

Analisis Sentimen terhadap Toko Online menggunakan Naïve Bayes pada Media Sosial Twitter

Kautsar Ramadhan S¹, Kemas Muslim L²

^{1,2}Fakultas Informatika, Universitas Telkom, Bandung

¹kautsarramadhan@students.telkomuniversity.ac.id, ²kemasmuslim@telkomuniversity.ac.id

Abstrak

Twitter merupakan salah satu Media Sosial yang digunakan untuk menyalurkan opini dan membahas berbagai masalah yang ada disekitar. Salah satu masalah yang banyak dibahas adalah Toko Online yang merupakan salah satu sistem terbaru dalam berbelanja. Toko Online banyak digunakan oleh konsumen karena mudah dan cepat dalam proses transaksinya. Pada tahun 2017, Tokopedia dan Lazada menjadi salah satu situs Toko Online yang sering dikunjungi dan banyak melakukan transaksi. Analisis sentimen merupakan suatu cara untuk membandingkan pendapat banyak orang. Penelitian ini bertujuan untuk menentukan sentimen yang terjadi pada setiap komentar mengenai Tokopedia dan Lazada pada media sosial Twitter. Sentimen yang digunakan akan dibagi menjadi 2 kelas, yaitu positif dan negatif serta menggunakan Naïve Bayes Classifier sebagai algoritmanya. Hasil penelitian diperoleh akurasi 75.26% untuk Lazada dan 83.97% untuk Tokopedia.

Kata kunci : *Analisis Sentimen, Naïve Bayes Classifier, Tokopedia, Lazada*

Abstract

Twitter is one of the Social Media that is used to channel opinions and discuss various problems that are around. One problem that is widely discussed is the Online Store which is one of the newest systems in shopping. Online stores are widely used by consumers because they are easy and fast in the transaction process. In 2017, Tokopedia and Lazada became one of the Online Shop sites that were frequently visited and made many transactions. Sentiment analysis is a way to compare the opinions of many people. This study aims to determine the sentiments that occur in every comment about Tokopedia and Lazada on Twitter social media. The sentiment used will be divided into 2 classes, namely positive and negative and using the Naïve Bayes Classifier as the algorithm. The results of the study obtained 75.26% accuracy for Lazada and 83.97% for Tokopedia.

Keywords: *Sentiment Analysis, Naïve Bayes Classifier, Tokopedia, Lazada*

1. Pendahuluan

1.1 Latar Belakang

Di zaman yang sudah modern ini, penggunaan internet sudah menjadi hal yang wajar bagi masyarakat dunia untuk melakukan apapun yang diinginkan. Dalam perkembangannya, internet bukan hanya digunakan sebagai media informasi dan media komunikasi saja. Namun, internet juga dapat menambah pendapatan seseorang. Apalagi di zaman seperti sekarang ini, manusia menyukai semua hal yang berbau praktis dan otomatis untuk menjalankan kelangsungan hidupnya terutama dalam hal menjalankan transaksi jual-beli [12].

Belakangan ini, Toko Online kerap mewarnai perdagangan dunia maya. Toko Online merupakan suatu perdagangan elektronik dimana konsumen langsung membeli barang dari penjual melalui sebuah situs website pada internet yang dimana transaksinya dilakukan tanpa layanan perantara. Sebenarnya cara belanja seperti ini sudah ada sejak 1979, diciptakan oleh Michael Aldrich dari Inggris. Baru pada Maret 1981 sistem ini mulai diperkenalkan kepada masyarakat oleh Thomson Holidays. Saat ini, di Indonesia jumlah pengunjung shopping online bisa mencapai 73% perharinya. Dengan masing-masing konsumen 65% wanita dan 35% kaum pria [12].

