

# PERANCANGAN SISTEM ANALISIS SENTIMEN MASYARAKAT PADA SOSIAL MEDIA DAN PORTAL BERITA

Rizqon Sadida<sup>1)</sup>, Moh Royandi Azkia<sup>2)</sup>, Beda Puspita Candra<sup>3)</sup>,  
Novandi Rezeki<sup>4)</sup>, M. Ozzy Calvin Rendy<sup>5)</sup>

<sup>1), 2), 3), 4), 5)</sup> Magister Teknik Informatika STMIK AMIKOM Yogyakarta

Jl Ring road Utara, Condongcatur, Sleman, Yogyakarta 55281

Email : rizqon.s@students.amikom.ac.id<sup>1)</sup>, royandiazkia@gmail.com<sup>2)</sup>, bedapuspitacandra@gmail.com<sup>3)</sup>,  
novandinovan@gmail.com<sup>4)</sup>, m.ozzy.r@gmail.com<sup>5)</sup>

## Abstrak

Menjelang pemilihan Gubernur DKI Jakarta saat ini tengah menjadi trending topic diberbagai media massa dan juga di sosial media. yang mana pada sosial media tersebut terdapat percakapan-percakapan yang di aspirasikan masyarakat, dari percakapan tersebut bisa didapatkan sebuah data tekstual yang mengandung sentimen berupa persepsi dari pengguna internet terhadap para calon gubernur.

Dengan adanya data dari sosial media tersebut dapat dilakukan sebuah analisis sentimen atau opinion mining yaitu sebuah proses yang memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi. Analisis ini menerapkan metode Natural Language Processing (NLP) menggunakan algoritma N-Gram untuk menganalisa bahasa dan Support Vector Machine (SVM) untuk mengklasifikasikan konten dari bahasa manusia yang ada di media sosial dan portal berita.

Hasil dari rancangan ini diharapkan memberikan gambaran tentang penerapan sistem analisis sentimen masyarakat yang bersumber dari sosial media dan portal berita. Hasil dari proses berupa informasi grafik sentimen positif, negatif, dan netral dari media online.

**Kata kunci:** Analisis Sentimen, Natural Language Processing, N-Gram, Support Vector Machine

## 1. Pendahuluan

Sosial media dan portal berita tidak hanya menjadi salah satu wadah populer untuk memperbincangkan suatu masalah, tetapi juga menjadi wadah untuk menampung sentimen masyarakat tentang sesuatu objek yang dianggap viral.

Sosial media dan portal berita merupakan sumber data yang sangat besar, dimana data tersebut berbentuk UGC (User Generated Content). Data tersebut tentu sangat berguna untuk berbagai keperluan, salah satunya adalah pilgub DKI 2017. Akan tetapi karena data tersebut sangat banyak dan besar, manusia tidak mampu untuk mengelompokkan data tersebut satu-persatu.

Perancangan sebuah sistem yang mampu menganalisis sentimen masyarakat pada sosial media dan portal berita dianggap sebagai sebuah salah satu solusi untuk mengatasi masalah diatas. Analisis sentimen masyarakat

pada facebook, twitter, kompas, dan detik dengan menerapkan Natural Language Processing (NLP) menggunakan Support Vector Machine (SVM) dan N-Gram sebagai algoritma untuk memprediksi nilai dari sebuah kata.

## Rumusan Masalah

Untuk merumuskan masalah yang ada agar dapat dijadikan acuan penelitian yaitu bagaimana memberikan gambaran tentang rancangan sistem analisis sentimen yang dapat dijadikan sebagai alat untuk mengukur sentiment masyarakat tentang sesuatu objek?

## Tujuan Penelitian

Dengan rumusan masalah tersebut tujuan dapat dicapai adalah untuk memberi gambaran tentang rancangan sebuah sistem yang mampu menganalisis sejauh mana sentimen masyarakat tentang objek.

