# Aplikasi Analisa Sentimen Pada Komentar Berbahasa Indonesia Dalam Objek Video di Website YouTube Menggunakan Metode Naïve Bayes Classifier

Maximillian Christianto.1, Justinus Andjarwirawan 2, Alvin Tjondrowiguno 3
Program Studi Informatika, Fakultas Teknologi Industri, Universitas Kristen Petra
Jl. Siwalankerto 121-131 Surabaya 60236
Telp. (031)-2983455, Fax. (031)-8417658x
E-Mail: agustinusmax678@gmail.com[1], justin@petra.ac.id [2], alvin.nathaniel@petra.ac.id

## ABSTRAK

Dari minat masyarakat di Indonesia yang semakin tinggi akan penggunaan YouTube, hal ini memicu munculnya *content creator* yang memilih YouTube sebagai media untuk berkarya. Sehingga para *content creator* berlomba - lomba untuk menghasilkan karya video yang dapat dinikmati pengguna YouTube. Berbagai metode dilakukan oleh para *content creator* untuk meningkatkan kualitas video yang dihasilkan.

Proses dilakukan pada skripsi ini merupakan proses mengolah data mentah yang telah dikumpulkan dari *YouTube*, sebelum dilakukan proses *training* dan klasifikasi. Proses data *preprocessing* perlu dilakukan untuk mengatasi data mentah yang bervariasi dan tidak konsisten sehingga dapat mempengaruhi proses *training* dan proses klasifikasi. Data *preprocessing* yang dilakukan pada skripsi ini meliputi *Tokenization, Stopwords Removal, Stemming*. Proses klasifikasi adalah proses dimana algoritma klasifikasi dijalankan kepada komentar yang digunakan sebagai *input* data komentar.

Aplikasi yang digunakan pada karya tulis ilmiah ini berhasil menghasilkan nilai *smoothing*, dimana nilai tersebut menunjukan bahwa komentar termasuk kedalam pengelompokan *class* sentimen positif, sentimen negatif atau data bukan bahasa Indonesia.

**Kata Kunci**: teks, API YouTube, klasifikasi sentiment

## *ABSTRACT*

*From the increasing interest of the people in Indonesia in the use of YouTube, this has triggered the emergence of content creators who choose YouTube as a medium for work. So that the content creators are competing to produce video works that can be enjoyed by YouTube users. Various methods are used by content creators to improve the quality of the video produced.*

*The process carried out in this thesis is the process of processing raw data that has been collected from YouTube, before the training and classification process. The process of data preprocessing needs to be done to overcome the raw data that is varied and inconsistent so that it can affect the training process and the classification process. Data preprocessing conducted in this thesis includes Tokenization, Stopwords Removal, Stemming. The classification process is the process by which the classification algorithm is run on comments that are used as input data for comments.*

*Applications used in scientific papers have succeeded in producing smoothing values, where the value shows that the comments belong to the classifications of positive sentiments, negative sentiments or non-Indonesian data.*

***Keywords:*** *texts, API YouTube, classification sentiment*

## 1. PENDAHULUAN

Dalam era perkembangan teknologi informasi yang semakin pesat di Indonesia. Masyarakat Indonesia tidak dapat lepas dari media sosial dalam kehidupan sehari-hari. Media sosial telah menjadi sarana bagi masyarakat untuk saling berkomunikasi dengan orang lain dengan beragam cara, salah satunya melalui video pada *website* YouTube. Berbagai sumber informasi dan hiburan dapat diakses tanpa terbatas akan jarak. Selain itu konten maupun pesan dapat diterima oleh orang lain secara cepat.

Dari minat masyarakat di Indonesia yang semakin tinggi akan penggunaan YouTube, hal ini memicu munculnya *content creator* yang memilih YouTube sebagai media untuk berkarya. Sehingga para *content creator* berlomba - lomba untuk menghasilkan karya video yang dapat dinikmati pengguna YouTube. Berbagai metode dilakukan oleh para *content creator* untuk meningkatkan kualitas video yang dihasilkan.

Untuk meningkatkan kualitas video, upaya umum yang sering dilakukan oleh *content creator* adalah mengecek komentar penonton secara manual. Hal ini dilakukan dengan membaca satu per satu komentar pada video yang diunggah ke *website* YouTube. Komentar yang telah dibaca menjadi bahan evaluasi *content creator* untuk meningkatkan kualitas video yang dibuat. Pada kenyataannya data komentar video yang masuk begitu banyak. Pengecekan manual kurang efektif karena menggunakan banyak sumber daya manusia yang intensif, dan membutuhkan banyak waktu.

