudul sentimen analysis untuk memanfaatkan saran kuesioner dalam evaluasi pembelajaran dengan menggunakan *naïve bayes classifier* menyebutkan metode *NBC* dapat memperoleh akurasi mencapai 85,95% (Hamzah, 2014). *Naive Bayes Clasifier* merupakan salah satu metode *machine learning* yang menggunakan perhitungan probabilitas. keuntungan dari penggunaan metode *Naïve Bayes* adalah hanya membutuhkan sejumlah kecil data pelatihan untuk memperkirakan parameter (sarana dan varians dari variabel-variabel) yang diperlukan untuk klasifikasi. Karena variabel independen diasumsikan, hanya varian dari variabel untuk masing-masing kelas harus ditentukan dan tidak seluruh matriks

ISBN 978-602-99334-5-1

kovariansi. Konsep dasar yang digunakan oleh *Bayes* adalah Teorema peluang bersyarat *Bayes* berikut:

$$P(A|B) = P(A)\ P(B|A) \quad (1)$$

Peluang kejadian A bersyarat B ditentukan dari peluang A dan peluang B bersarat A. Persamaan ini dikembangkan menjadi persamaan berikut:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

Pada pengaplikasiannya persamaan ini dapat digunakan untuk klasifikasi dokumen dengan melakukan perhitungan nilai probabilitas. Klasifikasi dilakukan untuk menentukan kategori dari suatu dokumen. Persamaan ini dapat diubah kedalam bentuk persamaan berikut:

$$P(Ci|D) = \frac{P(D|Ci)P(Ci)}{P(D)} \quad (3)$$

Metode *Naive Bayes* menganggap setiap variabel berdiri bebas satu sama lain dan tidak ada keterkaitan dengan variabel lain, sehingga sebuah dokumen akan dianggap sebagai kumpulan dari kata-kata yang menyusun dokumen tersebut, dan tidak memperhatikan urutan kemunculan kata pada dokumen. Sehingga perhitungan probabilitas dapat dianggap sebagai hasil perkalian dari probabilitas kemunculan kata-kata pada dokumen. Persamaan *Naive Bayes Classifier* dapat dilihat pada persamaan-persamaan berikut:

$$P(Ci) = \frac{fD(Ci)}{|D|} \quad (4)$$

Keterangan :
P(Ci) : Probabilitas dari suatu kategori dokumen.
fD(Ci) : Frekuensi dokumen yang memiliki kategori *Ci*.
|D| : Jumlah seluruh dokumen latih.

$$P(Wkj|Ci) = \frac{f(Wkj|Ci) + 1}{f(Ci) + |W|} \quad (5)$$

keterangan:
P(Wkj|Ci):Probabilitas kemunculan kata Wkj pada suatu dokumen dengan kategori kelas Ci.
Wkj : frekuensi kata ke-k pada setiap kategori.
W : jumlah kata pada dokumen test.
f(Ci) : frekuensi dokumen berkategori kelas Ci.
Pada persamaan ini terdapat suatu penambahan angka 1 pada pembilang untuk mengantisipasi jika terdapat suatu kata pada dokumen uji yang ber nilai nol (0) karena tidak terdapat pada dokumen latih.

3. HASIL DAN PEMBAHASAN

3.1. Hasil

Berdasarkan dari hasil analisis dan pengujian yang telah dilakukan dengan 100 data uji, penggunaan metode *Naïve Bayes Classifier* (*NBC*) dengan pre-processing yang digunakan menghasilkan rata-rata akurasi 95%. Dengan menggunakan *Naïve Bayes Classifier (NBC)* pada sistem ini dapat dihasilkan klasifikasi sentimen positif dan negatif. Contoh hasil yang didapatkan dari penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Hasil Klasifikasi Analisis Sentimen Asuransi

No	Teks Ulasan	Hasil Klasifikasi Sentimen
1	Saya sdh 6 tahun ikut. Saat mengklaim polis knapa susah sekali . Sya hrus nunggu brpa lama spya uang bisa ditransfer hm apa saya juga termasuk korban seperti kebanyakan yg sya liat	Negatif
2	Asuransi yang oke bngtttt pkony saya cinta prudential .yg lain sieh ??	Positif

3.2. Pembahasan

Setelah melalui tahap pra-proses yang terdiri dari *case folding, pendekatan rule based(parser), convert negation,filtering dan pendekatan rule based(leksikal),* dilakukan perhitungan menggunakan metode Naïve Bayes Classifier dengan menghitung nilai probabilitas teks berdasarkan basis pengetahuan yang dimiliki. Terdapat dua contoh teks yang akan diklasifikasi seperti yang dapat dilihat pada Tabel 2.

