

Sentimen Analisis Berinternet Pada Media Sosial dengan Menggunakan Algoritma Bayes

Ahmad Fauzi¹, Muhammad Faittullah Akbar², Yudhi Ferdi Andri Asmawan³

¹Universitas Bina Sarana Informatika
e-mail: Ahmad.fzx@bsi.ac.id

²Universitas Bina Sarana Informatika
e-mail: Muhammad.mtl@bsi.ac.id

³Universitas Bina Sarana Informatika
e-mail: Yudhi.yed@bsi.ac.id

Abstrak

Dalam penelitian analisis sentimen ini proses klasifikasi dokumen di bagi dalam dua kelas, yaitu kelas sentimen positif dan negatif. Media sosial telah menyediakan tempat bagi pengguna web untuk mengungkapkan berbagai pemikiran, pendapat dan menyampaikan berita tentang topik yang berbeda dalam sebuah acara. Data diperoleh dari jejaring sosial Instagram berdasarkan komentar yang ada. Penelitian ini bertujuan untuk menentukan sentimen publik terhadap universitas yang disampaikan di Instagram, sehingga membantu usaha untuk melakukan riset marketing atas opini publik. Klasifikasi algoritma seperti Naive Bayes (NB) yang diusulkan oleh banyak peneliti untuk digunakan dalam analisis sentimen teks. Algoritma Naive Bayes dan metodenya, akan diuji dengan dua masukan dengan menggunakan komentar tokenize and Transform Case yang positif (100 teks komentar) dan negatif (100 komentar teks), akurasi yang di dapat algoritma Naive Bayes 76.50% +/- 7.76% (mikro: 76.50). Hasil penelitian menunjukkan bahwa Naive Bayes (NB) mendapatkan hasil terbaik dan akurat.

Kata kunci: sentimen analisis, Naive Bayes, sosial media

Abstract

In this sentiment analysis study the document classification process was divided into two classes, namely positive and negative sentiment classes. Social media has provided a place for web users to express their thoughts, opinions and deliver news about different topics in an event. Data obtained from Instagram social networks based on existing comments. This research aims to determine the public sentiment towards the university that was delivered on Instagram, thus helping businesses to conduct marketing research on public opinion. Classification algorithms such as Naive Bayes (NB) were proposed by many researchers to be used in text sentiment analysis. The Naive Bayes algorithm and its methods, will be tested with two inputs using positive tokenize and Transform Case comments (100 comment texts) and negative (100 text comments), the accuracy obtained by the Naive Bayes algorithm is 76.50% +/- 7.76% (micro: 76.50). The results showed that Naive Bayes (NB) got the best and accurate results.

Keywords: sentiment analysis, Naive Bayes, social media

1. Pendahuluan

Jumlah pengguna media sosial kini berkembang pesat seiring kemajuan teknologi informasi yang bergerak cepat. Pengguna internet sekarang bebas berbagi informasi hingga membentuk sebuah opini publik yang memiliki berbagai makna. Dari Opini tersebut diyakini sebagai data yang akurat sebab opini publik terbentuk dari suatu luapan emosi, pendapat yang muncul, atau cerita yang sedang viral di media

sosial. Seperti dalam Twitter, Instagram, Facebook yang kerap menjadi kanal bagi warganet untuk 'curhat'.

Selain para netizen, partai politik maupun perusahaan dapat menggunakan data tersebut sebagai sumber informasi penting mengenai tanggapan dari masyarakat, mengukur tingkat popularitas dan layanan. Namun mengukur suatu sentimen komentar di suatu media sosial bukan hal yang mudah. Kendala yang

dihadapi dalam media sosial yakni menggunakan bahasa formal sehingga dibutuhkan analisis agar opini yang berada di media sosial tersebut dapat diartikan dan digunakan dalam pengambilan keputusan.

Media Sosial yang saat ini ada terdiri dari berbagai media online seperti situs media jejaring sosial, aplikasi media jejaring sosial, game, dan media online lainnya. Macam-macam media sosial yang disediakan: *Path, Line, Instagram, Whatsapp, Blackberry Messenger, Twitter, LinkedIn, Facebook, Wikipedia, Google Plus, youtube* dan masih banyak lagi lainnya. Pada umumnya berita yang disampaikan dalam portal tersebut terdiri dari beberapa kategori seperti berita politik, layanan sosial, ekonomi, olahraga, kesehatan, dan lain-lain. Tentu saja informasi yang akan didapat bukan saja pemberitaan yang bersifat positif saja tetapi ada juga yang bersifat negatif (Aji, Budilaksono, n.d.).