Untuk menjawab permasalahan yang ada, diperlukan aplikasi otomatis yang dapat mengkategorikan komentar. *Text mining* penting dalam analisis sentimen sebagai pengidentifikasi emosional suatu pernyataan. Dengan melakukan analisa sentimen positif, dan negatif *content creator* dapat lebih mudah memilah komentar dari video yang diunggah. Diharapkan dengan adanya aplikasi penganalisis sentimen komentar secara otomatis, dapat membantu *content creator* dalam meningkatkan kualitas video yang dihasilkan..

## 2. LANDASAN TEORI

### 2.1 *Natural Language Procesing (NLP)*

*Natural Language Processing* (NLP) adalah bidang ilmu komputer dan teknik yang telah dikembangkan dari studi bahasa

dan linguistik komputasi dalam bidang *Artificial Intelligence*. Tujuan dari *Natural Language Processing* adalah merancang dan membangun aplikasi yang memfasilitasi interaksi manusia dengan mesin dan perangkat lain melalui bahasa alami manusia [10].

*Natural Language Processing* bekerja untuk mengenali dan memahami bahasa manusia. Dengan melakukan beberapa tahap yang mencerminkan ilmu bahasa yaitu *syntax*, *semantics* dan *pragmatics*[4]. Penjelasan ilmu bahasa tersebut dapat diuraikan sebagai berikut:

- *Syntax*
  Syntax merupakan bagian dari bahasa alami yang mengekspresikan proposisi, ide atau pemikiran dan mengungkapkan sesuatu. Pada umumnya syntax merupakan kata dari sebuah kalimat dengan urutan yang linear.
- *Semantics*
  *Semantics* merupakan penjelasan arti dari kalimat dalam suatu bahasa. Penjelasan arti suatu kalimat didapat melalui identifikasi struktur syntax yang membentuk suatu kalimat. Sehingga dengan melakukan identifikasi dari syntax suatu kalimat, maka didapatkan suatu makna dari kalimat tersebut.
- *Pragmatics*
  Pragmatiks menjelaskan bagaimana pernyataan yang ada berhubungan dengan keadaan yang ada. Dengan mempertimbangkan berbagai aspek seperti konteks kalimat dan tujuan dari speaker. Maka makna ucapan atau teks sesuai dalam konteks yang ingin disampaikan.

## 2.2 *Text Mining*

Teks merupakan data yang tidak terstruktur yang terbentuk dari susunan string yang biasa disebut dengan kata. Susunan kata tersebut memiliki arti yang kemudian digabungkan menjadi suatu kalimat. Informasi dari kalimat tersebut tidak dapat ditangkap oleh komputer. Oleh karena itu dengan melakukan *Text Mining* sebagai suatu proses mengekstrak informasi atau makna melalui data teks komputer dapat mengetahui makna dari suatu data teks. Hal yang biasa dilakukan adalah dengan *Classification*, *Clustering* dan *Association* [5].

## 2.3 Tanggapan

Tanggapan juga merupakan suatu pengalaman tentang objek peristiwa atau hubungan yang diperoleh dengan menggunakan informasi dan menafsirkan pesan. Proses menafsirkan pesan tersebut menyangkut masuknya pesan dan informasi ke dalam otak manusia. Melalui presepsi inilah manusia terus – menerus mengadakan hubungan dengan lingkungan, dimana hubungan ini dilakukan lewat indra manusia yaitu pengelihatan, pendengaran, peraba, perasa, dan penciuman[1].

## 2.4 Naïve Bayes Classifier

*Naïve Bayes Classifier* merupakan suatu metode klasifikasi yang menggunakan perhitungan probabilitas. Dengan penggunaan statistik berdasarkan teorema *bayes* yang mengasumsikan bahwa keberadaan atau ketiadaan dari suatu kelas dengan fitur lainnya. Pengklasifikasian statistik juga dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class*.

*Naïve Bayes* merupakan metode klasifikasi sangat sederhana dan efisien. Metode klasifikasi *Naïve Bayes* juga merupakan metode populer untuk klasifikasi teks dan memiliki peforma yang baik, namun juga terdapat kekurangan yang sangat sensitif dalam pemilihan fitur. Fitur yang terlalu banyak tidak hanya meningkatkan waktu perhitungan tapi juga dapat menurunkan akurasi klasifikasi[8].

