TESIS – IS185401

PENGARUH PEMBOBOTAN *SOCIAL NETWORK ANALYSIS* TERHADAP *SENTIMENT ANALYSIS* UNTUK SUBJEK ORANG

IKHWAN RUSTANTO
05211550012002

DOSEN PEMBIMBING
Nur Aini Rakhmawati, S.Kom., M.Sc Eng., Ph.D
198201202005012001

PROGRAM MAGISTER
DEPARTEMEN SISTEM INFORMASI
FAKULTAS TEKNOLOGI ELEKTRO DAN INFORMATIKA CERDAS
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
2020

THESIS – IS185401

THE IMPACT OF SOCIAL NETWORK ANALYSIS WEIGHTING ON SENTIMENT ANALYSIS FOR SUBJECT PEOPLE

IKHWAN RUSTANTO
05211550012002

SUPERVISOR
Nur Aini Rakhmawati, S.Kom., M.Sc Eng., Ph.D
198201202005012001

MASTER PROGRAM
DEPARTEMENT OF SISTEM INFORMASI
FACULTY OF INTELLEGENT ELECTRICAL AND INFORMATICS TECHNOLOGY
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
2020

(Halaman sengaja dikosongkan)

LPP/20/116

LEMBAR PENGESAHAN TESIS

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Sistem Informasi (M.Kom.)
di
Institut Teknologi Sepuluh Nopember

Oleh:
Ikhwan Rustanto
NRP: 05211550012002

Tanggal Ujian: 04 Agustus 2020
Periode Wisuda: September 2020

Disetujui oleh:
Pembimbing:

Nur Aini Rakhmawati, Ph.D.
NIP: 198201202005012001 .....

Penguji:

Ahmad Muklason, Ph.D.
NIP: 198203022009121009 .....

Faizal Mahananto, S. Kom, M.Eng., Ph.D.
NIP: 198510312019031009 .....

Surabaya, 24 Agustus 2020
Kepala Departemen Sistem Informasi
Fakultas Teknologi Elektro dan Informatika Cerdas

Dr. Mujahidin, S.T., M.T.
NIP. 197010102003121001

SURAT PERNYATAAN BEBAS PLAGIARISME

Saya yang bertandatangan di bawah ini :

Nama	: Ikhwan Rustanto
Tempat/Tanggal Lahir	: Surabaya, 23 Desember 2020
NRP	: 05211550012002
Program Studi	: S2 Sistem Informasi
Judul Tesis	: Pengaruh Pembobotan Social Network Analysis Terhadap Sentiment Analysis Untuk Subjek Orang.

Dengan penuh kesadaran saya telah memahami sebaik-baiknya dan menyatakan bahwa karya ilmiah Tesis ini bebas dari segala bentuk plagiat. Apabila dikemudian hari terbukti adanya indikasi plagiat dalam karya ilmiah ini, maka saya bersedia menerima sanksi sesuai dengan Peraturan Perundang-undangan yang berlaku

Surabaya, 14 Agustus 2020
Pembuat Pernyataan

METERAI TEMPEL
94881AFF737098251
6000
ENAM RIBU RUPIAH

Ikhwan Rustanto
05211550012002

PENGARUH PEMBOBOTAN *SOCIAL NETWORK ANALYSIS* TERHADAP *SENTIMENT ANALYSIS* UNTUK SUBJEK ORANG

Nama Mahasiswa : Ikhwan Rustanto
NRP : 05211550012002
Pembimbing : Nur Aini Rakhmawati, S.Kom, M.Sc Eng., Ph.D

ABSTRAK

Berdasarkan data dari Kementrian Komunikasi dan Informatika (Kominfo) per Februari 2018, tercatat pengguna internet di Indonesia mencapai angka 142,36 juta. Angka ini menunjukkan peningkatan sebesar 54,68% dibandingkan dengan awal tahun 2016. Peningkatan pengguna internet di Indonesia ini menunjukkan era keterbukaan informasi yang semakin melebar. Keterbukaan informasi menyebabkan informasi terhadap kinerja pemerintahan semakin mudah diperoleh oleh seluruh lapisan masyarakat.

Hal ini mendorong penelitian tentang analisa sentimen terhadap kinerja pemerintahan, dalam hal ini adalah Pemerintah Provinsi Jawa Timur (Pemprov Jatim). Penelitian yang ada tentang analisa sentimen, umumnya hanya fokus pada kalimat atau postingan yang menjadi bahan analisa sentimen. Sehingga kurang sekali memperhatikan subjek atau orang yang melakukan posting. Dengan semakin maraknya penggunaan akun palsu maupun bot, membuat kredibelitas dari pembuat opini menjadi semakin diragukan. Dalam analisa media sosial terdapat metode untuk menganalisa keberadaan akun palsu ataupun bot, salah satunya *social network analysis*.

Berdasarkan hal tersebut, penelitian ini akan mengembangkan model kombinasi *sentiment analysis* dan *social network analysis* untuk mengetahui pengaruh pembobotan s*ocial network analysis* terhadap analisa sentimen masyarakat pada Pemprov Jatim. Penelitian ini akan menggunakan nilai sentralitas dari *social network analysis* sebagai penambah bobot dalam *sentiment analysis*. Dalam hal ini, penggunaan algoritma *Naïve Bayes* akan dibandingkan dengan S*upport Vectore Machine* untuk membagi opini masyarakat berdasarkan *dataset* dari *Twitter* dan *Instagram* Pemprov Jatim dan 5 media berita online yang memberitakan Pemprov Jatim. Hasil percobaan menunjukkan hasil berkebalikan dari yang diharapkan, metode kombinasi memberikan penurunan akurasi untuk kedua algoritma.

Kata Kunci: *Sentiment analysis*, *Social network analysis*, *Naïve Bayes*, *Support Vectore Machine*

(Halaman sengaja dikosongkan)

THE IMPACT OF SOCIAL NETWORK ANALYSIS WEIGHTING ON SENTIMENT ANALYSIS FOR SUBJECT PEOPLE

By : Ikhwan Rustanto
Student Number : 05211550012002
Supervisor : Nur Aini Rakhmawati, S.Kom, M.Sc Eng., Ph.D

ABSTRACT

Based on data from the Ministry of Communication and Information (Communication and Information) as of February 2018, recorded internet users in Indonesia reached 142.36 million. This figure shows an increase of 54.68% compared to the beginning of 2016. The increase in internet users in Indonesia shows an era of information openness that is increasingly widened. Openness of information makes information on government performance more easily obtained by all levels of society.

This encourages research on sentiment analysis of government performance, in this case the Provincial Government of East Java (East Java Provincial Government). Existing research on sentiment analysis, generally only focus on the sentence or post that is the subject of sentiment analysis. So as not to pay attention to the subject or the person making the post. With the increasingly widespread use of fake accounts and bots, making the credibility of opinion makers becomes increasingly doubtful. In social media analysis there are methods to analyze the existence of fake or bot accounts, one of which is social network analysis.

Based on this, this study will develop a combination model of sentiment analysis and social network analysis to determine the effect of weighting social network analysis on community sentiment analysis in the East Java Provincial Government. This study will use the centrality of social network analysis as an added weight in sentiment analysis. In this case, the use of the Naïve Bayes algorithm will be compared with the Support Vectore Machine to share public opinion based on a dataset from Twitter and Instagram of the East Java Provincial Government and 5 online news media that preach the East Java Provincial Government. Experimental results show the opposite of the expected results, the combination method gives decreased accuracy for both algorithms.

Kata Kunci: *Sentiment analysis*, *Social network analysis*, *Naïve Bayes*, *Support Vectore Machine*

KATA PENGANTAR

Puji syukur kehadirat Allah SWT atas berkat Rahmat dan Ridho-Nya sehingga penulis dapat menyelesaikan tesis yang berjudul "PENGARUH PEMBOBOTAN SOCIAL NETWORK ANALYSIS TERHADAP SENTIMENT ANALYSIS UNTUK SUBJEK ORANG". Selama proses penyelesaian tesis ini, penulis memperoleh banyak bimbingan, bantuan, dan dukungan dari berbagai pihak. Sehingga pada kesempatan ini, penulis mengucapkan terima kasih yang sebesar-besarnya kepada

1. Kedua orang tua, Istri Fithriyah C. Ummah, Ananda Kifah R. Amany, Nada F. Amany, Salwa H. Amany dan seluruh keluarga yang telah memberikan doa, semangat, motivasi, dan senantiasa memberi dukungan kepada penulis dalam kondisi apapun.
2. Ibu Nur Aini Rakhmawati, S.Kom, M.Sc.Eng. Ph.D. selaku dosen pembimbing yang telah sabar dan telaten mendukung, membimbing serta membagikan ilmu dan waktunya kepada penulis dalam penyelesaian penelitian.
3. Bapak Ahmad Mukluson, Ph D. dan Faizal Mahananto, S.Kom., M.Eng., Ph.D. selaku dosen penguji yang telah memberikan kritik dan saran untuk perbaikan penelitian ini.
4. Seluruh bapak ibu dosen dan karyawan di Departemen Sistem Informasi, Fakultas Teknologi Elektro dan Informatika Cendas, Institut Teknologi Sepuluh Nopember.
5. Teman-teman S2 Sistem Informasi angkatan 2015 Gasal, 2015 Genap dan 2016 Gasal dan Keluarga Lab. S2 ADDI yang telah memberikan bantuan dan dukungan kepada penulis.
6. Semua pihak yang belum disebutkan, yang telah membantu dan mendukung penulisan tesis ini.