Dalam Surat Edaran (SE) Kapolri soal penanganan ujaran kebencian Nomor SE/06/X/2015 ("Surat Edaran Kapolri Mengenai Penanganan Ujaran Kebencian.Pdf," 2015). Dalam surat edaran tersebut, disebutkan bahwa persoalan ujaran kebencian semakin mendapatkan perhatian masyarakat baik nasional atau internasional seiring meningkatnya kepedulian terhadap perlindungan hak asasi manusia (HAM). Berikut poin-poin krusial dalam SE tersebut:

Bentuk Ujaran Kebencian disebutkan bahwa "ujaran kebencian dapat berupa tindak pidana yang diatur dalam Kitab Undang-Undang Hukum Pidana (KUHP) dan ketentuan pidana lainnya di luar KUHP, yang berbentuk antara lain: (1) Penghinaan, (2) Pencemaran nama baik, (3) Penistaan, (4) Perbuatan tidak menyenangkan, (5) Memprovokasi, (6) Menghasut, (7) Menyebarkan berita bohong dan semua tindakan di atas memiliki tujuan atau bisa berdampak pada tindak diskriminasi, kekerasan, penghilangan nyawa, dan atau konflik sosial" ("SURAT EDARAN KAPOLRI MENGENAI PENANGANAN UJARAN KEBENCIAN.pdf," 2015).

Aspek Ujaran Kebencian Selanjutnya, bahwa ujaran kebencian sebagaimana dimaksud di atas bertujuan untuk menghasut dan menyulut kebencian terhadap individu dan atau kelompok

masyarakat dalam berbagai komunitas yang dibedakan dari aspek: (1) Suku, (2) Agama, (3) Aliran keagamaan, (4) Keyakinan atau kepercayaan, (5) Ras, (6) Antargolongan, (7) Warna kulit, (8) Etnis, (9) Gender, (10) Kaum difabel, (11) Orientasi seksual ("Surat Edaran Kapolri Mengenai Penanganan Ujaran Kebencian.Pdf," 2015).

Media Ujaran Kebencian disebutkan bahwa "ujaran kebencian sebagaimana dimaksud di atas dapat dilakukan melalui berbagai media, antara lain: (1) Dalam orasi kegiatan kampanye, (2) Spanduk atau banner, (3) Jejaring media sosial, (4) Penyampaian pendapat di muka umum (demonstrasi), (5) Ceramah keagamaan, (6) Media massa cetak atau elektronik, (7) Pamflet. dengan memperhatikan pengertian ujaran kebencian di atas, perbuatan ujaran kebencian apabila tidak ditangani dengan secara efektif, efisien, dan sesuai dengan ketentuan peraturan perundang-undangan, akan berpotensi menimbulkan tindak kriminal, kekerasan, dan memunculkan konflik sosial yang meluas ("Surat Edaran Kapolri Mengenai Penanganan Ujaran Kebencian.Pdf," 2015).

Penelitian Terkait

Banyak ide-ide telah muncul selama bertahun-tahun tentang bagaimana mencapai hasil yang berkualitas dari sistem klasifikasi Web, sehingga ada pendekatan yang berbedayang dapat digunakan untuk permasalahan analisis sentimen, seperti *Naïve Bayes* (NB) dan *Bayesian Networks*, *NNs*, *DTs*, *Support Vector Machines* (SVM) dll. *Mode I* (*Naïve Bayes*) NB yang populer dalam aplikasimachine learning, karena kesederhanaan mereka dalam membiarkan setiap atribut berkontribusi terhadap keputusan akhir yang sama dan independen dari atribut lainnya (Xhemali, Hinde and Stone, 2009). Berbagai metode *machine learning* telah digunakan untuk penentuan atribut dengan menggunakan *naïve Bayes* (Saleh, El, & Menai, 2014).