## Penelitian Sebelumnya

Noviah Dwi Putranti dan Edi Winarko (2014) Universitas Gadjah Mada melakukan penelitian berjudul Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan Maximum Entropy dan Support Vector Machine. Data opini diperoleh dari jejaring sosial Twitter berdasarkan query dalam Bahasa Indonesia. Penelitian ini bertujuan untuk menentukan sentimen publik terhadap objek tertentu yang disampaikan di Twitter dalam bahasa Indonesia, sehingga membantu usaha untuk melakukan riset pasar atas opini publik.[1] Syahfitri Kartika Lidya, Opim Salim Sitompul dan Syahril Efendi (2015) Universitas Sumatra Utara melakukan penelitian berjudul Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (Svm) Dan K-Nearest Neighbor (K-Nn). menggunakan teks Bahasa Indonesia yang terdapat di website berupa artikel berita, kemudian metode K-Nearest Neighbor akan mengklasifikasi langsung pada data pembelajaran agar dapat menentukan model yang akan dibentuk oleh metode Support Vector Machine untuk menentukan kategori dari data baru yang ingin ditentukan kategori tekstual, yaitu kelas sentimen positif, negatif dan netral.[2]

AndiNurulHidayat (2015) STMIK Bina Mulia melakukan penelitian berjudul Analisis Sentimen Terhadap Wacana Politik Pada Media Masa Online Menggunakan Algoritma Support Vector Machine Dan

*Naive Bayes*. Mengatakan bahwa manfaat analisis sentimen dalam dunia politik antara lain untuk membantu dalam menganalisis kebijakan publik pemerintah serta memberikan efisiensi waktu dan efisiensi kerja bagi para penyedia berita dalam mengklasifikasikan berita dan membantu pencari berita untuk mendapatkan wacana berita politik harian yang mereka inginkan.[3]

**Natural Language Processing (NLP)**

*Natural Language Processing (NLP)* merupakan salah satu cabang ilmu AI yang berfokus pada pengolahan bahasa natural. Bahasa natural adalah bahasa yang secara umum digunakan oleh manusia dalam berkomunikasi satu sama lain. Bahasa yang diterima oleh komputer butuh untuk diproses dan dipahami terlebih dahulu supaya maksud dari *user* bisa dipahami dengan baik oleh komputer.[4]

**N-Gram**

N-gram adalah potongan *n* karakter dalam suatu string tertentu atau potongan *n* kata dalam suatu kalimat tertentu.[5] Misalnya dalam kata “Belajar” akan didapatkan n-gram sebagai berikut.

- Unigram : B, E, L, A, J, A, R
- Bigram : \_B, BE, EL, LA, AJ, JA, R\_
- Trigram : \_BE, BEL, ELA, LAJ, AJA, JAR AR\_, R \_ \_

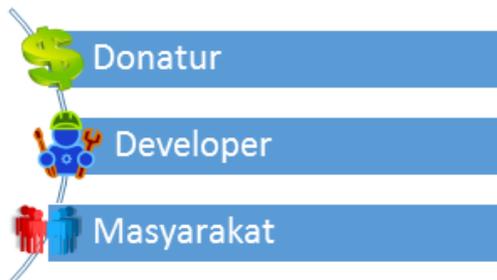
Contoh pemotongan N-gram berbasis karakter :  
 Karakter blank “\_” digunakan untuk merepresentasikan spasi di depan dan diakhir kata.

**Support Vector Machine (SVM)**

*Support Vector Machine (SVM)* adalah sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi-fungsi linier dalam sebuah ruang fitur (feature space) berdimensi tinggi, dilatih dengan algoritma pembelajaran yang didasarkan pada teori optimasi dengan mengimplementasikan learning bias yang berasal dari teori pembelajaran statistik.[6]

**2. Pembahasan**

**Arsitektur Stakeholder**



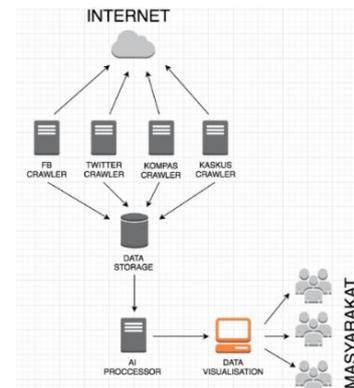
Gambar 1. Stakeholder

Pada gambar 1 diatas ada 3 buah elemen yang menyusun arsitektur stakeholder tersebut yaitu donatur, developer dan masyarakat.