Penulis menyadari bahwa tesis ini masih jauh dari sempurna, oleh karena itu peneliti menerima kritik dan saran yang membangun untuk perbaikan di masa mendatang. Semoga penelitian ini dapat memberikan wawasan dan manfaat yang berguna bagi pengembangan ilmu pengetahuan dan pembaca

Surabaya, Agustus 2020

Ikhwan Rustanto

(Halaman sengaja dikosongkan)

DAFTAR ISI

ABSTRAK vii
ABSTRACT 9
KATA PENGANTAR 10
DAFTAR ISI 12
DAFTAR TABEL 14
DAFTAR GAMBAR 15
BAB 1 16
PENDAHULUAN 16
1.1 Latar Belakang 16
1.2 Perumusan Masalah 21
1.3 Tujuan dan Manfaat Penelitian 22
1.4 Kontribusi Penelitian 22
1.4.1 Kontribusi Teoritis 22
1.4.2 Kontribusi Praktis 22
1.5 Keterbaruan *(Novelty)* 22
1.6 Batasan Penelitian 23
1.7 Sistematika Penulisan 23
BAB 2 26
KAJIAN PUSTAKA 26
2.1 Kajian Teori 26
2.1.1 *Sentiment Analysis* 26
2.1.2 *Sosial Network Analysis* 29
2.1.3 Fake Account Detection 30
2.2 Penelitian Sebelumnya 32
2.2.1 *Sentiment Analysis* 32
2.2.2 *Social Network Analysis* 35
2.2.3 Kombinasi *Sentiment Analysis* dan *Social Network Analysis* 38
2.2.4 Fake Account Detection 38
BAB 3 40
METODOLOGI PENELITIAN 40
3.1 Perumusan Masalah 41
3.2 Kajian Pustaka 41

3.3 Pengumpulan Data 41
3.4 Pengajuan Formula Kombinasi *SNA + SA* 44
3.3.1 Preprocessing 45
3.3.2 Teknik *SA* 46
3.3.3 Teknik *SA* kombinasi *SNA* 46
3.5 Verifikasi dan Validasi 47
3.6 Kesimpulan dan Saran 48
BAB 4 50
HASIL DAN PEMBAHASAN 50
4.1 Hasil Penelitian 50
4.1.1 Pengumpulan Data 50
4.1.2 Preprocessing 52
4.1.3 Social Netwok Analysis 54
4.1.4 Sentiment Analysis 55
BAB 5 73
KESIMPULAN DAN SARAN 73
5.1 Kesimpulan 73
5.2 Saran 73
DAFTAR PUSTAKA 76

DAFTAR TABEL

Tabel 3.1. Tabel Ringkasan dataset yang digunakan 42
Tabel 3.2. Tabel contoh data twitter dan instagram 43
Tabel 4.1. Contoh data hasil pelabelan 51
Tabel 4.2. Contoh hasil penarikan data 52
Tabel 4.3. Preprocessing penghapusan URL..... 52
Tabel 4.4. Preprocessing penghapusan user tag 53
Tabel 4.5. Preprocessing penghapusan stopword 53
Tabel 4.6. Preprocessing case folding 54
Tabel 4.7. Contoh data hasil perhitungan SNA..... 54
Tabel 4.8. Daftar percobaan 55
Tabel 4.9. Hasil tokenisasi Kata 56
Tabel 4.10. Contoh vektor Kata 56
Tabel 4.11. Contoh training Data dengan label kelas 57
Tabel 4.12. Hasil akhir klasifikasi 58
Tabel 4.13. Hasil percobaan pertama dengan Naïve Bayes 58
Tabel 4.14. Hasil percobaan kedua dengan SVM 60
Tabel 4.15. Contoh Kombinasi fitur Kata dan SNA 62
Tabel 4.16. Hasil akhir klasifikasi dengan fitur SNA 63
Tabel 4.17. Hasil percobaan ketiga dengan Naïve Bayes + SNA 63
Tabel 4.18. Rangkuman Percobaan ketiga dengan variasi perbandingan data 65
Tabel 4.19. Hasil percobaan keempat dengan SVM + SNA 66
Tabel 4.20. Rangkuman Percobaan keempat dengan variasi perbandingan data 68
Tabel 4.21. Hasil Rangkuman Percobaan 69
Tabel 4.22. Hasil Rangkuman Percobaan dengan data posting 69

DAFTAR GAMBAR

Gambar 2.1. Ilustrasi kinerja *SVM* 358
Gambar 2.2. *SNA* pada komunitas beladiri 35
Gambar 2.3. Visualiasi *centrality* 36
Gambar 2.4. *Sociogram* penelitian kelas 37
Gambar 3.1. Tahapan Metodologi Penelitian 40
Gambar 3.2. Reply Instagram terhadap @jatimpemprov 444
Gambar 3.3. Alur Formula Kombinasi 455
Gambar 4.1. Confusion matrix tanpa normalisasi percobaan pertama 59
Gambar 4.2. Confusion matrix normalisasi percobaan pertama 59
Gambar 4.3. Confusion matrix normalisasi percobaan kedua 61
Gambar 4.4. Confusion matrix normalisasi percobaan kedua 61
Gambar 4.5. Confusion matrix tanpa normalisasi percobaan ketiga 64
Gambar 4.6. Confusion matrix normalisasi percobaan ketiga 64
Gambar 4.7. Confusion matrix normalisasi percobaan keempat 67
Gambar 4.8. Confusion matrix normalisasi percobaan keempat 67

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Indonesia adalah sebuah negara yang memiliki pembagian daerah administrasi yang cukup banyak di dunia. Berdasarkan data dari Kementerian Dalam Negeri (Kemendagri), Indonesia terdiri dari 34 provinsi yang dibagi menjadi 416 kabupaten dan 98 kota (otda.kemendagri.go.id). Hal ini menyebabkan pengelolaan administrasi di Indonesia menjadi sebuah tantangan tersendiri. Masing – masing daerah memiliki kewenangan untuk mengatur daerah administrasinya, sehingga pengelolaan setiap daerah menjadi unik, sesuai karakter dan infrastruktur di masing-masing daerah.

Berkembangnya teknologi informasi (TI) menjadikan setiap informasi tentang suatu daerah dapat diakses dengan mudah, kapanpun dan dimanapun. Pengguna internet di Indonesia semakin meningkat dengan semakin banyaknya media yang dapat digunakan untuk mengakses internet. Berdasarkan data dari Kementrian Komunikasi dan Informatika (Kominfo) per Februari 2018, tercatat pengguna internet di Indonesia mencapai angka 142,36 juta. Angka ini menunjukkan peningkatan sebesar 54,68% dibandingkan dengan awal tahun 2016 (kominfo.go.id).

Peningkatan pengguna internet di Indonesia menunjukkan era keterbukaan informasi yang semakin melebar. Keterbukaan informasi menyebabkan informasi terhadap kinerja pemerintahan semakin mudah diperoleh oleh seluruh warga negara melalui dunia maya. Hal ini menjadi salah satu pendorong untuk menyebarkan informasi terkait kinerja pemerintahan melalui media *online* yang ada. Selain itu, UU no 14 tahun 2008 tentang keterbukaan informasi publik mengharuskan setiap lembaga pemerintahan dan pemerintah daerah untuk memberikan informasi terkait kebijakan, rencana dan realisasi dalam kinerja pemerintahan.

Hampir seluruh pemerintah memiliki media, baik berupa media berita *online* maupun media sosial, untuk menyebarkan informasi kinerja pemerintahan. Selain untuk menyebarkan informasi, media-media tersebut dapat digunakan untuk mendapatkan umpan balik dari masyarakat. Melalui umpan balik tersebut,

pemerintah dapat mengukur tingkat kepuasan dari masyarakat terhadap kinerja pemerintahan.

Penelitian ini akan menggunakan studi kasus pada Pemerintah Provinsi Jawa Timur (Pemprov Jatim). Pemprov Jatim memiliki beberapa akun resmi untuk menampung umpan balik dari masyarakat, yaitu akun *twitter @jatimpemprov*. Terdapat juga beberapa akun *twitter* dari satuan kerja di bawahnya, seperti akun *twitter* Hubungan Masyarakat *@humasprovjatim*, akun *twitter* Badan Kepegawaian Daerah *@bkdjatim*, dll. Selain itu, Pemprov Jatim juga memiliki akun di media sosial lain, yaitu *instagram* dengan akun *@humasprovjatim dan @jatimpemprov.* Pemberitaan tentang Pemprov Jatim juga bisa didapatkan dari media *online* yang lain seperti akun *twitter* dari *@jawaposco, @sbotv, @detikcom, @protalSURYA* dan *@e100.* Dan akun *instagram* dari *@jawapos, @sbotvofficial, @detikcom, @suryaonline* dan *@esuarasurabayamedia.*

Umpan balik dari masyarakat memiliki dua arti. Pertama adalah kepuasan, yang dapat diukur dari opini positif. Kedua adalah ketidakpuasan, yang dapat diukur dari opini negatif. Opini positif ataupun negatif dapat dikategorikan dari umpan balik dan laporan yang masuk pada akun-akun tersebut, melalui teknik *sentiment analysis (SA)*. Teknik *SA* atau sebagian penelitian menyebut sebagai *opinion mining*, merupakah salah satu teknik untuk menganalisa opini, sentimen, penghargaan, emosi terhadap sebuah produk, layanan, organisasi, individu beserta atribut mereka (Shepelenko, 2017).

Penelitian tentang *SA* banyak menggunakan metode *machine learning* dalam melakukan klasifikasi sentimen dari topik yang dibahas. Algoritma yang paling banyak dibahas dengan menggunakan pendekatan *machine learning* ini adalah algoritma *Naïve Bayes (NB)*.

Nurhuda *et al* (2016) menggunakan algoritma *NB* untuk mengukur sentimen masyarakat Indonesia pada pemilihan umum presiden RI tahun 2014. Penelitian tersebut menandai *Part-of-Speech (POS)* dengan menggunakan *Hidden Markov Model POS Tagging* untuk model statistik.

Shepelenko (2017) menggunakan *NB* dan *Convolution Neural Network (CNN)* untuk menghitung sentimen berdasarkan dua *dataset* yang berbeda. Shepelenko menggunakan *dataset* dari *twitter* dan *review* film. Hasil penelitian

tersebut menunjukkan bahwa *CNN* memerlukan *resource* komputer yang besar dibandingkan *NB*.

Faradhillah *et al* (2016) melakukan *sentiment analysis* dengan menggunakan perbandingan *NB* dengan *Support Vector Machine (SVN)*. Penelitian tersebut menggunakan *dataset twitter* dengan dipecah menjadi tiga *dataset* dengan tahap *preprocessing* yang berbeda.