Tabel 2. Contoh Teks Uji

No	Teks Uji
1	Saya sdh 6 tahun ikut. Saat mengklaim polis knapa susah sekali . Sya hrus nunggu brpa lama spya uang bisa ditransfer hm apa saya juga termasuk korban seperti kebanyakan yg sya liat
2	Asuransi yang oke bngtttt pkony saya cinta prudential .yg lain sieh ??

1. Teks uji 1 :

Saya sdh 6 tahun ikut. Saat mengklaim polis knapa susah sekali . Sya hrus nunggu brpa lama spya uang bisa ditransfer hm apa saya juga termasuk korban seperti kebanyakan yg sya liat

a. Hasil *pre-processing* teks uji 1

Pada teks uji 1 dilakukan tahapan *pre processing* yang meliputi *case folding,* pendekatan *rule based* dengan tahap parser dan leksikal, tahap *convert negation* dan *filtering* sehingga didapatkan hasil seperti berikut.

Tabel 3. Hasil Pre-processing Teks 1

Teks	Hasil Pre-processing	
Saya sdh 6 tahun ikut. Saat mengklaim polis knapa susah sekali . Sya hrus nunggu brpa lama spya uang bisa ditransfer hm apa saya juga termasuk korban seperti kebanyakan yg sya liat	mengklaim polis susah nunggu lama	uang ditransfer korban

b. Perhitungan nilai probabilitas kata pada teks 1

Tabel 4. Frekuensi Kemunculan Kata Pada Teks 1

No	Kata	Frekuensi
1	**mengklaim**	1
2	**polis**	1
3	**susah**	1
4	**nunggu**	1
5	**lama**	1
6	uang	1
7	ditransfer	1
8	korban	1

Perhitungan nilai probabilitas setiap kata adalah sebagai berikut :

ISBN 978-602-99334-5-1

a. Kategori positif

P(mengklaim|positif) = 0.050
P(polis|positif) = 0.050
P(susah|positif) = 0.050
P(nunggu|positif) = 0.050
P(lama|positif) = 0.050
P(|positif)
=0.050*0.050*0.050*0.050*0.050
= 0.0000003125
Nilai probabilitasnya
P(positif)*P(|positif)
= 0.5*0.0000003125 **=<u>0.00000015625</u>**

b. Kategori negatif

P(mengklaim|negatif) = 0,090
P(polis|negatif) = 0,090
P(susah|negatif) = 0,090
P(nunggu|negatif) = 0,090
P(lama|negatif) = 0,090
P(|negatif)
= 0,090*0,090*0,090*0,090*0,090
= 0.0000059049
Nilai probabilitasnya
P(negatif)*P(|negatif)= 0.5*0.0000059049
= <u>0.00000295245</u>

Dari perhitungan tersebut dapat disimpulkan kategori dari teks uji 1 termasuk **kategori sentimen negatif** karena memiliki nilai probabilitas paling tinggi yaitu **0.00000295245.**

2. Teks uji 2 :
   Asuransi yang oke bngtttt pkony saya cinta prudential .yg lain sieh ??

a. Hasil *pre-processing* teks uji 2

Pada teks uji 2 dilakukan tahapan *pre processing* yang meliputi *case folding*, pendekatan *rule base* dengan tahap parser dan leksikal, tahap *convert negation* dan *filtering* sehingga didapatkan hasil seperti berikut.

Tabel 5. Hasil Pre-processing Teks 2

Teks	Hasil Pre-processing
Asuransi yang oke bngtttt pkony saya cinta prudential .yg lain sieh ??	asuransi oke cinta prudential

b. Perhitungan nilai probabilitas kata pada teks 2

Tabel 6. Frekuensi Kemunculan Kata Pada Teks 2

No	Kata	Frekuensi
1	**asuransi**	1
2	oke	1
3	**cinta**	1
4	**prudential**	1

Perhitungan nilai probabilitas setiap kata adalah sebagai berikut :

a. **Kategori positif**

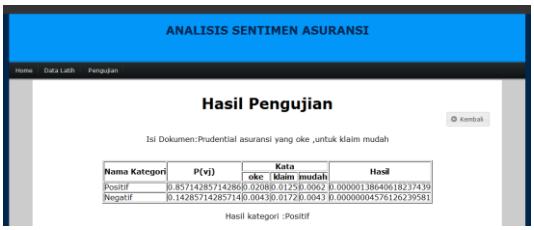
P(asuransi|positif) = 0.100
P(cinta|positif) = 0.100
P(prudential|positif) = 0.100
P(|positif)
= 0.100*0.100*0.100
= 0.001
Nilai probabilitasnya
P(positif)*P(|positif)
= 0.5*0.001 = **<u>0.005</u>**