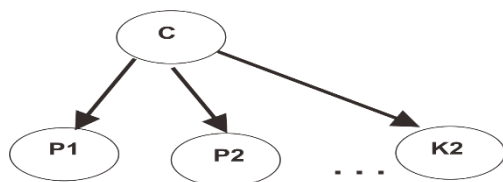
Proses klasifikasi teks dapat dibagi menjadi dua fase *klasifikasi* (Darujati et al., 2012) yaitu:

1. Fase *information retrieval* (IR) untuk mendapatkan data numerik dari sebuah dokumen teks. Langkah pertama yang dilakukan pada fase ini adalah *feature extraction*. Merupakan pendekatan yang umum

digunakan adalah distribusi frekuensi kata. Nilai angka yang diperoleh dapat berupa berapa kali suatu kata muncul di dalam sebuah dokumen, 1 (satu) jika kata ada didalam dokumen atau 0 (nol) jika tidak ada (biner), atau jumlah kemunculan kata pada awal dokumen. *Feature* yang diperoleh dapat direduksi agar dimensi vektor menjadi lebih kecil. Beberapa pendekatan *feature reduction* dapat diterapkan seperti menghapus *stop-words*, *stemming*, *statistical filtering*. (Feldman, Goldenberg, & Netzer, 2010)

2. Fase klasifikasi utama yaitu ketika suatu algoritma yang memproses sebuah data untuk memutuskan kategori dimana teks itu ditempatkan. Algoritma klasifikasi yang merupakan salah satu kajian di bidang statistika dan machine learning yang dapat diterapkan pada fase ini, di antaranya *Naive Bayesian*, *Decision Tree*, *Neural Network*, *kNearest Neighbor* dan *Support Vector Machines*. (Darujati et al., 2012).

Klasifikasi–klasifikasi *Bayes* merupakan klasifikasi statistik yang dapat memprediksi atribut kelas suatu anggota probabilitas. Untuk sebuah klasifikasi *Bayes* sederhana yang lebih dikenal sebagai *Naive Bayesian Classifier* dapat diasumsikan bahwa efek dari suatu nilai atribut sebuah kelas yang diberikan adalah bebas dari atribut-atribut lain. Dalam *Naive Bayes* diasumsikan prediksi atribut adalah tidak tergantung pada kelas atau tidak dipengaruhi atribut lain (Darujati et al., 2012), dengan gambaran sebagai berikut:



Gambar 1. Klasifikasi Naive Bayes sebagai jaringan bayes dengan atribut prediksi (P1, P2,.....Pk) dan kelas (C). (Darujati et al., 2012)

C adalah anggota kelas dan X adalah variabel acak sebuah vektor sebagai

atribut nilai yang diamati. c mewakili nilai label kelas dan x mewakili nilai atribut vector yang diamati. Jika diberikan sejumlah x tes untuk klasifikasi maka probabilitas tiap kelas untuk atribut prediksi vektor yang diamati adalah:

$$p(C = c|X = x) = \frac{p(C = c)p(X = x|C = c)}{p(X = x)}$$

X = x adalah mewakili kejadian dari Jumlah dari untuk semua kelas adalah 1

$$X_1 = x_1 \wedge X_2 = x_2 \wedge \dots \wedge X_k = x_k$$

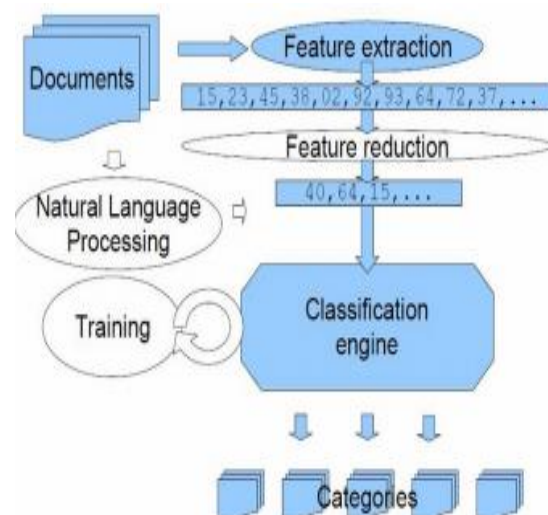
$$p(C = c|X = x)$$

Pada tahap evaluasi sistem, perhitungan menggunakan tabel kontingensi yang diberikan (Ling, Kencana and Oka, 2014):

#	Relevant	Not Relevant
Retrieved	True Positive (TP)	False Positive (FP)
Not Retrieved	False Negatif (FN)	True Negatif (TN)

Tabel 1. Tabel Kontingensi Evaluasi Kinerja Sistem (Ling, Kencana and Oka, 2014)

Proses klasifikasi teks melibatkan banyak teknik IR mulai dari preprocessing, pengukuran kemiripan selama operasional klasifikasi sampai dengan evaluasi kinerja algoritma.