Dalam keterangan resminya kepada [Liputan6.com](http://liputan6.com), Rabu (23/8/2017), ilmuOne Data mengungkapkan beberapa Toko Online terbaik di Indonesia yang terdiri dari e-Commerce dan marketplace. IlmuOne Data mendefinisikan *marketplace* sebagai fasilitator pembelanjaan online yang tidak memiliki inventarisasi barang sendiri. Beberapa temuan kunci dari studi ini menunjukkan *Lazada* memimpin seluruh e-Commerce dengan 21,2 juta pengunjung unik, sementara *Tokopedia* memimpin marketplace dengan angka 14,4 juta [15].

Adapun dampak lain dari penggunaan Internet yaitu penggunaan media sosial yang terus meningkat. Menurut penelitian yang dilakukan perusahaan asal Inggris, We Are Social, dari total populasi Indonesia sebanyak 265,4 juta jiwa, pengguna aktif media sosialnya mencapai 130 juta dengan penetrasi 49% [17]. Salah satu media terpopuler pada penelitian tersebut yaitu Twitter. Berdasarkan data perusahaan tersebut, pengguna media sosial twitter memiliki 27% pengguna aktif [17]. Banyak pengguna yang senang memainkan media sosial dan menjadikan media sosial sarana untuk menuangkan opininya dan membahas berbagai masalah yang ada di sekitar. Opini dan masalah tersebut dapat diolah menjadi suatu informasi dengan suatu cara yaitu analisis sentimen.

Analisis sentimen merupakan suatu aktivitas menganalisa suatu perasaan, emosi dan opini seseorang yang di ekspresikan dalam teks. Tugas dasar dalam analisis sentimen adalah mengelompokkan sentimen yang terbagi menjadi sentimen positif dan negatif [13]. Pada penelitian ini dibuat sistem dengan tujuan untuk menghasilkan informasi sentimen mengenai opini masyarakat terhadap Toko Online Lazada dan Tokopedia yang mengarah ke sentimen positif dan negatif menggunakan metode Naïve Bayes Classifier yang merupakan salah satu Teknik pembelajaran dari Text Mining.

1.2 Topik dan Batasannya

Topik yang diangkat untuk dijadikan bahan analisis yaitu tweet yang berisikan tentang *Lazada* dan *Tokopedia*. Kata kunci yang digunakan untuk mencari tweet tentang toko online adalah “LazadaID” dan “Tokopedia”.

Batasan yang dilakukan dalam pengerjaan Tugas Akhir ini yaitu pengambilan data twitter yang dilakukan menggunakan teknik *crawling*. Data yang diambil hanya data tweet saja, dan bahasa tweet yang dipakai menggunakan Bahasa Indonesia.

1.3 Tujuan

Tujuan dari penulisan Tugas Akhir ini yaitu membandingkan klasifikasi data tweet yang sudah di dapat menjadi sentimen positif dan negatif serta membandingkan seberapa besar tingkat akurasi yang didapat dari hasil prediksi klasifikasi antara Toko Online Lazada dan Tokopedia.

1. Studi Terkait

2.1 Crawling Tweet

Crawling Data adalah suatu metode untuk mengumpulkan atau mengambil data dari suatu blog, Sosial Media atau situs lainnya. Dalam hal ini Twitter menjadi target untuk pengambilan dataset. Crawling dilakukan dengan memasukan kata kunci “Tokopedia” dan “Lazada” yang kemudian data set tersebut disimpan. Kemudian pengklasifikasian atau pengkategorian dataset dilakukan dengan preprocessing data terlebih dahulu.

2.2 Preprocessing Tweet

Preprocessing adalah tahapan dimana data atau informasi diatur dari data tekstual semi terstruktur dan tidak terstruktur menjadi sebuah format yang lebih terstruktur yang nantinya akan mempermudah pemakaian. Tujuan dari preprocessing adalah melakukan *seleksi komentar* yang mengandung kata atau hashtag (#), karena Twitter terdapat fungsi retweet, yaitu memberikan komentar terhadap tweet komentar seseorang, karena komentar tweet akan mengganggu dalam proses analisis Sentimen tweet. Jadi dalam preprocessing ini komentar tweet di hapus [16].