b. **Kategori negatif**

P(asuransi|negatif) = 0.045
P(cinta|negatif) = 0.045
P(prudential|negatif) = 0.045
P(|negatif)
= 0.045*0.045*0.045
= 0.000091125
Nilai probabilitasnya
P(negatif)*P(|negatif)
= 0.5*0.000091125**= <u>0.0000455625</u>**

Dari perhitungan tersebut dapat disimpulkan kategori dari teks 2 termasuk **kategori sentimen positif** karena memiliki nilai probabilitas paling tinggi yaitu **0.005.**

Hasil program dari proses pengujian yang telah dilakukan dapat dilihat pada Gambar 2.



Gambar 2. Program sentimen hasil pengujian

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan yang berjudul Text Mining Dalam Analisis Sentimen Asuransi Menggunakan Metode *Naïve Bayes Classifier* dapat disimpulkan bahwa penelitian ini telah menghasilkan sebuah sistem analisis sentimen asuransi menggunakan. Sistem yang dibangun telah mampu mentransformasi sentimen yang berupa teks ulasan pada fanpage prudential Indonesia sehingga dapat menampilkan informasi sentimen masyarakat terhadap asuransi yang bersifat positif dan negatif. Penggunaan metode *naive bayes classifier* pada penelitian ini tidak menjamin ketepatan dalam proses klasifikasi. Akurasi pada sistem yang dibangun mencapai 95%.

Berdasarkan hasil pengujian yang telah dilakukan dapat disimpulkan bahwa proses pre-processing kata pada dokumen dan semakin banyak data latih pada masing-masing kategori sentimen dapat berpengaruh pada akurasi yang diperoleh akan semakin baik, proses pengujian dengan pra-proses dan metode yang digunakan menjadi kurang akurat karena kurangnya pengetahuan pada data latih dan filtrasi kata yang belum maksimal.

Saran yang dikemukakan diharapkan dapat bermanfaat sebagai acuan bagi penelitian selanjutnya yang akan mengembangkan penelitian ini, diantaranya:

1. Selain dengan tahap pra-proses yang telah digunakan pada penelitian ini dapat ditambahkan kombinasi tahap pra-proses lain sehingga dapat menghasilkan filtrasi kata yang lebih baik dan hasil yang di peroleh untuk nilai akurasi dapat lebih meningkat.
2. Melakukan kombinasi metode naive bayes dengan metode lain untuk mendapatkan hasil klasifikasi lebih baik lagi.
3. Menggunakan fitur API *facebook* untuk memperoleh data dari berbagai fanpage di dalam *facebook* sehingga dapat diperoleh data untuk berbagai bisnis.

DAFTAR PUSTAKA

Aditia Rakhmat, A. (2014). Analisis sentimen terhadap acara televisi berdasarkan opini publik. *Jurnal Ilmiah Komputer dan Informatika(KOMPUTA)* , 1-6.

Elly Indrayuni, M. (2015). Penerapan character N-gram untuk sentiment analysis review hotel menggunakan algoritma naive bayes. *Konferensi Nasional Ilmu Pengetahuan dan Teknologi(KNIT)* , 88-93.

Ema Utami, S. (2007). Pendekatan Metode Rule Based Untuk Mengalihbahasakan Teks Bahasa Inggris Ke Teks Bahasa Indonesia. *Jurnal Informatika , 8*, 42-53.

Hamzah, A. (2014). Sentimen analysis untuk memanfaatkan saran kuesioner dalam evaluasi pembelajaran dengan menggunakan naive bayes classifier(NBC). *Prosiding Seminar Nasional Aplikasi Sains & Teknologi(SNAST)* , 17-24.

Noviah Dwi Putranti, E. (2014). Analisis sentimen twitter untuk teks berbahasa Indonesia dengan maximum entropy dan support vector machine. *IJCCS , 8*, 91-100.

Nurirwan Saputra, T. (2015). Analisis sentimen data presiden Jokowi dengan pre-processing normalisasi dan stemming menggunakan metode naive bayes dan SVM. *Jurnal Dinamika Informatika , 5*.

ISBN 978-602-99334-5-1