- Donatur disini berperan sebagai sumber dana untuk melakukan perancangan sistem analisis sentimen masyarakat yang akan dibangun. Dalam kasus pemilihan gubernur DKI jakarta 2017 ini donatur adalah lembaga penyedia quick count atau media berita cetak dan eletronik.
- Developer sendiri adalah stakeholder yang bertanggungjawab penuh dalam merancang sistem yang akan dibangun sampai kemudian diimplementasikan sesuai dengan kebutuhan yang disepakati sebelumnya dengan donatur.
- Masyarakat dijadikan sebagai sumber data yang nantinya akan diolah oleh sistem yang kemudian akan menghasilkan informasi untuk kepentingan yang ada. Data sentimen masyarakat tersebut didapatkan melalui sosial media dan portal berita.

**Arsitektur Teknologi**

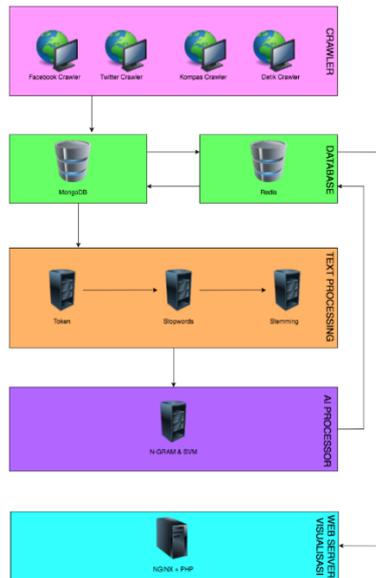
Pada gambar 2 dijelaskan arsitektur teknologi yang digunakan untuk merancang sistem analisis sentimen, dimana data-data diambil dari internet melalui bot crawl yang kemudian akan diolah pada sebuah server oleh AI Processor.



Gambar 2. Arsitektur Teknologi

**Proses Olah Data**

Pada gambar 3, dijelaskan terdapat lima bagian yang berperan dalam rancangan sistem sentiment analysis ini, antara lain crawling, database, text processing, ai processor, dan web server. adapun penjelasannya dibawah ini.



Gambar 3. Olah Data

**Crawling**

Pada tahap ini, sistem akan mengambil data dari source yang telah di tentukan seperti facebook, twitter, detik, dan kompas. karena data yang di ambil akan sangat banyak, maka dalam melakukan process crawling akan di sediakan masing-masing satu server untuk satu source. Adapun kebutuhan untuk crawling akan dijelaskan pada tabel 1. untuk kebutuhan hardware crawling dijelaskan pada tabel 2, dimana pada rancangan digunakan Digital Ocean sebagai Platform Crawling Server, alasannya mengapa menggunakan Digital Ocean dikarenakan benefit yang diberikan lebih besar dan biaya yang lebih sedikit bila dibandingkan dengan platform lain seperti Microsoft Azure, AWS, Linode, dan Vultr. perbandingan ini dapat dilihat pada gambar 4.[9]

Provider	Microsoft Azure	Amazon Web Services	Digital Ocean	Linode	Vultr
Plan	A1 Small	T2-micro	1GB	1GB	1GB
Cost per Month (\$)	\$38	\$15	\$10	\$10	\$10
Memory (GB)	1.75GB	1GB	1GB	1GB	1GB
Disk Space (GB)	30GB	8GB	30GB	24GB	20GB
CPU Cores	1	1	1	1	1
Data Transfer (GB)	100GB	200GB	2TB	2TB	2TB
Virtualization	Hyper-V	Xen	KVM	KVM	KVM