Selain itu juga dilakukan *cleansing* untuk menghilangkan kalimat yang terdapat noise, yaitu kesalahan acak atau varian dalam variable terukur [14]. Kata yang dihilangkan adalah karakter HTML, kata kunci, ikon emosi, hashtag (#), username (@username), url (<http://website.com>), dan email (nama@website.com). Lalu ada *parsing* untuk memecah dokumen menjadi sebuah kata dengan melakukan analisa terhadap kumpulan kata dengan memisahkan kata tersebut dan menentukan struktur sintaksis dari tiap kata tersebut [18]. Kemudian ada *normalisasi kalimat* untuk menormalkan kalimat sehingga kalimat gaul menjadi normal [19], sehingga bahasa gaul tersebut dapat dikenali sebagai bahasa yang sesuai dengan KBBI. Serta menghilangkan huruf yang berulang dan menghilangkan emoticon.

2.3 Naïve Bayes Classifier

Metode *Naïve Bayes Classifier* (NBC), merupakan suatu metode klasifikasi sederhana berdasarkan aplikasi teorema Bayes dengan asumsi antar variable penjelas saling bebas (independen). Teorema bayes berfungsi untuk memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya. Metode pengklasifikasian tersebut dikemukakan oleh ilmuwan Inggris Thomas Bayes. Berikut ini rumus yang digunakan:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{1}$$

keterangan:

- X = peubah acak
- C = kelas
- $P(C|X)$ = probabilitas C didasarkan pada kondisi X
- $P(X|C)$ = probabilitas X di bawah kondisi C
- $P(C)$ = probabilitas C
- $P(X)$ = probabilitas X

Perhitungan yang dibutuhkan dalam NBC yaitu probabilitas tiap kelas, rumus mencari probabilitas tiap kelas dinotasikan sebagai berikut:

$$P(C) = \frac{N_c}{N} \tag{2}$$

Keterangan:

- $P(C)$ = probabilitas C
- N_c = jumlah total kelas C
- N = jumlah total keseluruhan kelas

2.4 Evaluasi Performansi

Evaluasi Performansi dilakukan untuk melihat performansi model klasifikasi dalam mencari hasil klasifikasi. Cara yang sering digunakan adalah menghitung akurasi, *precision*, *recall* dan *f-measure*. Untuk mempermudah perhitungan dibuatkan *Confusion Matrix* yang didapat dari hasil prediksi dan data aktual.

Tabel 1. Confusion Matriks

	Kelas Positif Prediksi	Kelas Negatif Prediksi
Kelas Positif Aktual	True Positif (TP)	False Negatif (FN)
Kelas Negatif Aktual	False Positif (FP)	True Negatif (TN)

akurasi merupakan persentase dari suatu prediksi model yang sudah dibuat.

$$akurasi = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{3}$$

precision merupakan tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem.

$$precision = \frac{TP}{(FP+TP)} \tag{4}$$

Recall merupakan kualitas seberapa lengkap hasil relevan yang ditampilkan oleh sistem prediksi kelas.