Penggunaan *NB* yang lain dilakukan oleh Lestari *et al* (2017) dengan *dataset twitter*. Penelitian tersebut tidak hanya menggunakan *dataset twitter* secara *tekstual*, tetapi juga *non-textual* yaitu dari pembobotan *emoji*.

Afshoh *et al* (2017) menggunakan *NB* dengan dikombinasikan dengan *chi-square* sebagai ekstraksi fitur untuk mengukur sentimen masyarakat terhadap kenaikan harga rokok.

Penelitian (Indrayuni, 2016) menggunakan algoritma *NB* dalam pendekatan *machine learning*, yang ditambahkan dengan penggunaan algoritma *Support Vector Machine (SVM)* untuk menganalisa *review* hotel. Penelitian tersebut dilakukan dengan menggunakan seleksi fitur *particle swam optimization*.

Penelitian Devi *et al* (2016) menggunakan *SVM* dengan *feature selection* dan *semantic analysis* untuk melihat sentimen terhadap film. Penelitian tersebut menggunakan *dataset review* film sebagai *input data training*. Data *review* film diberikan *score* berdasarkan kata yang telah di-*tag*. *Score* inilah kemudian menjadi *input* untuk proses *SVM* dengan menggunakan *kernel hyper parameter*.

Selain *NB* dan *SVM*, penelitian oleh Winarko (2017) menggunakan algoritma *Deep Believe Network (DBN)* sebagai *classifier*. Penelitian tersebut membandingkan *DBN* dengan *SVM* dan *NB* dengan menggunakan *dataset twitter*.

Selain pendekatan *machine learning*, pendekatan lain yang digunakan adalah pendekatan *lexicon*. Pada pendekatan ini, *sentiment analysis* dilakukan dengan orientasi kontekstual sentimen berdasarkan pada jumlah orientasi sentimen di setiap kata atau kalimat. Pendekatan ini biasanya menggunakan kamus untuk mendukung klasifikasi sentimen.

Bravo-Marquez *et al* (2016) menggunakan pendekatan *lexicon-based* dengan algoritma *Annotate-Sample-Average(ASA)* untuk *generate training* data dengan *dataset twitter*.

Kusumawati *et al* (2017) menggunakan pendekatan *lexicon based* untuk melakukan *sentiment analysis* pada *dataset twitter* untuk melihat sentimen masyarakat terhadap kebijakan kenaikan harga rokok. Penelitian tersebut menggunakan *lexicon based* dengan kamus dari *SentiWordnet* dengan ditranslasikan ke bahasa Indonesia, karena *SentiWordnet* hanya menyediakan kamus dalam bahasa Inggris.

Penelitian-penelitian yang terkait dengan teknik *SA* lebih fokus pada opini yang ada pada kalimat atau komentar, kurang mempertimbangkan hubungan antar individu ataupun antar entitas dalam dunia sosial. Selain itu teknik *SA* kurang sekali memperhatikan subjek atau orang yang melakukan *posting*. Di sisi lain, dengan semakin maraknya penggunaan *automatic program* untuk posting atau biasa disebut *bot*, membuat kredibelitas dari pembuat opini menjadi semakin berkurang. Karena pada dasarnya pengguna bot dapat menyebarkan berbagai macam informasi baik informasi palsu untuk penggiringan opini, contohnya *black campaign* (Pratama and Rakhmawati, 2019).

Teknik yang dapat digunakan untuk mengetahui ciri khas (*properties*) apakah sebuah akun merupakan akun asli atau sebuah *bot* salah satunya adalah teknik *social network analysis (SNA)*. Dengan mengetahui *properties* dari sebuah akun, akan diketahui kredibelitas dari akun tersebut. Karena akun yang memiliki sedikit jaringan ataupun pengikut (*follower*) merupakan salah satu ciri khas dari akun *bot*. Teknik ini juga dapat digunakan untuk menggambarkan model hubungan antara individu dengan individu yang lain, melalui relasi di media sosial mereka.

Penelitian *SNA* berkaitan dengan menghitung dan menggambarkan hubungan antara individu, akun, organisasi, maupun entitas yang dapat digambarkan dalam bentuk graf yang disebut *sociogram*. Dalam *sociogram*, satu akun diwakili oleh *node* yang terhubung dengan *node* lain. Hubungan antara satu *node* dengan *node* lain digambarkan sebagai sebuah *edge*. Semakin banyak hubungan yang dimiliki, maka semakin banyak *edge* yang dimiliki. Penelitian *SNA* menggunakan teorema *graf* untuk menggambarkan jejaring sosial antar *node*, termasuk dalam menganalisa sentralitas dari jaringan.

Penelitian dengan *SNA* untuk memetakan komunitas *online* dilakukan oleh Yuliana *et al* (2015) dengan menggunakan *dataset facebook group*. Penelitian

tersebut menggunakan *microsoft NodeXL* untuk memetakan data pertemanan dalam grup komunitas pecinta beladiri Indonesia. Penelitian tersebut juga menggunakan metode *Fruchterman Reingold* untuk memodelkan hubungan antar *node*-nya.

Penelitian *SNA* yang lain adalah penelitian tentang interaksi antara aktif dalam penyebaran informasi dalam kegiatan *Jazz Goto Campus (JGTC)* oleh Oktora dan Alamsyah (2014). Penelitian tersebut menggunakan *dataset* dari *Application Programming Interface (API) twitter*, kemudian menggunakan *tools Gephi* untuk menghitung dan menganalisa *tweet* yang berhubungan dengan JGTC. Analisa pada penelitian tersebut mendapatkan akun-akun yang paling berpengaruh maupun menjadi *trigger* untuk penyebaran informasi kegiatan tersebut. Dengan menggunakan analisa sentralitas, didapatkan juga akun-akun yang paling aktif dalam penyebaran informasi, maupun akun yang paling mudah menyebarkan informasi.

Penelitian oleh Grunspan *et al* (2014) menggunakan *twitter* dan juga data survei sebagai *dataset* dalam penelitian. Penelitian tersebut disusun untuk memberikan gambaran model interaksi antar siswa dan hubungannya dalam meningkatkan nilai akhir ujian.

Penelitian-penelitian pada pembahasan sebelumnya cenderung hanya berfokus pada satu pendekatan saja. Pendekatan gabungan antara *SA* dan *SNA* pernah dilakukan oleh Bermingham *et al* (2009) dengan menggunakan *dataset* dari *group youtube*. Penelitian tersebut menggunakan pendekatan *leksikal-based* dan *SNA* untuk menggambarkan potensi radikalisme. *SNA* pada penelitian tersebut digunakan untuk mengidentifikasi perbedaan penyebaran topik radikalisme pada gender tertentu.

Penelitian lain yang menggabungkan kedua analisa yaitu oleh Gryc dan Moilanen (2014). Penelitian tersebut menggunakan *hybrid machine learning* dan *logic-based framework* sebagai *classifier*. Penelitian tersebut menggunakan *tools Weka* untuk mengekstrak *feature* dari *SNA* dengan *dataset posting blog* yang berhubungan dengan Barack Obama pada masa pemilihan presiden Amerika Serikat.

Fornacciari *et al*. (2015) pada penelitiannya juga mengkombinasikan *SNA* dan *SA* dengan *dataset twitter*. Berkebalikan dalam penelitian yang lain, penelitian

tersebut menggunakan asosiasi sentimen sebagai *node* dari *graf* yang menjadi model koneksi *SNA*. Penelitian tersebut menggunakan algoritma *Multinomial Naïve Bayes.*

Dari uraian di atas, penelitian tentang *SA* banyak berfokus pada aspek algoritma, *feature* maupun *dataset* yang digunakan dalam penelitian, dengan mengabaikan informasi terkait subjek akun pemilik opini maupun jaringan akun pemilik opini.

Untuk itulah, penelitian untuk mengetahui pengaruh pembobotan SNA terhadap SA terkait subjek orang perlu dilakukan dalam rangka untuk mengetahui bagaimana teknik *SNA* terkait subjek orang dapat mempengaruhi pengukuran opini masyarakat terhadap Pemprov Jatim.

1.2 Perumusan Masalah

Berdasarkan uraian yang telah dilakukan, beberapa poin yang menjadi perhatian adalah: (1) Perlu sebuah teknik kombinasi yang dapat mengukur pengaruh pembobotan *SNA* dari subjek orang pada *SA* terhadap Pemprov Jatim, (2) Bagaimana teknik gabungan *SA* dan *SNA* dapat meningkatkan akurasi dalam analisa opini terhadap Pemprov Jatim.

Oleh sebab itu, pertanyaan utama yang menjadi motivasi peneliti untuk mengadakan penelitian ini adalah:

- Bagaimana pengaruh pembobotan teknik *SNA* terhadap teknik *SA*?
- Bagaimana teknik *SNA* dapat meningkatkan akurasi dari teknik *SA* dalam analisa sentiment pada akun *twitter* dan *instagram* Pemprov Jatim dan 5 media *online* lain yang juga memberitakan Pemprov Jatim?

Untuk menjawab pertanyaan tersebut, diusulkan beberapa strategi penelitian sebagai berikut:

a. Menentukan teknik *SNA* dan *SA* yang akan digunakan untuk menganalisa akun *twitter* dan *instagram* Pemprov Jatim dan 5 media *online* lain.
b. Membuat instrumen penilaian performa untuk menguji algoritma.

c. Menguji masing-masing algoritma untuk mengukur performa.
d. Menarik kesimpulan dan alur perkembangan penelitian berikutnya.

1.3 Tujuan dan Manfaat Penelitian

Berdasarkan perumusan masalah di atas, maka tujuan penelitian tesis ini adalah untuk menghasilkan teknik kombinasi antara *SNA* dan *SA* untuk mengukur pengaruh pembobotan *SNA* terhadap *SA* terhadap subjek orang pada akun media sosial Pemprov Jatim dan 5 media *online* lain.

Manfaat penelitian ini adalah untuk mendapatkan hasil pengukuran sentimen yang lebih akurat, dengan cara memasukkan hubungan antara pemilik opini, melalui kombinasi dengan teknik *SNA.*

1.4 Kontribusi Penelitian

Kontribusi penelitian ini dapat dibedakan menjadi kontribusi teoritis dan kontribusi praktis.