Gambar 4. Perbandingan Cloud Server

Tabel 1. Kebutuhan Software Crawling Server

Nama	Keterangan
Sistem Operasi	Centos 7
Bahasa Pemrograman	Phyton 2.7
Python Library	Scrappy

Tabel 2. Kebutuhan Hardware & Cost Crawling Server

Nama	Keterangan
Platform	Digital Ocean

Name	1GB
Memory	1GB
Storage	30GB/SSD
Data Transfer	2TB
Total Instances	4

Dalam melakukan pengambilan data, Crawling untuk facebook menggunakan Facebook Graph, dan Twitter Menggunakan Twitter API, keduanya merupakan web service yang telah disediakan oleh masing-masing vendor. Sedangkan untuk kompas dan detik menggunakan Scrappy, salah satu library yang ada di python yang berfungsi untuk mengambil data berdasarkan tag html. keterangan pengambilan data source dapat dilihat pada tabel 3 dibawah ini.

Tabel 3. Metode Scrapping

Source	Metode	Target
Facebook	Facebook Graph API	Status, Komentar
Twitter	Twitter REST API	Tweet
Kompas	Scrappy	Komentar
Detik	Scrappy	Komentar

Pada tabel diatas terdapat kolom bernama Target yang berarti sumber data yang diambil. Sebagai contoh kasus adalah facebook, maka yang di ambil untuk *sentiment analysis* berupa status dan komentar pada facebook, begitu pula dengan source lainnya.

**Database**

Database merupakan unsur penting dalam perancangan sistem ini. Karena sistem ini berjalan realtime, dan data yang diolah akan sangat banyak, dan berlangsung cepat, maka dibutuhkan suatu sistem yang dapat mengampu aktifitas tersebut. Oleh karena itu kami menggunakan dua DBMS untuk menyimpan data, antara lain adalah Redis dan MongoDB. Redis adalah sistem storage yang menyimpan datanya pada memory [http://redis.io/topics/introduction], karena data yang di simpan dalam memory lebih cepat dibandingkan dengan hardisk, karena memory tidak ada mekanisme untuk mencari data[9], atas dasar tersebut maka kami merancang redis untuk dipasang sebagai front data storage, dimana aktifitas I/O data berlangsung dengan cepat. MongoDB adalah salah satu DBMS yang bersifat No-SQL, kelebihan dari MongoDB ini adalah karena mendukung replikasi data dan mudah untuk melakukan scaling. singkatnya DBMS sangat berguna untuk digunakan pada database yang besar, mengingat sistem ini akan mengambil data yang sangat banyak dari beberapa source, server database ini akan membutuhkan sistem khusus yang mendukung untuk scaling. Karena itu digunakan Digital Ocean Block Storage. kelebihan block storage ini adalah kapasitas hardisk yang dinamis hingga dapat disesuaikan dengan kebutuhan. adapun spesifikasi dari server database akan di jelaskan pada tabel 4 dan 5 berikut.

Tabel 4. Spesifikasi Database Server (MongoDB)

Nama	Keterangan
Nama	Paket 1GB
Memory	1GB
Storage	30GB
Block Storage	100GB

Tabel 5. Spesifikasi Database Server (Redis)

Nama	Keterangan
Nama	1GB
Memory	1GB
Storage	30GB

**Text Processing**

Pada bagian ini akan membahas penerapan NLP dalam pengolahan text, sebelum masuk pada tahap penilaian sentiment, data yang diambil harus diolah terlebih dahulu agar memudahkan untuk pengambilan *sentiment score*, adapun contoh data yang akan diolah dapat dilihat pada gambar 5.