$$recall = \frac{TP}{(FN+TP)} \tag{5}$$

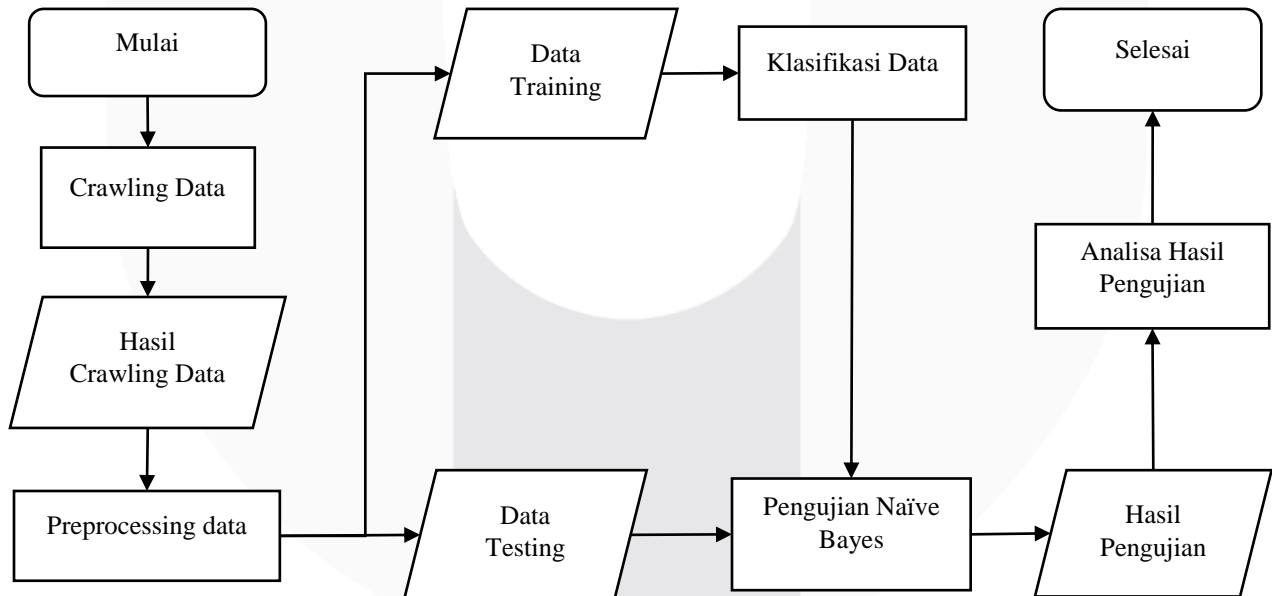
F-measure merupakan salah satu perhitungan evaluasi dalam informasi temu kembali yang mengkombinasikan recall dan precision.

$$f - measure = \frac{(2 \cdot recall \cdot precision)}{(recall+precision)} \tag{6}$$

dimana TP adalah hasil prediksi positif dan nilai aktual positif, FP adalah hasil prediksi positif dan nilai aktual negatif, FN adalah hasil prediksi negatif dan aktual positif, dan TN adalah hasil prediksi negatif dan nilai aktual negatif.

3 Sistem yang Dibangun

Pada tugas akhir ini akan diberikan gambaran langkah-langkah dalam mendeteksi kelas sentimen menggunakan metode Naïve Bayes. Berikut gambaran dari proses mendeteksi kelas sentimen dengan metode Naïve Bayes yang akan digunakan pada tugas akhir ini:



Gambar 1. Flowchart sistem yang akan digunakan

Berikut penjelasan pada Gambar 1:

1. Crawling data dari Twitter dengan kata kunci “Lazada” dan “Tokopedia”.
2. Setelah mendapatkan hasil crawling data, dilakukan pre-processing secara manual untuk membersihkan data tweet mentah menjadi data yang terstruktur.
3. Lalu menentukan data training dan data testing yang sudah di pre-processing secara manual.
4. Pada data training dilakukan klasifikasi data untuk menentukan sentimen.
5. Langkah selanjutnya dilakukan pengujian Naïve Bayes Classifier terhadap data training yang sudah diklasifikasi dan data testing.
6. Setelah mendapatkan hasil dari pengujian, hasil dari pengujian Naïve Bayes Classifier di Analisa.