1.4.1 Kontribusi Teoritis

Kontribusi teoritis penelitian ini terletak pada pengembangan metode kombinasi antara teknik *SNA* dan *SA* dalam pengukuran sentimen dari dua atau lebih *dataset* yang berbeda.

1.4.2 Kontribusi Praktis

Kontribusi praktis penelitian ini yaitu (1) Meningkatkan kinerja proses *SA* pada analisa sentiment masyarakat terhadap pemprov Jatim dengan cara menambahkan teknik *SNA* pada pembobotan *SA.* (2) Memberikan gambaran opini masyarakat Jawa Timur terhadap pemprov Jatim melalui media sosial.

1.5 Keterbaruan *(Novelty)*

Selama ini, penelitian untuk mengetahui sentimen masyarakat hanya dilakukan dengan teknik *SA.* Penelitian ini akan memberikan keterbaruan dengan mengetahui pengaruh pembobotan *SNA* terhadap subjek orang dapat meningkatkan akurasi dari *SA.*

1.6 Batasan Penelitian

Penelitian ini memiliki ruang lingkup yang menjadi batasan penelitian. Batasan penelitian ini adalah:

1. Sumber yang digunakan adalah data akun *twitter* dan *instagram* dari Pemprov Jatim dan 5 media media *online* lain yang melakukan pemberitaan tentang Pemprov Jatim. Sumber tersebut telah di*crawling* mulai bulan Oktober 2018.
2. Data jejaring sosial yang digunakan adalah berdasarkan indeks keterhubungan pada akun *twitter* dan *instagram* dari Pemprov Jatim dan 5 media *online* lain.

1.7 Sistematika Penulisan

Sistematika penulisan laporan penelitian yang dilakukan adalah sebagai berikut:

Bab I: Pendahuluan

Bab ini terdiri dari latar belakang penelitian, perumusan masalah, tujuan, manfaat, batasan penelitian, kontribusi penelitian dan sistematika penulisan.

Bab II: Kajian Pustaka

Bab ini berisi tinjauan pustaka dan penelitian-penelitian yang sudah ada mengenai *sentimen analysis* dan *social network analysis*.

Bab III: Metodologi Penelitian

Bab ini mengulas tentang sistematika tahapan-tahapan yang digunakan untuk melakukan penelitian.

Bab IV: Kerangka Konseptual dan Hipotesis Penelitian

Bab ini membahas konsep dan konstruk model yang akan dikaji dalam penelitian ini. Hipotesis dan instrumen penelitian disajikan dalam bab ini.

Bab V: Analisis dan Hasil Penelitian

Bab ini berisi analisis deskriptif maupun statistik dari hasil penelitian. Pada bab ini juga diuji hipotesis yang telah ditentukan pada bab IV.

Bab VI: Kesimpulan dan Saran

Bab ini terdiri dari kesimpulan penelitian dan saran untuk penelitian selanjutnya.

Daftar Pustaka

Berisi daftar referensi yang digunakan dalam penelitian ini, baik jurnal, buku maupun berita terkait teknik *SA* dan teknik *SNA*.

(Halaman ini sengaja dikosongkan)

BAB 2
KAJIAN PUSTAKA

2.1 Kajian Teori

2.1.1 *Sentiment Analysis*

Sentiment Analysis (SA) adalah salah satu cabang dari *opinion mining (OM)*. Sedangkan *OM,* adalah salah satu sub dari *computational linguistic* yang mengambil fokus pada ekstraksi dari opini yang ada di web (Osimo and Mureddu, 2012). *SA* lebih fokus pada keberpihakan *(polarity)* dan bobot keberpihakan *(weight of polarity)* dari sebuah opini. Sebuah opini dari teks yang ada akan ditentukan apakah bersifat positif, negatif atau netral terhadap topik yang dibahas. Opini teks disebut *binary classification* jika hanya memiliki dua opini, yaitu positif dan negatif. Opini teks yang memiliki lebih dari dua opini dapat disebut dengan *multi-class classification* (Shepelenko, 2017). Dalam beberapa penelitian, *SA* disamakan dengan *OM,* karena *expression* yang muncul antara *SA* dan *OM* saling menggantikan, dan memiliki pengertian yang sama (Medhat *et al*., 2014).

Berdasarkan asal ekstraksi opini, *SA* dapat dikategorikan menjadi tiga level (Medhat *et al*., 2014; Shepelenko, 2017), yaitu:

- *Document level*: pada level ini opini diekstrak dari satu dokumen utuh. Satu dokumen dianggap hanya mewakili satu opini.
- *Sentence level*: pada level ini opini diekstrak dari satu kalimat. Satu kalimat dianggap adalah dokumen pendek yang bisa bermakna subjektif maupun obyektif. Kalimat subjektif dianggap sebagai opini.
- *Aspect level*: pada level ini, opini diekstrak berdasarkan aspek entitas.

Dalam survei penelitian terhadap teknik-teknik *SA,* Medhat *et al* (2014) mengkategorikan *SA* menjadi dua pendekatan. Yaitu pendekatan *machine learning* dan pendekatan *lexicon-based*. Penelitian ini akan menggunakan pendekatan *machine learning*.

SA dengan pendekatan *machine learning* diawali dengan melakukan *training* terhadap algoritma *machine learning* dengan menggunakan *dataset* awal sebelum digunakan untuk menguji data yang sebenarnya. *Dataset* awal ini sudah memiliki

label opini untuk masing-masing *input*, sehingga akan menjadi *predefined data training* untuk algoritma *machine learning*. Setelah *training* terhadap algoritma selesai, maka algoritma tersebut dapat digunakan untuk melakukan analisa terhadap *dataset* uji atau data sebenarnya.

Berdasarkan penelitian Medhat *et al*. (2014), metode *supervised learning* banyak digunakan dalam penelitian *SA.* Sehingga pembahasan pada penelitian ini akan difokuskan pada *machine learning* dengan metode *supervised learning*. Metode *supervised learning* bergantung pada jumlah dokumen *training* yang sudah diberi label. Metode *supervised learning* memiliki banyak model *classifier*. Namun, ada dua *classifier* yang paling banyak digunakan dalam penelitian *SA,* yaitu *probabilistic classifier* dan *linear classifier*. Dalam *probabilistic classifier,* ada beberapa algoritma yang sering digunakan antara lain *Naïve Bayes, Bayesian Network* dan *Maximum Entropy*. Sedangkan dalam *linear classifier* terdapat *Support Vector Machine* dan *Neural Network*. Detail untuk setiap algoritma akan dibahas pada pembahasan selanjutnya.

2.1.1.1 *Naïve Bayes classifier*

Naïve Bayes classifier (NB) adalah algoritma paling sederhana dan paling banyak digunakan. Algoritma ini bekerja dengan menghitung *probability class* yang muncul berdasarkan distribusi kata dalam sebuah dokumen. Model ini bekerja dengan menggunakan *Bag of Word (BOW) feature extraction* yang tidak memperhitungkan posisi kata dalam sebuah dokumen. Model ini menggunakan teorema Bayes dalam melakukan prediksi terhadap sebuah set *feature* yang telah diberikan label.

$$P(label \mid features) = \frac{P(label) * P(features \mid label)}{P(features)}$$

$P(label)$ adalah *prior probability* dari label atau probabilitas hipotesis, $P(feature|label)$ adalah *prior probability* pada *feature set* yang telah diberikan label, $P(features)$ adalah *prior probability* yang muncul pada *feature set* yang diberikan. Rumus ini menjelaskan bahwa peluang masuknya *sample* label tertentu pada kelas *posterior* (yang akan muncul), adalah peluang munculnya kelas *prior* dikali kemungkinan munculnya karakteristik *sample*

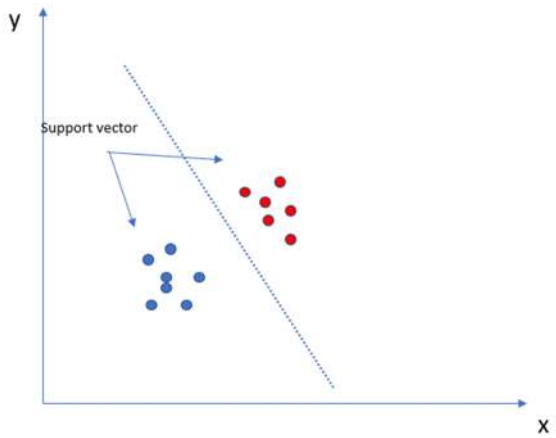
pada kelas yang diberikan label (*likehood*) dibagi peluang munculnya karakteristik tersebut secara global (*evidence*).

2.1.1.2 *Bayesian Network*

Bayesian Network (BN) yang disebut juga dengan *Belief Network* atau *Probabilistic Network,* adalah algoritma yang menggambarkan interaksi antar variabel dengan model grafik. *BN* menggambarkan hubungan antara variabel dengan bentuk *graf*, dimana variabel (dilambangkan X) dan probabilitasnya (dilambangkan $P(X)$) digambarkan sebagai *node*. Jika variabel X memiliki pengaruh terhadap variabel Y, maka akan digambarkan *node* x terhubung dengan *node* y. Dan ini dapat dinyatakan dengan $P(Y|X)$.

2.1.1.3 *Support Vector Machine*

Support Vector Machine (SVM) merupakan algoritma yang tujuan utamanya adalah untuk menentukan pemisah linear yang dapat memisahkan kelas yang berbeda dengan ruang antara kelas yang terbaik. SVM berada dalam satu kelas *Artificial Neural Network (ANN)* yaitu berada di dalam kelas *supervised learning* dalam pendekatan *machine learning.* Pada metode SVM, data akan dibagi menjadi beberapa kelas berdasarkan pembagi linear dengan nilai masing-masing fitur sebagai koordinat pada ruang n-dimensional. Tujuan algoritma ini adalah menemukan *hyper-plane* yang membagi kelas dengan akurasi yang baik. Kinerja SVM dapat digambarkan pada gambar 2.1.