*Bayesian classification* yang didasarkan pada teorema *bayes* yang memiliki kemampuan klasifikasi dengan akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* dengan data yang besar. Berikut bentuk umum dari teorema *bayes*:

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

**Rumus 1. Rumus Naïve Bayes**

P(H|X) = Probabilitas hipotesis H berdasar kondisi X (posteriori probability)
P(H) = Probabilitas hipotesis H (prior probability)
P(X|H) = Probabilitas X berdasar kondisi pada hipotesis H
P(X) = Probabilitas dari X

Untuk menjelaskan metode *Naive Bayes*, proses klasifikasi memerlukan sejumlah penunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis. Oleh karena itu, persamaan Naïve Bayes di atas disesuaikan sebagai berikut [6]

$$P(C \mid F1 \dots Fn) = \frac{p(C) \times p(F1 \dots Fn|C)}{p(F1 \dots Fn)}$$

**Rumus 1. Rumus Naïve Bayes**

$P(C \mid F1 \dots Fn)$ = Nilai probabilitas class C dari karakteristik $F1 \dots Fn$
$p(C)$ = Probabilitas class C
$F1 \dots Fn$ = Karakteristik petunjuk untuk melakukan klasifikasi
$p(F1 \dots Fn|C)$ = Nilai probabilitas karakteristik $F1 \dots Fn$ pada class C
$p(F1 \dots Fn)$ = Probabilitas karakteristik $F1 \dots Fn$

dalam suatu dataset, pemilihan data akan menyebabkan kemungkinan adanya nilai nol pada model probabilitas[2]. Nilai nol tersebut menyebabkan Naïve Bayes Classifier tidak dapat melakukan klasifikasi data yang diinputkan, oleh karena itu diperlukan metode yang dapat dilakukan untuk menghindari nilai nol dalam perhitungan probabilitas. *Laplacian Smoothing* atau yang biasa dikenal dengan *add one smoothing* dapat menjadi solusi untuk menghindari munculnya angka nol, karena dalam perhitungannya parameter $n_k$ ditambah dengan 1.

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

**Rumus 1. Rumus Naïve Bayes Classifier *Add One Smooth Method***

P(Fk | Ci) = Nilai probabilitas Fk pada class Ci
nk = Jumlah sampel data atribut nk
n = Jumlah sampel data dalam class dari atribut nk
|Vocabulary| = Jumlah sampel data

## 2.5 *K-fold Cross Validation*

K-fold cross validation adalah teknik yang dapat digunakan apabila memiliki jumlah data yang terbatas (jumlah instance tidak banyak). K-fold cross validation merupakan salah satu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan cara melakukan perulangan dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut input yang acak. K-fold cross validation diawali dengan membagi data sejumlah n-fold yang diinginkan. Dalam proses cross validation data akan dibagi dalam n buah partisi dengan

ukuran yang sama D1,D2,D3..Dn selanjutnya proses testing dan training dilakukan sebanyak n kali. Dalam iterasi ke-I partisi akan menjadi data testing dan sisanya akan menjadi data training. Untuk penggunaan jumlah fold terbaik untuk uji validitas, dianjurkan menggunakan 10-fold cross validation dalam model[9].

## 2.6 *Confusion Matrix*

*Confusion matrix* adalah suatu metode yang digunakan untuk melakukan perhitungan akurasi pada *text mining*[3]. *Confusion matrix* digambarkan dengan tabel yang menyatakan jumlah data uji benar diklasifikasikan dan data uji yang salah diklasifikasikan. Berikut adalah tabel dari *confusion matrix*:

| *Confusion Matrix* | | Prediksi | |
|---|---|---|---|
| | | Positif "+" | Negatif "-" |
| Aktual | Positif "+" | TP (True Positive) | FN (False Negative) |
| | Negatif "-" | FP (False Positive) | TN (True Negatif) |

Tabel 1. Tabel Confusion Matrix

## 2.7 *Precision*

*Precision* didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih. *Precision* dapat diartikan kecocokan atau presentase antara fakta data yang benar dengan prediksi positif.