3.1 Crawling Tweet

Crawling Data adalah suatu metode untuk mengumpulkan atau mengambil data dari suatu blog, Sosial Media atau situs lainnya. Dalam hal ini Twitter menjadi target untuk pengambilan dataset. Crawling dilakukan dengan memasukan kata kunci “LazadaID” dan “Tokopedia” yang kemudian data set tersebut disimpan. Tweet yang diambil menggunakan Bahasa Indonesia. Total data Lazada dan Tokopedia yang akan diambil sebanyak 1200 tweet untuk setiap kata kunci. Contoh tweet hasil crawling ditampilkan pada table 2 yang diambil secara acak dari data asli.

Tabel 2. Sampel Tweet

No.	Tweet
1	@LazadaID ikutan flash salea berasa dikerjain sama 我法慎鸵非法我法非垢析作我法慎鸵非法我法非垢析作 #lazada
2	adaketupat diantara laz lazadaid sahabat ikutan yuk ❤️ @LazadaID
3	thr buat beli donat ice cream coklat dan biskuit wafer ☺️ @tokopedia #tokopedia #thr
4	kog gak ad kiriman yaperasaan saya sudah isi data 我法慎鸵非狂我法非羴恻羴 apa salah mohon tanggapannya ☺️ @tokopedia

3.2 Preprocessing Tweet

Sebelum dilakukan pengolahan, data yang disimpan perlu di processing terlebih dahulu untuk “dibersihkan”. Pre-processing adalah tahapan dimana data atau informasi diatur dari data tekstual semi terstruktur dan tidak terstruktur menjadi sebuah format yang lebih terstruktur yang nantinya akan mempermudah pemakai. Kemudian data tersebut akan terbagi menjadi dua yaitu data training dan data testing.

Tabel 3. Preprocessing

No.	Tweet
1	ikutan flash sale berasa dikerjain sama
2	Ada ketupat diantara lazadaid sahabat ikutan yuk
3	thr buat beli donat ice cream coklat dan biskuit wafer
4	kok gak ada kiriman yaperasaan saya sudah isi data apa salah mohon tanggapannya

Pada table 3 merupakan hasil dari preprocessing dari table 2. Berikut ini merupakan tahapan yang dilakukan pada preprocessing tweet:

1. *Tokenizing* merupakan proses untuk memisahkan kalimat pada data menjadi beberapa kata.
2. *Cleansing* adalah proses pembersihan data teks yaitu dengan menghilangkan data yang tidak konsisten atau tidak relevan seperti emoticon, hashtag (#), username, url.
3. *Normalization* digunakan untuk menormalisasikan kalimat gaul menjadi kalimat yang sesuai dengan KBBI.
4. *Case folding* yaitu merubah bentuk kata yang memiliki imbuhan menjadi bentuk dasarnya.

3.3 Labelling

Pada proses labelling data tweet yang sudah dilakukan proses preprosessing diberikan label secara manual. Label ini digunakan sebagai klasifikasi kelas pada setiap tweet. Kelas yang digunakan yaitu kelas positif dan kelas negatif. Kelas positif diberi tanda "1" sedangkan kelas negatif diberi tanggal "0". Berikut ini contoh data yang diambil secara acak pada data Lazada dan Tokopedia sebelum dan sesudah diberikan label.

Tabel 4. Contoh Sampel Data sebelum pemberian label

No.	Tweet
1	ikutan flash sale berasa dikerjain sama
2	Ada ketupat diantara lazadaid sahabat ikutan yuk
3	thr buat beli donat ice cream coklat dan biskuit wafer
4	kok gak ada kiriman yaperasaan saya sudah isi data apa salah mohon tanggapannya

Tabel 5. Hasil pemberian Label pada Sampel Data

No.	Tweet	Kelas
1	ikutan flash sale berasa dikerjain sama	0
2	Ada ketupat diantara lazadaid sahabat ikutan yuk	1
3	thr buat beli donat ice cream coklat dan biskuit wafer	1
4	kok gak ada kiriman yaperasaan saya sudah isi data apa salah mohon tanggapannya	0

Setelah data sudah diberikan label secara manual seperti pada tabel 5, data yang sudah diberi label dibagi menjadi data training dan data testing. Untuk data training diambil 900 tweet dan data testing diambil 300 tweet secara manual.