Gambar 2.1. Ilustrasi kinerja SVM

2.1.2 *Sosial Network Analysis*

Social Network Analysis (SNA) adalah salah satu teknik untuk mengukur dan memetakan hubungan antara individu, organisasi, akun maupun entitas dalam memproses sebuah informasi melalui media sosial.

SNA memanfaatkan interaksi dalam media sosial dan menggambarkannya dalam bentuk *graf*, dimana *nodes* melambangkan individu, atau organisasi atau entitas yang terhubung satu dengan lainnya melalui interaksi pertemanan (relasi), *follow*, *comment* maupun sekedar *retweet*. Hubungan antar individu ini dapat digambarkan sebagai *edge* pada teorema *graf*.

Dalam jejaring sosial, sebuah *node* atau individu tidak hanya terhubung ke satu *node* lain. Tetapi bisa terhubung dengan beberapa *node* yang lain, sehingga akan membentuk jejaring sosial yang lain. Setiap individu bisa jadi memiliki banyak jejaring untuk mereka sendiri. Dalam *SNA,* dilakukan pemetaan dan pengukuran terhadap alur interaksi sehingga terbentuk pola alur interaksi dari sebuah *node* pada jejaring sosial mereka.

SNA memiliki fokus pada hubungan antara *node*, dan cara hubungan itu terjadi. Selain itu, *SNA* juga untuk mengetahui apakah hubungan terjadi dua arah, melalui media apa, melalui jejaring yang mana, siapa yang memiliki hubungan paling banyak, siapa yang terisolasi dalam jejaring tersebut, bagaimana *gap* dan *length* antara masing-masing *nodes*, dimana terjadi *bottleneck*, siapa yang menjadi *keyplayer* dalam sebuah jejaring, dan masih banyak hal lain terkait hubungan antara *nodes*.

Dalam *SNA,* ada beberapa pengukuran yang melibatkan hubungan antara individu atau *node*. Pengukuran ini digambarkan sebagai sentralitas individu. Pembahasan pada penelitian ini akan difokuskan pada tiga sentralitas,

1. *Degree centrality*, adalah jumlah koneksi yang dimiliki sebuah *node* pada seluruh jejaring yang dimilikinya.

2. *Closeness centrality*, adalah jarak antara *node* satu dengan yang lain pada seluruh jejaring, ukuran ini menggambarkan kedekatan antar *node*. Semakin dekat *node*, semakin kuat *node* tersebut dalam mempengaruhi *node* yang terhubung.
3. *Betweeness centrality*, adalah peran sebuah *node* dalam menjadi *bottleneck*. Artinya, semakin banyak jalan yang melewati persimpangan pada salah satu *node*, maka semakin penting arti dari persimpangan tersebut. Sehingga ketika terjadi perubahan, atau penghambatan informasi, dapat dilakukan analisa yang lebih dalam pada persimpangan tersebut.

Selain terminologi di atas, terdapat beberapa terminologi perhitungan lain terkait dengan *SNA*, antara lain:

1. *Eigenvector Centrality*, adalah versi *rekursif* dari *degree centrality* karena memberikan bobot lebih tinggi pada *node* yang terhubung pada *node* lain yang memiliki keterhubungan yang tinggi.
2. *PageRank*, adalah pengukuran yang biasa ditemui pada teknik *Search Engine Optimization (SEO)* untuk menghitung nilai dari suatu *page*. *PageRank* dapat digunakan sebagai jaringan *graph* yang memiliki arah. Prinsip dalam perhitungan ini adalah semakin banyak suatu *node* di*referer* oleh *node* lain, maka nilai *node* tersebut semakin penting.
3. *Clustering Coefficient*, adalah perhitungan *node* dalam proporsi keterhubungan antara *node* dalam kelompok. Semakin kuat keterhubungan antar *node* dalam satu kelompok maka semakin besar nilai *Clustering Coefficient*.

2.1.3 Fake Account Detection

Dunia media sosial yang semakin berkembang menyisakan sebuah permasalahan baru. Identitas yang dimiliki sebuah akun media social, belum tentu mencerimakan pemilik asli dari akun tersebut. Satu orang bahkan bisa jadi memiliki banyak akun dengan tujuan dan fungsi masing-masing. Dalam perkembanganya masing-masing platform media sosial memiliki tools untuk mencegah munculnya akun-akun palsu tersebut. Linkedin contohnya memilki beberapa level pencegahan, dari mulai *registration scoring*, *clustering*, *activity*

based scoring hingga *manual review* ("Automated Fake Account Detection at LinkedIn," n.d.).

Akun palsu dalam dunia sosial media bisa dibuat dan dijalankan oleh manusia, mesin/komputer, maupun *cyborg* (Van Der Walt and Eloff, 2018). Akun palsu dibuat manusia secara manual untuk tujuan marketing, penggiringan opini maupun untuk tujuan lainnya. Mesin / komputer dapat membuat dan menjalankan akun palsu sesuai dengan program yang dimasukkan ke dalam mesin tersebut, akun palsu yang dibuat oleh mesin biasa disebut *bot*. Sedangkan *cyborg* adalah perpaduan manusia dan mesin, dimana manusia menjalankan mesin untuk otomatisasi posting, otomatisasi rating dan sebagainya.

Deteksi akun palsu dapat menggunakan beberapa pendekatan dengan kelebihan dan kekurangannya, beberapa yang sering digunakan antara lain filtering, rules dan machine learning (Van Der Walt and Eloff, 2018).

2.1.3.1 Filtering

Filtering adalah teknik yang paling reaktif. *Filtering* digunakan untuk membedakan platform media sosial untuk membedakan antara akun palsu dengan akun asli. Ketika sebuah akun mengirimkan pesan yang memiliki karakteristik spam, maka akun tersebut akan di blacklist. Metode yang digunakan seperti dalam metode pemilihan spam pada email. Namun metode spam filtering menjadi lebih rumit dengan adanya *spammer* yang dapat menggunakan strategi pengiriman yang dinamis. Manusia dengan mudah dapat menghindari filtering ini, ketika satu akun di blacklist maka akan langsung membuat akun baru.

2.1.3.2 Rules

Rules digunakan untuk identifikasi akun palsu dengan cara mendeteksi beberapa pola yang dimiliki oleh akun palsu. Seperti contoh *rules* berbasis kata. Ketika sebuah akun memposting konten yang mengandung kata (misal MLM) sejumlah tertentu, maka akun tersebut akan dikategorikan sebagai akun palsu. Sebuah rules perlu terus diperbaharui, karena banyak kata baru yang muncul.

2.1.3.3 Machine learning

Pendekatan *Machine Learning* (*ML*) adalah pendekatan yang umum digunakan untuk mendeteksi akun palsu. Algoritma *supervised machine learning* perlu dataset berisi fitur yang memiliki label klasifikasi per fiturnya. fitur yang dimasukkan sebagai inputan akan digunakan sebagai inputan dalam proses pembelajaran model untuk memprediksi *outcome*. fitur yang digunakan dalam *ML* didapatkan dengan memgambil atribut dari akun yang dipelajari. Contoh fitur yang digunakan seperti jumlah teman, jumlah follower jumlah status dll.

2.2 Penelitian Sebelumnya

Pada pembahasan ini akan dijabarkan beberapa penelitian sebelumnya terkait *SA, SNA,* dan kombinasi keduanya.

2.2.1 *Sentiment Analysis*

Dalam penelitian *Sentiment Analysis (SA)*, beberapa penelitian mencoba untuk membandingkan algoritma *SA* dengan menggunakan beberapa pendekatan. Penggunaan *Naïve Bayes* dan *Support Vector Machine* paling banyak mendominasi penelitian yang ada. Paparan selanjutnya akan membahas beberapa penelitian yang dibagi berdasarkan algoritma dan *dataset* yang digunakan.

2.2.1.1 Algoritma *Naïve Bayes (NB)*

Algoritma ini banyak digunakan karena paling sederhana dan dipandang akurat. Pembahasan tentang algoritma *Naïve Bayes* berkaitan dengan berapa *dataset* yang digunakan, maupun membandingkan *Naïve Bayes* dengan algoritma *machine learning* yang lain, serta *feature extraction* dalam algoritma *Naïve Bayes.*

Nurhuda *et al*. (2016) melakukan penelitian *SA* dengan *Naïve Bayes*, menggunakan *dataset twitter* untuk mengukur sentimen masyarakat Indonesia terhadap calon presiden Republik Indonesia 2014. Nurhuda menandai *Part-of-Speech (POS)* dengan menggunakan alat *Hidden Markov Model (HMM) POS Tagging*. *HMM POS* adalah model statistik untuk menghitung probabilitas dari kejadian yang tidak diamati berdasarkan kejadian yang dapat diamati.

Shepelenko (2017) dalam penelitiannya menggunakan *Naïve Bayes (NB)* dan *Convolution Neural Network (CNN)* untuk menghitung sentimen dari *tweet* dan *review* film. Dalam penelitiannya, Shelepenko mengemukakan bahwa *CNN* lebih membutuhkan dan menghabiskan *resource* komputasi yang besar daripada *NB*, meskipun hasilnya lebih baik daripada *NB*.

Beberapa penelitian lain juga membandingkan antara algoritma *Naïve Bayes* dengan algoritma lain seperti yang dilakukan oleh Shepelenko. Faradhillah *et al*. (2016) melakukan penelitian pengklasifikasian opini masyarakat terhadap Pemerintah Kota Surabaya dengan menggunakan *Naïve Bayes* yang dibandingkan dengan *Support Vector Machine.* Penelitian tersebut menggunakan *dataset twitter* yang kemudian divisualisasikan dengan *R Shiny* menjadi sebuah *wordcloud* dan grafik *streamgraph*.

2.2.1.2 *Naïve Bayes* dengan *Chi-Square Feature Extraction*

Penelitian *SA* dengan menggunakan *Naïve Bayes* juga dilakukan oleh Afshoh *et al*. (2017) dengan menggunakan *dataset twitter* dan *chi-square* sebagai *feature extraction*. Dalam penelelitian ini, analisa sentiment digunakan untuk melihat melihat persepsi masyarakat terhadap wacana kenaikan harga rokok.