$$Precision = \frac{TP}{(TP + FP)}$$

**Rumus 2. Rumus perhitungan *f-measure***

## 2.8 *Recall*

*Recall* digunakan untuk mengukur data kelompok pola positif yang diklasifikasikan dengan benar atau presentase prediksi positif dengan fakta data yang bernilai positif.

$$Recall = \frac{TP}{(TP + FN)}$$

**Rumus 3. Rumus perhitungan *f-measure***

## 2.9 *F-Measure*

*F Measure* merupakan suatu pengukuran yang menggambarkan rata - rata antara *recall* dan *precision*.positif.

$$F\ Measure = 2\ x \frac{Recall\ x\ Precision}{Recall + Precision}$$

**Rumus 4. Rumus perhitungan *f-measure***

## 2.10 Akurasi

*Recall* digunakan untuk mengukur data kelompok pola positif yang diklasifikasikan dengan benar atau presentase prediksi positif dengan fakta data yang bernilai positif

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

**Rumus 5. Rumus perhitungan *f-measure***

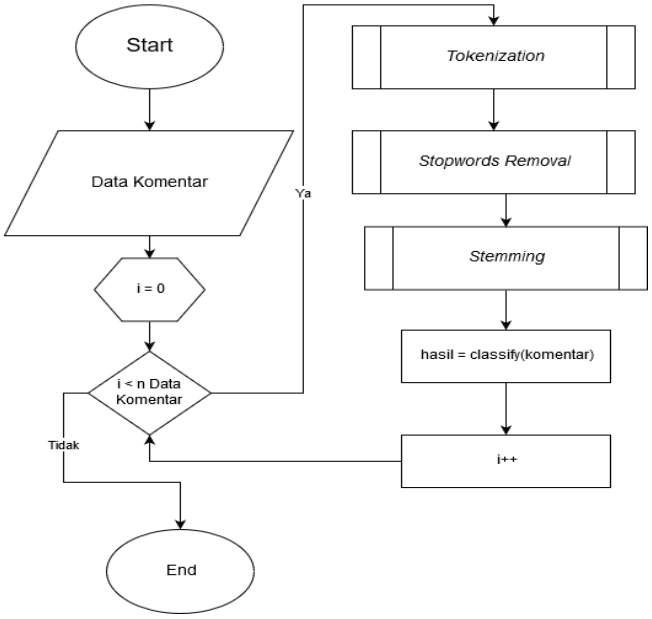
# 3. METODE

Garis besar proses yang dilakukan dapat dilihat pada Gambar 1. Terdapat empat proses utama yang terdiri dari *Tokenization, Stopwords Removal, Stemming dan Clasification.* Aplikasi dibuat menggunakan Sublime Text 3 sebagai *text editor* yang mendukung pengerjaan aplikasi dengan bahasa PHP, HTML, JavaScript dan Jquery. Bahasa pemrograman yang digunakan adalah PHP menggunakan *framework composer* untuk menginstall *library* yang diperlukan seperti *library* sastrawi, selain itu juga menggunakan database MySQL sebagai tempat menyimpan data train. Data testing yang akan diproses pada model didapat dari API YouTube. Proses klasifikasi naïve bayes dan perhitungan confusion matrix menggunakan library varunon.

Setelah pembuatan aplikasi dan pengujian pada skripsi ini, penulis melakukan beberapa uji coba yang dilakukan dalam pengujian ini. Pengujian dilakukan untuk mendapatkan metode data *preprocessing* terbaik. Jumlah data komentar yang dikumpulkan sebanyak 4229 (empat ribu dua ratus dua puluh sembilan). Dari data tesebut, 1023(seribu dua puluh tiga) data positif, 1030 (seribu tiga puluh) data negatif dan 2176 (dua ribu seratus tujuh puluh enam) data bukan bahasa Indonesia.

## 3.1 *Data Preprocessing*

Proses data *preprocessing* diperlukan karena data komentar dari *website* YouTube masih bersifat mentah. Data komentar yang bervariasi dan inkonsisten perlu dipersiapkan dengan serangkaian proses untuk mengurangi variasi dan menambah konsistensinya. MySQL digunakan untuk memuat data komentar, data kata singkatan dan data *stopwords*. Proses selanjutnya adalah *tokenization, stopwords removal* dan *stemming* yang dilakukan secara berurutan.