Gambar 5. Contoh Data

Adapun langkah-langkah text processing akan dijelaskan pada bagian di bawah ini :

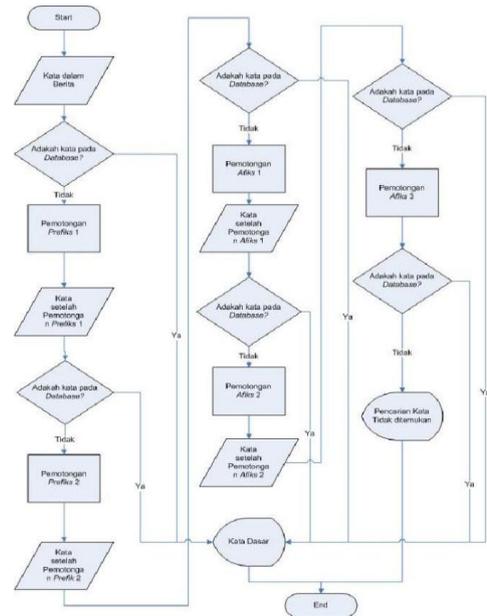
1. Tokenisasi

Tokenisasi berfungsi untuk memecah teks menjadi token-token yang berurutan. Kami memecah teks menggunakan whitespace, dan mengubahnya menjadi huruf kecil supaya seragam. Berdasarkan gambar 5, maka teks tersebut akan di pecah-pecah menjadi token sebagai berikut. "jangan", "sedih", "ya", "pak", "oyan", "moslem2", "moderat", "tetap", "pilih", "bapak", "kok", "untuk", "jakarta", "oyan", "pasti", "menang".

2. Stopwords Removal

Stopwords adalah kumpulan kata dasar (root word). pada semua implementasi NLP baik itu dalam bahasa inggris ataupun bahasa indonesia dibutuhkan stopwords removal. fungsi dari stopwords removal adalah untuk menghilangkan sejumlah kelas kata penghubung.[8] hal ini dilakukan untuk meningkatkan efektifitas sistem sehingga data yang akan di olah adalah konten yang dianggap penting saja. sebagai contoh kita akan mengambil data sebelumnya yang telah di tokenisasi. "maju", "terus", "pak", "oyan", "untuk", "jakarta", "oyan", "pasti", "menang". teks yang terdapat pada stopwordlist berupa "dua", "pak", "terus". maka dari proses stopword removal akan di dapati hasil berupa. "maju", "oyan", "pasti", "menang".

3. Stemming



Gambar 6. Algoritma Nazief

Stemming adalah suatu metode untuk mapping token ke bentuk dasarnya. Pada proses ini digunakan algoritma nazief yang dapat dilihat pada gambar 6.

Langkah-langkah dari algoritma tersebut dijelaskan dibawah ini : [7]

1. Mencari kata yang akan dicari sistem dalam kamus kata dasar. Jika ditemukan maka diasumsikan kata adalah root word. Maka algoritma berhenti.
2. Inflection Suffixes (“-lah”, “-kah”, “-ku”, “-mu”, atau “-nya”) dibuang. Jika berupa particles (“-lah”, “-kah”, “-tah” atau “-pun”) maka langkah ini diulangi lagi untuk menghapus Possesive Pronouns (“-ku”, “-mu”, atau “-nya”), jika ada.
3. Menghapus Derivation Suffixes (“-i”, “-an” atau “-kan”). Jika kata ditemukan di kamus, maka algoritma berhenti. Jika tidak maka ke langkah selanjutnya.
  - a. Jika “-an” telah dihapus dan huruf terakhir dari kata tersebut adalah “-k”, maka “-k” juga ikut dihapus. Jika kata tersebut ditemukan dalam kamus maka algoritma berhenti. Jika tidak ditemukan maka lakukan langkah b.
  - b. Akhiran yang dihapus (“-i”, “-an” atau “-kan”) dikembalikan.
4. Menghapus Derivation Prefix. Jika pada langkah 3 ada sufiks yang dihapus maka pergi ke langkah a, jika tidak pergi ke langkah b.
  - a. Periksa tabel kombinasi awalan-akhiran yang tidak diijinkan. Jika ditemukan maka algoritma berhenti, jika tidakpergi ke langkah.
  - b. For i = 1 to 3, tentukan tipe awalan kemudian hapus awalan. Jika root word belum juga ditemukan lakukan langkah 5, jika sudah maka algoritma berhenti. Catatan: jika awalan kedua sama dengan awalan pertama algoritma berhenti.
5. Melakukan Training Data.