2.2.1.3 *Naïve Bayes* dengan *dataset textual* dan *non-textual*

Kombinasi penggunaan *dataset* ketika menggunakan *Naïve Bayes* juga dilakukan dalam penelitian *SA*. Penelitian ini dilakukan oleh Lestari *et al*. (2017) dalam mengukur sentimen terhadap opini Pemilihan Kepala Daerah di DKI Jakarta. Penelitian tersebut menambahkan pembobotan *emoji* sebagai *dataset non-textual* dalam pengujian dan *training* algoritma *Naïve Bayes*.

2.2.1.4 Algoritma *Support Vector Machine*

Penelitian yang menggunakan *Support Vector Machine (SVM)* antara lain dilakukan oleh Indrayuni (2016) untuk menganalisa sentimen *review* hotel. Penelitian ini menggunakan *SVM* dengan berbasis *Particle Swam Optimization* sebagai *tools* untuk *feature selection*.

Penelitian Faradhillah *et al* (2016) menggunakan *SVM* yang dibandingkan dengan *Naïve Bayes* untuk menganalisa tingkat kepuasan masyarakat terhadap Pemerintah Kota Surabaya.

Penelitian lain dilakukan oleh Devi *et al*. (2016) yang menggunakan *SVM* dengan berbasis *Feature Selection* dan *Semantic Analysis*. Penelitian ini menggunakan *dataset review* film yang tidak terstruktur, yang diubah menjadi bentuk yang lebih terstruktur, dan kemudian diberikan *score* berdasarkan kata yang telah di-*tag*. *Score* inilah yang menjadi *input* untuk proses *SVM* dengan *kernel hyper parameter.* Dalam penelitiannya, Devi menggunakan *SVM* sebagai algoritma utama yang dibandingkan dengan *Naïve Bayes*. Penelitian tersebut menggunakan *Gaussian Radial Basis Function (RBF) kernel* dari *SVM* yang dimodifikasi berdasarkan *hyper parameter*.

2.2.1.5 Algoritma *Deep Belief Network*

Penelitian yang menggunakan algoritma *Deep Belief Network* (DBN) dilakukan oleh Winarko (2017) dengan melakukan analisa terhadap *dataset twitter*. Penelitian tersebut membandingkan penggunakan algoritma *DBN* dengan algoritma *SVM* dan *Naïve Bayes* untuk melakukan analisa terhadap *tweet* dari *dataset twitter*. *DBN* merupakan salah satu metode machine learning yang masuk dalam turunan *Deep Learning*.

2.2.1.6 *The Annotate-Sample-Average (ASA) Algorithm*

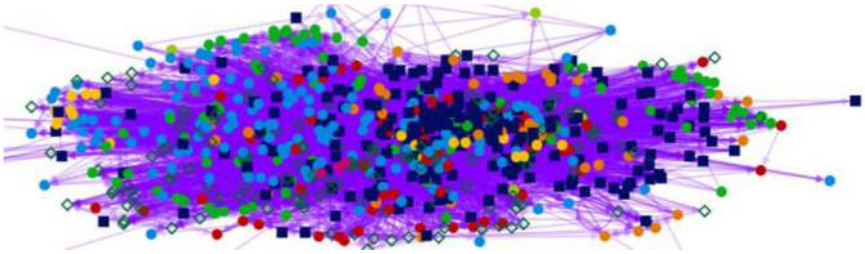
Berbeda dengan penelitian yang dibahas pada pembahasan sebelumnya, penelitian *sentiment analysis* Bravo-Marquez (2017) tidak menggunakan pendekatan *machine learning*. Penelitian tersebut mengajukan algoritma *Annotate-Sample-Average (ASA)* untuk *generate training data* dalam mendapatkan *polarity data twitter*. Metode ini menggunakan dua data *input*: 1) Data *corpus,* dan 2) Data opini *lexicon*. Data *corpus* didapatkan dari koleksi data *twitter* yang belum memiliki label. Data diambil melalui *API twitter* dengan bahasa dan topik yang sama dengan data opini *lexicon*. Data opini *lexicon* dapat didapatkan secara manual (penilaian manusia) maupun automatis (melalui *semantic network*, sinonim, antonim atau statsitik kata).

2.2.2 *Social Network Analysis*

Penelitian tentang *Social Network Analysis (SNA)* sangat bervariasi. Tidak hanya didominasi dengan penelitian yang berhubungan dengan media sosial, tetapi berhubungan juga dengan aspek dan disiplin ilmu yang lain.

2.2.2.1 *SNA* dengan *NodeXL* dan *dataset Facebook Group*

Penelitian dari Yuliana *et al*. (2015) menggunakan *SNA* untuk memetakan komunitas *online*. Penelitian tersebut menggunakan *Microsoft NodeXL* untuk memetakan komunitas *online* dengan *dataset Facebook Group*. Data diambil melalui *API Facebook* dan dipetakan menjadi 11 klaster yang memiliki tokoh sentral di masing-masing klasternya. Penelitian tersebut menggunakan algoritma *Clause-Newman-Moore* untuk membagi sentral dari jejaring sosial komunitas pecinta beladiri Indonesia. Penelitian tersebut menggunakan metode *Fruchterman Reingold* dalam aplikasi *NodeXL* untuk memodelkan hubungan antara *node*. Dari graf yang terbentuk (pada gambar 2.2), terlihat *node* yang memiliki koneksi berupa aktifitas maupun pertemanan yang lebih besar akan memiliki verteks yang lebih tebal. Selain itu, terlihat *nodes* yang memiliki bentuk yang berbeda-beda, yang mewakili klaster masing-masing.

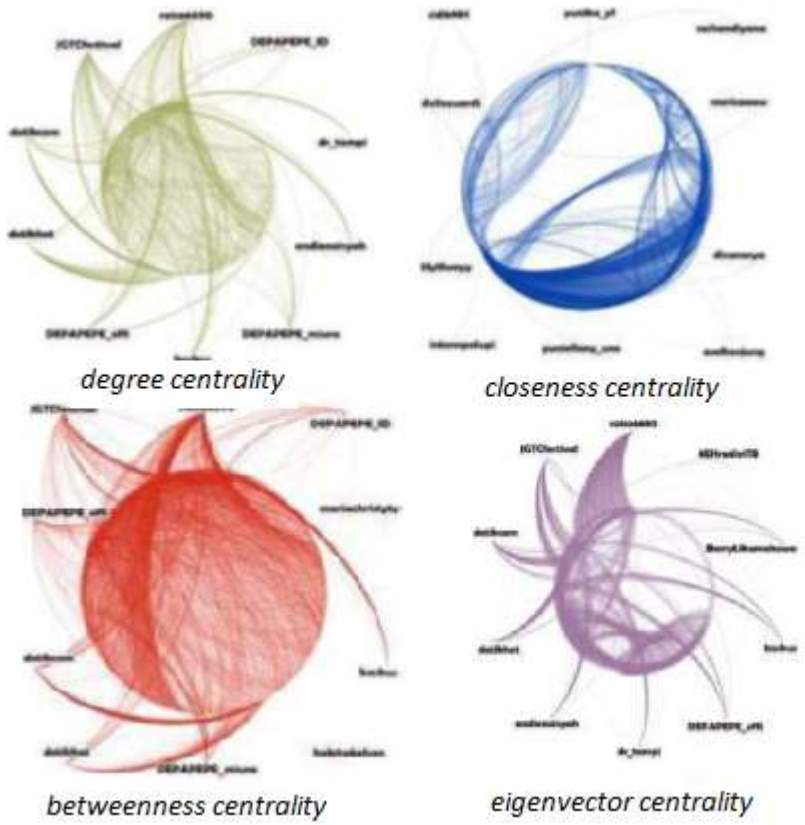


Gambar 2.2. *SNA* pada komunitas beladiri

2.2.2.2 *SNA* dengan *Gephi* dan *dataset Twitter*

Penelitian Oktora dan Alamsyah (2014) menggunakan *SNA* untuk melihat pola interaksi dan aktor yang paling berperan dalam sebuah kegiatan *Jazz Go To Campus (JGTC)* 2013. Dengan memanfaatkan *dataset API Twitter*, penelitian tersebut menggunakan aplikasi *Gephi* untuk menghitung dan menganalisa *tweet* yang berhubungan dengan *tag JGTC*. Dari 7642 *node* (akun) yang terlibat, dengan 7445 *edge* (interaksi), didapatkan aktor yang paling

berpengaruh adalah akun dari penyanyi terkenal yang juga merupakan bintang tamu dalam kegiatan tersebut.