**Gambar 1. Diagram klasifikasi komentar**

## 3.2 *Tokenizing*

*Tokenization* merupakan proses pemotongan kalimat pada sebuah set dokumen teks menjadi potongan kata – kata atau karakter yang sesuai dengan kebutuhan sistem. Pemecahan kalimat dan kata dilakukan berdasarkan pada spasi di dalam kalimat atau paragraf. Dalam tahapan *tokenization* juga melakukan penghilangan karakter-karakter tertentu seperti tanda baca dan mengubah semua kata menjadi huruf kecil (*lowercase*). Potongan-potongan tersebut dikenal dengan istilah token [7].

## 3.3 *Stemming*

*Stemming* adalah proses yang menyediakan pemetaan varian morfologi yang berbeda dari kata-kata ke dalam basis / kata umum[12]. Proses ini juga dikenal sebagai *conflation*. Berdasarkan asumsi bahwa istilah yang memiliki *stem* biasanya memiliki arti yang sama, proses *stemming* banyak digunakan dalam *information*

*retrieval* atau pangambilan informasi sebagai salah satu cara untuk meningkatkan kinerja pengambilan. Selain kemampuannya untuk meningkatkan kinerja pengambilan informasi, proses *stemming* yang dilakukan pada saat pengindeksan juga akan mengurangi ukuran *file* indeks.

Algoritma stemming untuk bahasa yang satu berbeda dengan algoritma stemming untuk bahasa lainnya. Sebagai contoh Bahasa Inggris memiliki morfologi yang berbeda dengan Bahasa Indonesia sehingga algoritma stemming untuk kedua bahasa tersebut juga berbeda. Pada teks berbahasa Inggris, proses yang diperlukan hanya proses menghilangkan sufiks. Sedangkan pada teks berbahasa Indonesia lebih rumit/kompleks karena terdapat variasi imbuhan yang harus dibuang untuk mendapatkan *root word* dari sebuah kata[13].

## 3.4 *Stopwords Removal*

*Stopwords Removal* adalah suatu proses menghilangkan kata-kata umum yang tidak memiliki makna. Penghilangan beberapa kata kerja, kata sifat, dan kata keterangan lainya dapat dimasukan ke dalam daftar *stopwords*. Karakteristik utama dalam pemilihan *stopword* biasanya adalah kata yang mempunyai frekuensi kemunculan yang tinggi misalnya kata penghubung seperti “dan”, “atau”, “tapi”, “sehingga” dan lainnya. Tujuan utama dari *stopwords Removal* adalah menghilangkan kata-kata yang tidak memiliki makna, sehingga dapat meningkatkan kecepatan dan peforma pemrosesan. *Stopword* termasuk penentu, konjungsi, preposisi dan sejenisnya[11]

# 4. PENGUJIAN

Pengujian dilakukan untuk mendapatkan metode *preprocessing* mana yang perlu dilakukan dari beberapa metode yang telah diusulkan, sehingga menghasilkan akurasi klasifikasi tertinggi. Durasi dari setiap proses juga akan dihitung sebagai bahan pertimbangan tambahan.

Pengujian pada skripsi ini dengan pengujian *k-fold cross validation* dan pengujian data tes diluar *dataset training*. Pada pengujian *k-fold cross validation* dataset diacak agar tidak berurutan, kemudian data train yang telah diacak tersebut dibagi menjadi 5, 10 dan 20 subset data yang kemudian dilakukan iterasi sebanyak nilai bagi data tersebut. Setiap iterasi dilakukan pengambilan satu *subset* data sebagai data tes, kemudian *subset* data lainnya dilakukan *training* kembali. Pengujian dilakukan dengan menggunakan data tes yang telah diambil diklasifikasikan dengan model yang telah di dapat dari proses *training*. Pada pengujian *k-fold cross validation* akan didapatkan hasil akurasi, *precision, recall, fscore* dan waktu proses yang dibutukan. Hasil pengujian k-fold cross validation dapat dilihat dibawah ini :

**Tabel 2. Hasil 5-*Fold Cross Validation***

| Pre processing | Akurasi | Precision | Recall | Fscore | Waktu |
|---|---|---|---|---|---|
| Sebelum *Preprocessing* | 0,898 | 0,762 | 0,807 | 0,783 | 156,09 s |
| *Stemming* | 0,904 | 0,773 | 0,823 | 0,797 | 147,58 s |
| *Stopwords Removal* | 0,892 | 0,775 | 0,779 | 0,777 | 140,75 s |