Gambar 2. Diagram proses klasifikasi teks (Darujati et al., 2012) (Darujati and Gumelar, 2012)

Pre-Processing pada pengujiannya Analisis Sentimen Menggunakan Metode *Naïve Bayes Classifier* Dengan Seleksi Fitur Chi Square (Ling & Oka, 2014) sebagai berikut:

Tokenization adalah tugas pemotongan urutan karakter dan sebuah set dokumen yang diberikan menjadi potongan-potongan kata atau karakter yang sesuai dengan kebutuhan sistem. Potongan-potongan tersebut dikenal dengan istilah token

Stemming merupakan salah satu proses dari mengubah token yang berimbuhan menjadi kata dasar, dengan menghilangkan semua imbuhan yang ada pada token tersebut. Pentingnya stemming dalam proses pembuatan sistem adalah untuk menghilangkan imbuhan pada awalan dan akhiran. Berdasarkan hasil proses tersebut, akan didapatkan sebuah informasi mengenai banyaknya fitur yang muncul dalam sebuah dokumen (Kontopoulos, Berberidis, Dergiades, & Bassiliades, 2013).

Stopwords dapat diartikan sebagai menghilangkan karakter, tanda baca, serta kata-kata umum yang tidak memiliki makna atau informasi yang dibutuhkan. *Stopwords* umumnya digunakan dalam Penelitian tentang klasifikasi sentimen terhadap *review* film telah dilakukan oleh Dhande dan Patnaik (2014) dengan menggunakan algoritma *Naive Bayes*, *Neural Network*, dan *Naive Bayes Neural Classifier*. Dari hasil penelitian akhir yang diuji menggunakan ketiga algoritma tersebut menyebutkan bahwa *Naive Bayes* menghasilkan akurasi yang lebih tinggi dibandingkan *Neural Network*. Dan algoritma *Naive Bayes Neural Classifier* yang merupakan penggabungan antara metode *Naive Bayes* dan *Neural Network* menghasilkan akurasi yang paling tinggi diantara kedua algoritma tersebut. *Naive Bayes* memiliki beberapa keunggulan seperti sederhana, cepat dan akurasi yang tinggi (Dehkharghani, Mercan, Javeed, & Saygin, 2014).

Banyak peneliti telah melakukan klasifikasi sentimen dengan menggunakan *Naive Bayes*. Namun klasifikasi ini memiliki keterbatasan utama yang tidak mungkin selalu memenuhi asumsi independensi antara atribut. Dan ini mempengaruhi tingkat akurasi klasifikasi.

K-fold Cross-validation merupakan teknik validasi dengan membagi data awal secara

acak kedalam k bagian yang saling terpisah atau "*fold*" (Ling & Oka, 2014). Grafik *Receiver Operating Characteristics* (ROC) adalah teknik untuk memvisualisasikan, mengorganisasikan dan memilih pengklasifikasi berdasarkan kinerja mereka [2]. Kurva ROC digunakan untuk mengukur nilai *Area Under Curve* (AUC). Kurva ROC memiliki properti yang menarik: mereka tidak sensitif terhadap perubahan distribusi kelas. Jika proporsi positif terhadap kasus negatif berubah dalam satu set tes, kurva ROC tidak akan berubah. Untuk melihat mengapa demikian, dapat dilihat pada *confusion matrix* (Indrayuni et al., 2015)

Penelitian lain yang pernah dilakukan (Kang, Yoo, & Han, 2012) adalah analisa sentimen pada *review* restoran menggunakan algoritma *Naive Bayes* dengan fitur unigrams dan bigrams untuk meningkatkan akurasi *Naive Bayes*. Pada penelitian ini dengan menerapkan metode senti leksikon yaitu fitur unigrams dan bigrams, menunjukkan bahwa selisih akurasi antara *class* positif dan negatif sekitar 3,6% dibandingkan dengan penggunaan *Naive Bayes* saja.