Gambar 2.3. Visualiasi *centrality*

Gambar 2.3 menggambarkan visualisasi *centrality* yang didapatkan dalam penelitian tersebut, sehingga didapatkan hasil analisa sebagai berikut:

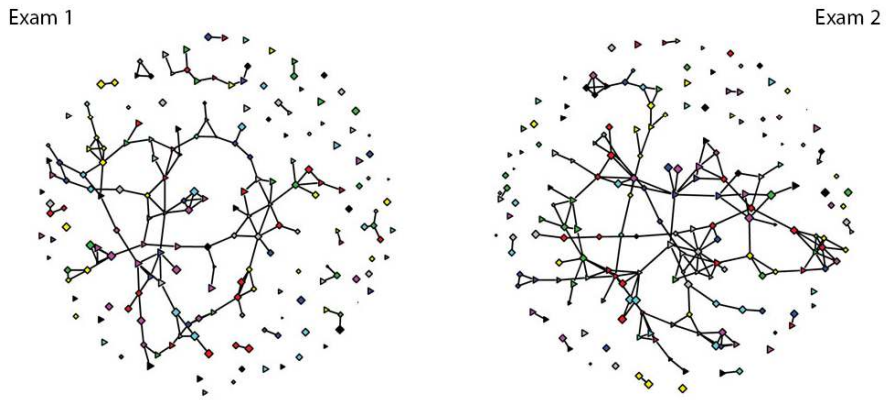
1. Berdasarkan *degree centrality*, ditunjukkan bahwa *node* yang memiliki peran paling besar dalam penyebaran informasi dalam penelitian tersebut adalah *node* yang memiliki *edge* yang paling besar. Dalam hal ini *node* tersebut adalah artis bintang tamu dalam kegiatan tersebut.
2. *Closeness centrality* digunakan untuk *node* yang memiliki jangkauan *node* yang pendek yang artinya *node* tersebut lebih cepat dalam menyebarkan informasi. Dalam hal ini, semakin mendekati nilai 1 atau sama dengan 1 maka semakin besar *closeness node* tersebut. Dari hasil analisa *centrality* tersebut, didapatkan selisih antara peringkat 1 dan peringkat 2 *closeness*

centrality adalah 0,024. Hal ini menyebabkan *node* dengan *closeness centrality* 0,857 lebih mudah untuk berhubungan dengan *node* lain saat menggunakan *mention, retweet* maupun *reply.*

3. Analisa *betweeness centrality* dapat digunakan untuk mengetahui posisi dari *node* yang menjadi jembatan dalam *network* tersebut. Dalam hal ini, akun bintang tamu dalam kegiatan tersebut sangat berperan menjadi jembatan antar *node* dengan nilai *betweeness* 0,124.
4. Analisa *eigenvector centrality* memiliki fungsi untuk menemukan *node* yang memiliki performa paling baik dalam *network*. Penelitian tersebut mendapatkan nilai *eigenvector centrality* paling tinggi (yaitu nilai 1) dimiliki oleh akun bintang tamu kegiatan tersebut, yang diikuti oleh akun resmi kegiatan tersebut. Hal ini menunjukkan bahwa akun bintang tamu tersebut memiliki interaksi yang cukup aktif dan tinggi dengan *node* lain.

2.2.2.3 *SNA* dengan *dataset* survei kelas

Grunspan *et al*. (2014) meneliti tentang penggunaan *SNA* dalam memahami pembelajaran kelas. Penelitian tersebut menggunakan data yang disusun dari kelas biologi dan memberikan gambaran model interaksi antar siswa, serta hubungan antar interaksi tersebut dengan *outcome* yang didapat dalam pembelajaran kelas. Data yang digunakan dalam penelitian tersebut adalah data survey. Survei dilakukan pada seluruh kelas untuk saling memberikan penilaian antara siswa terkait pertemanan pra-kelas dan tentang siswa pintar di kelas.



Gambar 2.4. *Sociogram* penelitian kelas

Sociogram pada gambar 2.4 menggambarkan *network* pada ujian pertama dan ujian kedua. Siswa laki-laki digambarkan sebagai segitiga dan siswa perempuan berbentuk permata. Warna pada masing-masing *node* menggambarkan lab yang mereka ambil. *Edge* pada masing-masing *node* menggambarkan relasi atau hubungan pembelajaran pada ujian pertama dan ujian kedua.

2.2.3 Kombinasi *Sentiment Analysis* dan *Social Network Analysis*

Penelitian kombinasi *Sentiment Analysis (SA)* dan *Social Network Analysis (SNA)* telah dilakukan oleh beberapa penelitian sebelumnya. Bermingham *et al* (2009) menggunakan *dataset* dari *youtube group* dan komentar dari video untuk analisa potensi radikalisme *online*. Penelitian tersebut menggunakan analisa *leksikal* terhadap *corpus* yang telah di-*crawling*. *SNA* pada penelitian tersebut digunakan untuk mengidentifikasi perbedaan penyebaran sebuah topik terkait radikalisme pada gender tertentu.

Gryc and Moilanen (2014) dalam penelitian *SA* mereka, menggunakan *corpus* 2,8 juta *blog post* dari 16.741 *blogger* selama April 2008 sampai dengan Mei 2009. Penelitian tersebut menganalisa sentimen terhadap presiden Amerika Barack Obama dengan *modelling blogsphere* terhadap akun terkait Obama. Pendekatan yang dilakukan dalam penelitian tersebut adalah *hybrid machine learning* dan *logic-based framework* yang bekerja dengan tiga *distinct* level untuk analisanya. Yaitu dengan *Shallow Document Clasification, Deep Entity-Level sentiment scoring* dan *Social Network Modelling*. Penelitian tersebut menggunakan *Naïve Bayes Multinomial (NBM)* dan *Logistic Regression Model* dengan *tool WEKA*.

Fornacciari *et al*. (2015) pada penelitiannya mengkombinasikan *SNA* dan *SA* dengan menggunakan *dataset twitter*. Penelitian tersebut mengasosiasikan sentimen sebagai *nodes* dari *graf* yang menjadi model koneksi *SNA. Classifier* yang digunakan dalam penelitian tersebut berdasarkan pada algoritma *Multinomial Naïve Bayes*.

2.2.4 Fake Account Detection

Penelitian terkait *fake account detection* banyak menggunakan algortima *Machine Learning* (*ML*).

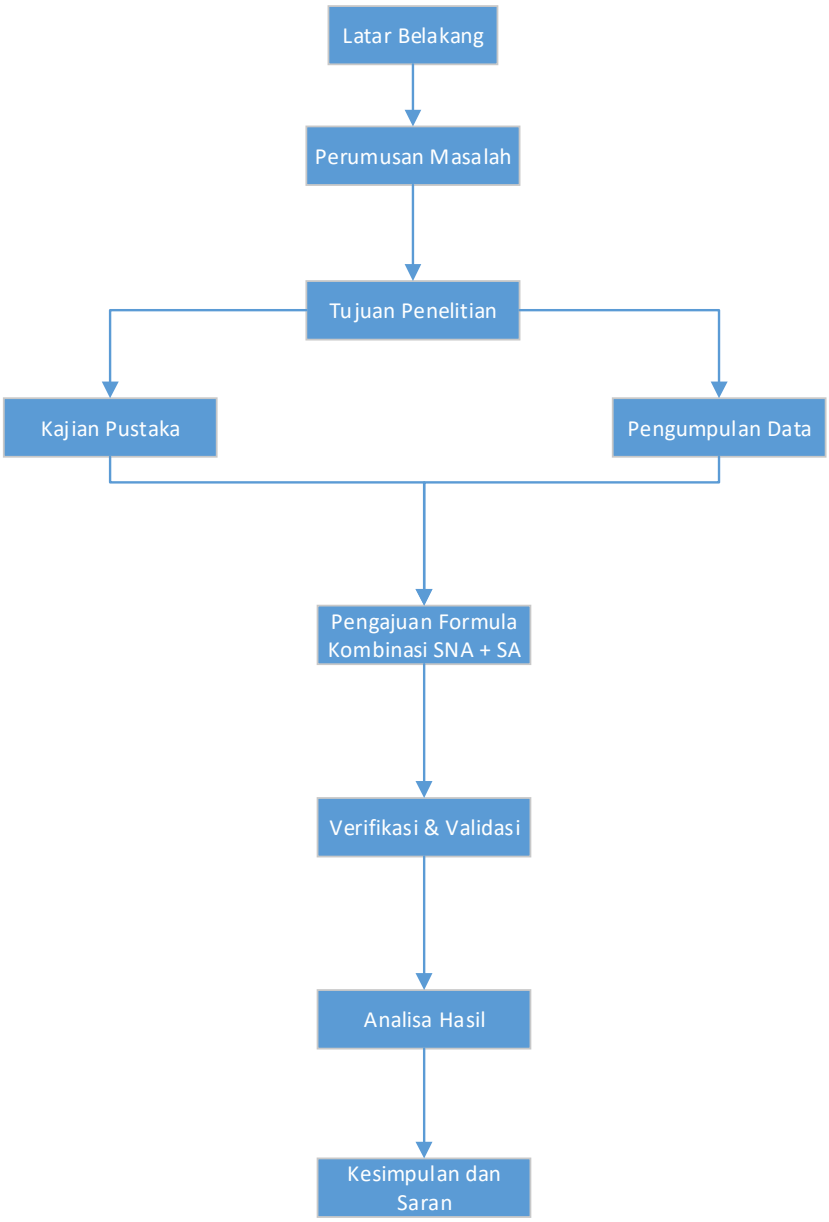
Penelitian oleh Pratama and Rakhmawati (2019) menggunakan algoritma *Support Vector Machine (SVM)* dan *Random Forest* untuk identifikasi *bot*. Data yang digunakan adalah data *twitter* dari pendukung kandidat presiden Indonesia pada pilpres 2019. Hasil yang didapatkan menunjukkan tahun pembuatan akun memiliki kontribusi paling tinggi pada identifikasi *bot*.

Van Der Walt and Eloff (2018) menggunakan *SVM machine learning* untuk identifikasi *fake identities*. Van Der Walt meneruskan penelitian yang sebelumnya yaitu identifikasi *fake account* dengan menggunakan atribut yang ditemukan terkait akun tersebut pada *twitter*. Pada penelitian ini Van Der Walt menggunakan *machine learning* dengan *engineered feature* (misal rasio *friend-to-follow*, hasil perhitungan dari atribut *friend-count* dan *follower-count*) untuk mendeteksi *fake account*. Penelitian ini menghasilkan *engineered fitur* yang digunakan untuk mendeteksi *bot* tidak efektif untuk mendeteksi *fake account* yang dibuat oleh manusia. Hasil yang terbaik didapatkan berupa F1 *score* 49.75%. penelitian ini menunjukan perilaku yang dilakukan manusia untuk membuat *fake account* berbeda dengan *bot* yang membuat *fake account*, sehingga seharusnya memiliki model yang berbeda.

El Azab et al., (2016) melakukan penelitian untuk identifikasi *fake account* pada *twitter*. Pada penelitian diawali dengan menentukan sejumlah set faktor utama yang memberikan pengaruh terhadap identifikasi *fake account* di *twitter*, yang kemudian faktor yang ditemukan diterapkan menggunakan teknik klasifikasi yang berbeda. Fitur yang digunakan didapat dari berbagai penelitian yang lain dan sudah difilter dengan *extensive analysis* dan diberikan pembobotan. Dari 22 atribut yang digunakan, hanya 7 atribut yang efektif untuk identifikasi akun palsu. Teknik klasifikasi yang digunakan pada penelitian ini antara lain *Random Fores*, *Decision Tree*, *Naive Bayes*, *Neural Network* dan *SVM.*

BAB 3
METODOLOGI PENELITIAN

Pada Bab ini dijelaskan metodologi dari penelitian yang dilakukan. Metode penelitian yang dilakukan mengacu pada penelitian kuantitatif. Tahapan penelitian yang akan dilakukan ditunjukkan pada gambar 3.1.