4 Evaluasi

Dari pengujian yang sudah dilakukan dalam mencari probabilitas, confusion matrix, akurasi, precision, recall, dan f-measure didapatkan nilai seperti berikut:

4.1. Hasil Pengujian Probabilitas

Tabel 6. Probabilitas Kelas Data *Training Lazada*

Kelas	Banyak Tweet	Probabilitas Kelas
Positif	598	0.664
Negatif	302	0.335
Jumlah	900	

Tabel 7. Probabilitas Kelas Data *Training Tokopedia*

Kelas	Banyak Tweet	Probabilitas Kelas
Positif	722	0.802
Negatif	178	0.197
Jumlah	900	

Dengan menggunakan persamaan (2) pada tabel 6 dan 7 didapat probabilitas tiap kelasnya. Untuk *Lazada* di dapat 0.664 pada kelas positif dan 0.335 pada kelas negatif. Sedangkan pada *Tokopedia* di dapat 0.802 untuk probabilitas kelas positif dan 0.197 untuk kelas negatif.

4.2 Hasil Pengujian Performansi

Pada tahapan evaluasi performansi dibuatkan *confusion* matriks Lazada dan Tokopedia terhadap sentimen positif seperti pada tabel berikut.

Tabel 8. *Confusion* Matriks Lazada

	Prediksi Positif	Prediksi Negatif
Aktual Positif	175	25
Aktual Negatif	42	58

Tabel 9. *Confusion* Matriks Tokopedia

	Prediksi Positif	Prediksi Negatif
Aktual Positif	211	10
Aktual Negatif	34	45

dari tabel 8 dan 9 dapat dihitung nilai akurasi, *recall*, *precision* dan *f-measure* berdasarkan sentimen positif dan hasilnya adalah sebagai berikut:

Tabel 10. Nilai Akurasi, *Recall*, *Precision*, *F-measure*

	<i>Lazada</i>	<i>Tokopedia</i>
Akurasi	75.26%	83.97%
Recall	87.50%	95.48%
Precision	80.65%	86.12%
F-measure	83.93%	90.56%

Pada tabel 10 nilai akurasi, *recall*, *precision*, *f-measure* untuk prediksi berdasarkan sentimen positifnya. Tokopedia memiliki nilai persentase lebih besar dibandingkan data Lazada.

4.3 Analisis Hasil Pengujian

Dari hasil pengujian yang telah dipaparkan sebelumnya pada tabel 10, penulis dapat menarik beberapa analisis mengenai hasil yang diperoleh dari akurasi, *recall*, *precision*, dan *f-measure* berdasarkan sentimen positifnya. Akurasi terbaik yang dihasilkan oleh algoritma adalah akurasi dari data *Tokopedia* yaitu sebesar 83.97% dibandingkan data *Lazada* yang bernilai 75.26%. Selain itu *recall* pada data *Tokopedia* lebih besar dengan nilai 95.48% dibandingkan data *Lazada* yang bernilai 87.50%, *precision* data *Tokopedia* lebih besar dengan nilai 86.12% dibandingkan data *Lazada* yang bernilai 80.65% dan *f-measure* data *Tokopedia* lebih besar dengan nilai 90.56% dibandingkan data *Lazada* yang bernilai 83.93%. Kurangnya atribut yang digunakan pada data training serta adanya perbedaan hasil dari pelabelan secara manual dengan hasil prediksi dari klasifikasi model merupakan kekurangan dari penelitian ini.