**Tabel 3. Hasil 10-*Fold Cross Validation***

| Pre processing | Akurasi | Precision | Recall | Fscore | Waktu |
|---|---|---|---|---|---|
| Sebelum *Preprocessing* | 0,901 | 0,773 | 0,813 | 0,791 | 82,32 s |
| *Stemming* | 0,910 | 0,777 | 0,828 | 0,801 | 77,71 s |
| *Stopwords Removal* | 0,900 | 0,779 | 0,781 | 0,779 | 68,58 s |

**Tabel 4. Hasil 20-*Fold Cross Validation***

| Pre processing | Akurasi | Precision | Recall | Fscore | Waktu |
|---|---|---|---|---|---|
| Sebelum *Preprocessing* | 0,902 | 0,776 | 0,811 | 0,797 | 41,19 s |
| *Stemming* | 0,91 | 0,78 | 0,83 | 0,80 | 39,43 s |
| *Stopwords Removal* | 0,892 | 0,776 | 0,781 | 0,777 | 33,58 s |

# 5. KESIMPULAN

Dari percobaan yang telah dilakukan dapat disimpulkan bahwa metode *preprocessing stemming* dapat meningkatkan akurasi dan mempercepat proses klasifikasi, sedangkan *stopword removal* menurunkan akurasi dan mempercepat proses klasifikasi.

Dari percobaan *k-fold cross validation* dengan pembagian data menjadi 5, 10 dan 20 *fold*, dengan metode klasifikasi *naïve bayes classifier* bergantung pada varias kata yang ada pada *dataset train*. Kesimpulan dari penelitian ini adalah semakin banyak variasi kata yang dipelajari oleh model, maka akurasi akan semakin baik.

# 6. REFERENCES


[1] Agustriono. Wiwit. 2012 Presepsi Guru Terhadap Komunikasi Kepala Sekolah Di Sekolah Menengah Atas Negeri 1 Tapung Kecamatan Tapung Kabupaten Kampar, (p.37). Pekanbaru, Indonesia: Universitas Islam Negeri Sultan Syarif Kasim Riau.

[2] Cahyanti. A. F., Saptono. R., Sihwi. S. W. 2015. Penentuan Model Terbaik Metode Naïve Bayes Classifier Dalam Menentukan Status Gizi Balita Dengan Mempertimbangkan Dependensi Parameter. Surakarta: Universitas Sebelas Maret

[3] Han. J. 2006. Data Mining: Concepts And Techniques. 2nd edition (p. 24). Amsterdam: University of illinois at Urbana-Champaign

[4] Indurkhya. N., & Damerau. F. J. 2010. Handbook Of Natural Language Processing. 2nd edition, (pp. 3-7).

[5] Jo. Taeho. 2019. Text Mining Concepts, Implementation, And Big Data Challange , (p.1). Seoul, Korea: Hongik University.

[6] Kusrini., & Luthfi. E. T. 2009. Algoritma Data Mining (pp. 190-191). Yogyakarta: ANDI.

[7] Manning, D. M., Raghavan, P., & Schutze, H. 2008. Introduction to Information Retrieval. Cambridge, United Kingdom: Cambridge University Press.

[8] Muthia, D. A. 2017. Analisis Sentimen Pada Review Restoran Dengan Teks Bahasa Indonesia Menggunakan Algoritma Naïve Bayes. Bekasi: Akademi Manajemen Informatika dan Komputer Bina Sarana Informatika

[9] Pitria. P. 2014. Analisis Sentimen Pengguna Twitter Pada Akun Resmi Samsung Indonesia Dengan Menggunakan Naïve Bayes. Bandung: Universitas Komputer Indonesia

[10] Pustejovsky. J., &Stubbs. Amber. 2012.Natural Language Annotation for Machine Learning, (pp. 1-2). United State of America: O'Reilly

[11] Riany. J., Fajar. Mohammad., & Lukman. M. P. 2016. Penerapan Deep Sentiment Analysis Pada Angket Penilaian Terbuka Menggunakan K-Nearest Neighbor. Makasar: STMIK Kharisma

[12] Tala, F. 2003. A study of stemming effects on information retrieval in Bahasa Indonesia. Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands. Amsterdam.

[13] Wahyudi. D., Susyanto. T., & Nugroho.. D. 2017. Implementasi Dan Analisi Algoritma Stemming Nazief & Adriani Dan Porter Dokumen Berbahasa Indonesia. Surakarta: STMIK Sinar Nusantara Surakarta