6. Jika semua langkah telah selesai tetapi tidak juga berhasil maka kata awal diasumsikan sebagai root word. Proses selesai.

Sebagai contoh untuk penerapan algoritma diatas berupa teks “saya menyukai orang yang mempunyai wibawa tinggi”, maka bila di terjemahkan mulai dari proses tokenisasi hingga stemming, token yang dihasilkan berupa “saya”, “suka”, “orang”, “punya”, “wibawa”. Adapun rancangan kebutuhan spesifikasi server yang di perlukan dalam memangun sistem ini dalam bagan text processing akan di jelaskan pada tabel 6.

Tabel 6. Spesifikasi Text Process Server

Nama	Keterangan
Nama	1GB
Storage	30GB
Memory	1GB
Instances	3

### AI Processor

Pada bagian AI Processor sistem ini akan membutuhkan spesifikasi server yang cukup besar dikarenakan data yang akan diolah akan berlangsung dengan cepat. Spesifikasi server paling tidak harus mempunyai 2 CPU Core, dengan RAM 2GB. maka dari pada itu spesifikasi lengkap akan diterangkan pada tabel 7 dibawah ini.

Tabel 7. Spesifikasi AI Processor Server

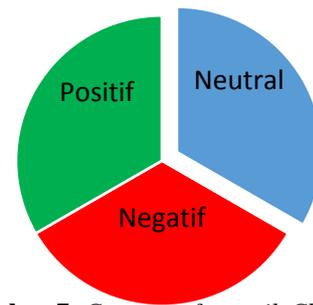
Nama	Keterangan
Nama	2GB
Storage	40GB
CPU	2 Cores

### Web Server & Visualisasi

Ini adalah bagian terakhir dari sistem ini dimana data yang telah diolah dan *sentiment analysis* yang didapatkan akan di gambarkan dalam bentuk grafik. grafik ini akan di tampilkan berbasis website, oleh sebab itu rancangan ini membutuhkan web server. adapun spesifikasi dari webserver ini akan di terangkan pada tabel 8. dan untuk library grafik yang akan kami gunakan adalah charts.js, sebuah library css untuk membuat grafik pada website, contoh grafik dapat dilihat pada gambar 7.

Tabel 8. Spesifikasi Web Server

Nama	Keterangan
Nama	1GB
Memory	1GB
Storage	30GB



Gambar 7. Contoh Infographic Charts.js

### 3. Kesimpulan

Penerapan sistem analisis sentimen masyarakat pada sosial media dan portal berita dinilai sangat efektif guna memberi sedikit informasi yang dapat dijadikan acuan dalam banyak hal sesuai dengan tujuan penerapan sistemnya. Maka dari itu perancangan sistem ini ditujukan untuk memberi gambaran umum untuk organisasi atau perseorangan yang ingin membangun sistem analisis sentimen.

Proses pengambilan data dilakukan oleh *bot crawler* pada sosial media dan portal berita. Kemudian data mentah akan disimpan pada *database* lalu diolah dengan *text processing* yang didalamnya terdapat beberapa metode yaitu *tokenisasi*, *stopwords removal* dan *stemming* menggunakan algoritma *naief*. Setelah data diolah pada *text processing* didapatkan kumpulan kata-kata baku yang kemudian diberikan nilai oleh *AI processor* yang nantinya akan dijadikan sebagai perhitungan guna memperoleh hasil sentimen tersebut mengandung arti positif, negatif atau netral pada sebuah objek. Hasil yang diperoleh dari proses pengolahan data tersebut dituangkan dalam bentuk informasi grafik yang tertera pada sebuah website.