5 Kesimpulan

Dari analisis percobaan yang dilakukan terkait sentimen terhadap Toko Online dengan metode Naive Bayes terhadap kata kunci Lazada dan Tokopedia. Maka dapat disimpulkan sebagai berikut:

1. Akurasi Naïve Bayes memberikan hasil pada Lazada sebesar 75.26% dan pada *Tokopedia* sebesar 83.97%. Pada pengolahan data *Lazada* kurang memuaskan karena data training sentimen sangat berbeda jauh perbedaannya antara data sentimen positif dan negatif.
2. Proses klasifikasi akan semakin akurat jika data training yang digunakan semakin banyak.
3. Proses pelabelan data menjadi dasar untuk membangun model yang lebih baik.

Daftar Pustaka

- [1]. Horn Christopher. 2010. Analysis and Classification of Twitter Message. Graz University of Technology
- [2]. Tan, Pang-Ning., Steinbach, Michael., dan Kumar., Vipin. 2006. Introduction to Data Mining. Boston: Addison Wesley.
- [3]. B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, pp. 1-135, 2008.
- [4]. Bifet, Albert., dan Frank, Eibe. 2010. Sentiment Knowledge Discovery in Twitter Streaming Data. Hamilton: Wakaito University.
- [5]. Horn Christopher. 2010. Analysis and Classification of Twitter Message. Graz University of Technology.
- [6]. Kunpeng, Zhang. A Systematic Framework for Sentiment Identification by Modeling User Social Effects. USA : Department of Information and Decision Sciences, University of Illinois at Chicago.
- [7]. Go A., Huang L., Bhayani R. 2009. Twitter Sentiment Analysis. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.
- [8]. Kamus Besar Bahasa Indonesia, Pengertian Sentimen, 2010.
- [9]. Yufis Azhar, Agus Zainal Arifin, Diana Purwitasari. 2013. "Otomatisasi Perbandingan Produk Berdasarkan Bobot Fitur pada Teks Opini".
- [10]. Pak, A., dan Paurobek, P., (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Universite de Paris-Sud, Laboratoire LIMSI-CNRS.
- [11]. Olson, David L.; & Delen, Dursun (2008); Advanced Data Mining Techniques, Springer, 1st edition (February 1, 2008), page 138, ISBN 3-540-76916-1.
- [12]. Risti, Khairunnisa N. dan Rahayu, Sri. "Shopping Online". Dinas Pendidikan Nasional. 2012. Blogspot 2 March 2012. <<http://wikispot-wikispot.blogspot.com/2012/03/karya-ilmiah-shopping-online.html#!/2012/03/karya-ilmiah-shopping-online.html>>, diakses tanggal 16 Agustus 2018.
- [13]. Kamus Besar Bahasa Indonesia, Pengertian Sentimen, 2010.
- [14]. J Ariawan, "Data Preprocessing." [Online]. Available: <<https://www.google.com/search?q=apa+itu+noise+dalam+data+mining&ie=utf8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a&channel=sb>>. [Accessed: 10-Des-2016].>, diakses tanggal 16 Agustus 2018.
- [15]. Librianty, A., 2017, "Ini 10 Toko Online Terbaik di Indonesia", [online], <<https://www.liputan6.com/tekno/read/3068210/ini-10-toko-online-terbaik-di-indonesia>>, diakses tanggal 16 Agustus 2018.
- [16]. Buntaro Ghulam, A., 2017, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter".
- [17]. Pertiwi, Kusuma, W., 2018, "Riset Ungkap Pola Pemakaian Medsos Orang Indonesia", [online], <<https://tekno.kompas.com/read/2018/03/01/10340027/riset-ungkap-pola-pemakaian-medsos-orang-indonesia>>, diakses tanggal 16 Agustus 2018.
- [18]. M. Yusuf Nur Sumarno Putro, (2011). "Analisis Sentimen pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine," masters, binus.
- [19]. N. Adiyasa, "Analisis Sentimen Pada Opini Berbahasa Indonesia Menggunakan Pendekatan Lexicon-Based," Catatan Kecil, 2011. [Online]. Available: <http://adiyasan.wordpress.com/2013/02/08/sentiment-analysis-menggunakanpendekatan-lexicon-based/>