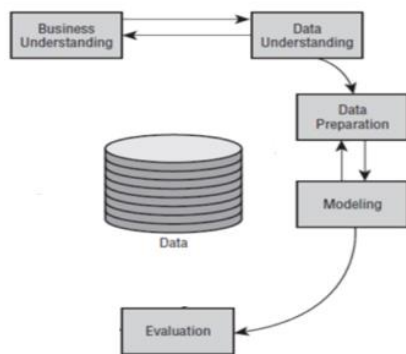
Pada penelitian ini menggunakan algoritma *Naive Bayes* untuk mendapatkan nilai akurasi tertinggi dengan model tersebut:

1. Banyaknya pemberitaan dimedia sosial yang bersifat negatif sehingga perlunya pembatasan atau penyaringan pada pemberitaan
2. Bagaimana tingkat akurasi analisa
3. sentimen yang dihasilkan algoritma *Naive Bayes*

Dan adapun manfaat dari penelitian ini adalah:

1. Manfaat dari penelitian ini adalah membantu para pembaca atau penerima informasi yang lebih sehat dan bersifat positif
2. Manfaat teoritis penelitian ini adalah memberikan bukti secara empiris untuk teori yang berkaitan sehingga dapat dijadikan sumbangan pemikiran untuk pengembangan teori berikutnya.

2. Metode Penelitian

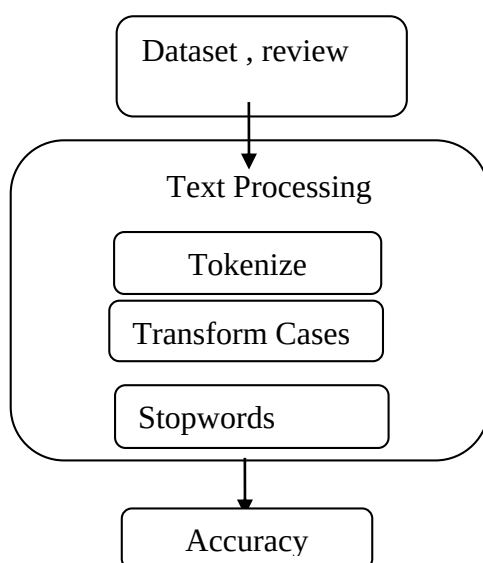


Gambar 4. Model modifikasi CRISP-DM pada penelitian

Metode yang dilakukan pada penelitian ini hanya menggunakan tahap *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* dan *Deployment*.

Penelitian yang dilakukan pada analisa sentimen dengan menggunakan Algoritma *Naive Bayes*(NB) pada pemberitaan “Surat Edaran (SE) Kapolri soal penanganan ujaran kebencian atau hate speech Nomor SE/06/X/2015” dengan mengukur akurasi data eksperimen yang dilakukan dalam penelitian dengan menggunakan aplikasi *RapidMiner Studio* 6.5.002.

Pada gambar 2 menggambarkan kerangka pemikiran yang diusulkan adalah sebagai berikut :



Gambar 3. Kerangka Pemikiran

Untuk melakukan eksperimen penelitian ini menggunakan data komentar review: <https://www.instagram.com/amikbsit.egal/>. Maka didapat 200 data *review* yang terdiri dari 100 *review* positif dan 100 *review* negatif. Data *review* positif disatukan dalam folder dengan nama *Positif*. Sedangkan data *review* negatif disatukan penyimpanannya dalam folder dengan nama *Negatif*. Tiap dokumen berekstensi .txt yang dapat dibuka menggunakan aplikasi *Notepad*.

Kemudian setelah mendapatkan data positif dan negatif dalam bentuk .txt dengan jumlah data yang sama, yaitu 100 *review* data positif maka yang bersifat negatif harus didapatkan 100 *review* data negatif, data akan diproses menggunakan Aplikasi *Rapid Miner*.

Pada aplikasi *Rapid Miner* untuk *text processing* yang peneliti menggunakan *Tokenize*, *TransformCases*, *Stopwords*(*Dictionary*). Pengujian yang dilakukan dengan memilih *feature selection Naive Bayes*(NB).

3. Hasil dan Pembahasan

Penelitian yang dilakukan menggunakan Aplikasi *RapidMiner Studio* 6.5.002 dengan menggunakan spesifikasi komputer CPU *Intel Core i3* 1.90GHz, RAM 4GB, dan *system operasi Microsoft Windows 7 Professional* 64-bit. Hasil pengujian *review* Pemberitaan “Awat, Penebar Kebencian di Media Sosial Bisa Diancam Pidana” dengan menggunakan teks bahasa Indonesia.