Beberapa keuntungan yang didapat dari penerapan sistem analisis sebagai berikut :

1. Hasil yang dari pengolahan data sentimen masyarakat digunakan sebagai tolak ukur untuk mengetahui keunggulan suatu objek.
2. Hasil dari pengolahan data sentimen dapat digunakan oleh berbagai kalangan seperti dosen, mahasiswa, lembaga survey, dan badan statistik sebagai acuan dalam mengambil keputusan.

### Daftar Pustaka

- [1] N.D. Putranti, E.Winarko, "Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan Maximum Entropy dan Support Vector Machine", *inIjccs 2014*, Vol.8, No.1, pp. 91~100, ISSN: 1978-1520, Januari 2013.
- [2] S.K. Lidya, O.S. Sitompul, S. Efendi, "Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (Svm) Dan K-Nearest Neighbor (K-Nn)", *in Sentika 2015*, ISSN: 2089-9815, Maret 28, 2015.
- [3] A.N. Hidayat, "Analisis Sentimen Terhadap Wacana Politik Pada Media Masa Online Menggunakan Algoritma Support Vector Machine Dan Naive Bayes", *inJesik 2015*, Vol.1, No.1, Januari-Juni, 2015.

- [4] D. Suhartono, *Natural Language Processing*, 2013, [Online], Available : <http://socs.binus.ac.id/2013/06/22/natural-language-processing> [Diakses : 25 Oktober 2016].
- [5] Y. Permadi, *Kategorisasi Menggunakan N-gram*, 2008, [Online], Available : [http://repository.ipb.ac.id/xmlui/bitstream/handle/123456789/1867/Pembahasan%20Permadi.%20Yudha\\_G2008-6.pdf?sequence=12](http://repository.ipb.ac.id/xmlui/bitstream/handle/123456789/1867/Pembahasan%20Permadi.%20Yudha_G2008-6.pdf?sequence=12) [Diakses : 26 Oktober 2016].
- [6] N. Christianini, J.S. Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge : Cambridge University Press, 2000.
- [7] S.A. Abdurrasyid, *Implementasi Dan Optimasi Algoritma Naziee Dan Adriani Untuk Stemming Dokumen Bahasa Indonesia*, Bandung : Universitas Telkom, 2012.
- [8] Q. Han, *CodeX: Combining an SVM Classifier and Character N-gram Language Models for Sentiment Analysis on Twitter Text*, Stuttgart : University of Stuttgart.
- [9] Stratusly, *Cloud Hosting Showdown – DigitalOcean vs AWS vs VULTR vs Linode vs Azure*, [Online], Available : <http://stratusly.com/digitalocean-vs-aws-vs-vultr-vs-linode-vs-azure/6> [Diakses : 3 November 2016].

### **Biodata Penulis**

**Rizqon Sadida**, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika STMIK AMIKOM Yogyakarta, lulus tahun 2015. Saat ini menempuh pendidikan Magister di STMIK AMIKOM Yogyakarta.

**Moh Royandi Azkia**, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika STMIK AMIKOM Yogyakarta, lulus tahun 2015. Saat ini menempuh pendidikan Magister di STMIK AMIKOM Yogyakarta.

**Beda Puspita Candra**, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika Universitas Nusantara PGRI Kediri, lulus tahun 2016. Saat ini menempuh pendidikan Magister di STMIK AMIKOM Yogyakarta.

**Novandi Rezeki**, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika STMIK AMIKOM Yogyakarta, lulus tahun 2016. Saat ini menempuh pendidikan Magister di STMIK AMIKOM Yogyakarta.

**M Ozzy Calvin Rendy**, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika STMIK AMIKOM Yogyakarta, lulus tahun 2015. Saat ini menempuh pendidikan Magister di STMIK AMIKOM Yogyakarta.