Hasil dari penelitian menggunakan *Text Processing* pada table 1 yang dilalui terdiri dari *Tokenize*, *Transform Cases*, *Stopwords*(*Dictionary*). Hasil pengolahan sebagai berikut:

<i>Text Review</i>	<i>Tokenize</i>	<i>Transform Cases</i>	<i>Stopwords(Dictionar)</i>
YG BEGINI DAPET SOSIAL MEDIA AWARD ? AMPUN MALU MALUIN	YG BEGINI DAPET SOSIAL MEDIA AWARD AMPUN MALU MALUIN	begini dapet Sosial media award ampun malu maluin	begini dapet sosial media award
wanjir ada aku wkwk, MIN katanya pihak penyelen ggara wajib pelajar semua min?	wanjir ada aku wkwk MIN katanya pihak penyelen ggara wajib pelajar semua min	wanjir ada aku wkwk, katanya pihak penyele nggara wajib pelajar semua	ada aku wkwk, katanya pihak penyelen ggara wajib pelajar semua
BSI keren coy!! Nambah banyak teman baru, salam dari tim kuning	BSI keren coy Nambah banyak teman baru salam dari tim kuning	bsi keren nambah banyak teman baru salam dari tim kuning	bsi keren nambah teman baru salam tim kuning

4. Kesimpulan

Pengujian model dengan menggunakan *Naive Bayes* (NB) dengan menggunakan data *review* analisis sentimen Etika berinternet dimedia sosial dari 200 data *review* dalam teks bahasa indonesia yang terdiri dari 100 *review* positif dan 100 *review* negatif.

Model yang dihasilkan mendapatkan hasil *accuracy* pengujian pertama dilakukan dengan menggunakan *Naive Bayes*(NB) yaitu mendapatkan nilai *accuracy* 76.50%+/-7.76(mikro: 76.50%) maka dapat disimpulkan pengujian analisa sentiment dengan menggunakan *Naive Bayes*(NB) sudah memenuhi syarat batas minimum akurasi sebuah penelitian. sehingga kita dapat memecahkan masalah analisa sentimen dengan memilih pemberitaan atau

informasi dimedia sosial khususnya lebih akurat.

Referensi

- Aji ,budilaksono, 2015. (n.d.). Aji dan Budilaksono-Analisis Sentimen Etika Berinternet Pada Media Sosial Dengan Menggunakan Algoritma Naive Bayes Dan Particle Swarm Optimization-2015.
- Darujati, C., Gumelar, A. B., Informasi, S., Komputer, F. I., Surabaya, U. N., Bayes, N., ... Learning, S. (2012). PEMANFAATAN TEKNIK SUPERVISED UNTUK, 16(1), 1–8.
- Dehkharghani, R., Mercan, H., Javeed, A., & Saygin, Y. (2014). Expert Systems with Applications Sentimental causal rule discovery from Twitter. *EXPERT SYSTEMS WITH APPLICATIONS*, 41(10), 4950–4958. <https://doi.org/10.1016/j.eswa.2014.02.024>
- Feldman, R., Goldenberg, J., & Netzer, O. (2010). Mine your own business: Market structure surveillance through text mining. *Marketing Science Institute, Special Report*, 10(3), 10–202. <https://doi.org/10.2139/ssrn.1816494>
- Indrayuni, E., Wahyudi, M., Informasi, S., Selatan, J., Komputer, I., & Selatan, J. (2015). Penerapan Character N-Gram Untuk Sentiment Analysis Review Hotel Menggunakan Algoritma Naive Bayes, 88–93.
- Kang, H., Yoo, S. J., & Han, D. (2012). Expert Systems with Applications Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems With Applications*, 39(5), 6000–6010. <https://doi.org/10.1016/j.eswa.2011.11.107>
- Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Expert Systems with Applications Ontology-based sentiment analysis of twitter posts. *Expert Systems With Applications*, 40(10), 4065–4074. <https://doi.org/10.1016/j.eswa.2013.01.001>

- Ling, J., & Oka, T. B. (2014). Analisis Sentimen Menggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Chi Square, 3(3), 92–99.
- Saleh, A., El, M., & Menai, B. (2014). Naïve Bayes classifiers for authorship attribution of Arabic texts. *Journal of King Saud University - Computer and Information Sciences*, 26(4), 473–484. <https://doi.org/10.1016/j.jksuci.2014.06.006>
- Surat Edaran Kapolri Mengenai Penanganan Ujaran Kebencian.Pdf. (2015).