Gambar 3.1. Tahapan Metodologi Penelitian

3.1 Perumusan Masalah

Tahapan perumusan masalah bertujuan untuk mencari *research question* yang akan dipecahkan melalui penelitian ini. Identifikasi masalah telah dibahas pada pembahasan sebelumnya. Berdasarkan pembahasan sebelumnya, tujuan yang ingin dicapai dalam penelitian ini adalah untuk mencari solusi dalam memperbesar akurasi dari teknik *Sentiment Analysis* dengan cara menggabungkan dengan teknik *Social Network Analysis.*

3.2 Kajian Pustaka

Pada tahapan ini dilakukan studi literatur pada buku, teks, jurnal dan artikel ilmiah yang mendukung tujuan penelitian. Teknik yang dilakukan adalah *snowbowling*, yaitu dengan melakukan penelusuran terhadap studi pustaka dari penelitian yang berkaitan dengan *Sentiment Analysis* dan *Social Network Analysis*. Kajian pustaka yang dilakukan meliputi teknik-teknik dalam *Social Network Analysis* dan *Sentiment Analysis*. Kajian lain yang juga dilakukan adalah terkait dengan uji coba *performance* yang perlu dilakukan untuk menguji kombinasi kedua teknik tersebut.

3.3 Pengumpulan Data

Pada tahapan pengumpulan data, dilakukan pengambilan data *twitter* dan *instagram* Pemprov Jatim, juga *twitter* dan *instagram* media *online* yang berhubungan dengan Pemprov Jatim. Detail *dataset* yang digunakan antara lain: *dataset twitter* yang diambil adalah *tweet mention, hashtag, retweet dan reply* dari *twitter* Pemprov Jatim *@jatimpemprov* dan *@humasprovjatim*. Untuk melengkapi *dataset* dari akun Pemprov Jatim, digunakan data *twitter* dari akun media *online* yang melakukan pemberitaan terkait Pemprov Jatim. Data dikumpulkan dari akun yang menggunakan *hashtag #PemprovJatim, #JatimPemprov, #JawaTimur* dan *#Jatim*, maupun yang melakukan *mention* terhadap akun *@jatimpemprov* dan *@humasprovjatim.* Untuk akun media lain yang dimonitor antara lain: *@jawaposco, @sbotv, @detikcom, @protalSURYA* dan *@e100.*

Selain itu, juga dilakukan *crawling* data komentar *instagram* terhadap akun *instagram @jatimpemprov* dan *@humasprovjatim*, serta terhadap 5 media lain yang memberitakan Pemprov Jatim. Senada dengan penanda di *twitter*, pemberitaan yang

berkaitan dengan Pemprov Jatim ditandai dengan adanya *tag #PemprovJatim, #JatimPemprov ,#JawaTimur* dan *#Jatim* atau melakukan *mention* terhadap akun *@jatimpemprov* dan *@humasprovjatim.* Untuk akun *instagram,* media yang dimonitor antara lain *@jawapos, @sbotvofficial, @detikcom, @suryaonline* dan *@esuarasurabayamedia.* Ringkasan *dataset* yang digunakan dapat dilihat pada tabel 3.1.

Tabel 3.1. Tabel Ringkasan dataset yang digunakan

Sumber	Akun	Keterangan
Twitter	*@jatimpemprov, @humasprovjatim*	*mention, hashtag, retweet dan reply*
Instagram	*@jatimpemprov, @humasprovjatim*	*mention, hashtag*
Twitter	*@jawaposco, @sbotv, @detikcom, @protalSURYA* dan *@e100*	*mention, hashtag #PemprovJatim, #JatimPemprov #Jatim #JawaTimur*
Instagram	*@jawapos,@sbotvofficial, @detikcom, @suryaonline* dan *@esuarasurabayamedia*	*mention, hashtag #PemprovJatim, #JatimPemprov #Jatim #JawaTimur*

Contoh data yang berasal dari *twitter* dan *instagram* bisa dilihat pada tabel di bawah 3.2.

Tabel 3.2. Tabel contoh data twitter dan instagram

Sumber	Akun	Keterangan
Twitter	*@jatimpemprov, @humasprovjatim*	Replying to @argesh_suta Hampir tiap ahir tahun. Terima kasih pak Karwo cs dah bantu mengurangi beban rakyat. @GusIpul_ID @JatimPemprov Walau jauh di Yamaguchi University Jepang sini, saya tetap haturkan Selamat HUT ke 73 Prov. Jawa Timur kita. Semoga lebih Jaya luar biasa menuju Jatim sebagai Smart Province dan Smart Governance. @e100ss @KhofifahIP @EmilDardak

		@JatimPemprov @humasprovjatim @jatim
Instagram	*@jatimpemprov, @humasprovjatim*	goenawanwst : Menjadikan jatim propinsi yg berdampak positif bagi Indonesia #prayforlombok #lombokberduka #DukaMerekaDukaKita #gempalombok #jatim #pemprovjatim
Twitter	*@jawaposco, @sbotv, @detikcom, @protalSURYA* dan *@e100*	Selama sepuluh tahun terakhir, pemerintahan Jawa Timur terbukti bersih, transparan, dan efektif kinerjanya. @GusIpul_ID @puti_soekarno @JatimPemprov @tribunjatim @detikcom #WargaPKBJatimCoblosGusIpul #JatimCoblosKudungAbang2 #27JuniCoblosGusIpulPuti #DebatPilgubJatim #JawaTimur
Instagram	*@jawapos,@sbotvofficial, @detikcom, @suryaonline* dan *@esuarasurabayamedia*	duwilestari94@noviarrah setuju bgtt 👍 makin naik umk sebakin bnyak pengangguran #umkjatim #pemprovjatim #jawatimur #pemerintahprovinsijawatimur #jatim #infojatim #infoumk #infojawatimur #kabarjawatimur #kabarjatim #jawatimurkeren #jawatimurhits

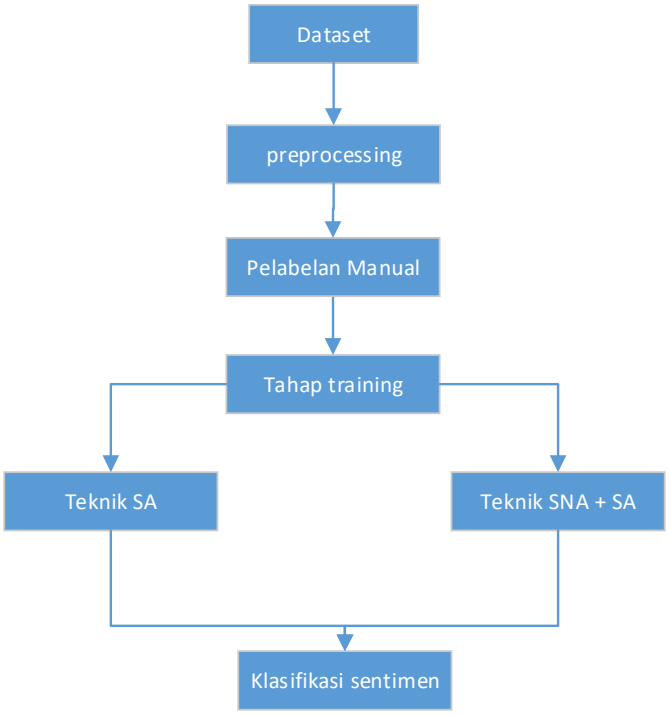
Contoh data instagram dari @jatimpemprov dan komentar yang memiliki sentiment dapat dilihat pada gambar 3.2.

Gambar 3.2. Reply Instagram terhadap @jatimpemprov

Pelaksanaan dilakukan dengan server yang berada di lab ADDI Jurusan Sistem Informasi ITS. Data yang didapatkan nantinya digunakan sebagai *dataset* untuk melakukan teknik dalam *Social Network Analysis* dan *Sentiment Analysis*. Dalam proses training *dataset* akan dipecah menjadi 3 *dataset* yaitu, (1) *dataset* twitter dan instagram pemprov Jawa Timur dan humas pemprov Jawa Timur, (2) *dataset* dari media berita terkait pemprov Jawa Timur,(3) *dataset* gabungan keduanya.

3.4 Pengajuan Formula Kombinasi *SNA + SA*

Pada tahapan ini dibuatlah desain system berupa alur dari data masuk hingga data ditraining dan divalidasi dengan *Social Network Analysis* dan *Sentiment Analysis.* Data awal (*dataset*) yang berasal dari twitter akan dilakukan proses *preprocessing* yang akan dijabarkan lebih detail, lalu *dataset* yang sudah dibersihkan melalui *preprocessing* akan dilakukan training dengan dua skenario. Untuk skema pembuatan sistem dapat dilihat pada gambar 3.3



Gambar 3.3. Alur Formula Kombinasi

3.3.1 Preprocessing

Pada tahap preprocessing ini, *dataset* twitter dan komentar instagram yang masuk akan dilakukan tahap pembersihan atau proses *preprocessing*. Dataset dalam penelitian ini akan melalui *preprocessing* untuk masing-masing metode. Proses *preprosesing* yang dilakukan untuk dataset *Sentiment Analysis* antara lain :

1. Menghapus *URL* dan *tag html*, adalah proses penghapusan *URL* dan *tag html* yang mungkin diikutkan dalam *tweet* atau komen instagram. Karena *URL* dan *tag html* tidak mengandung informasi yang bermanfaat untuk analisa.
2. Menghapus User akun, adalah proses penghapusan akun lain yang dimention oleh pembuat posting atau komentar. Biasanya diawali tanda '@' diikuti *username*. Missal @jatimprov. Karena kata ini kurang memberikan informasi yang dibutuhkan untuk analisa.

3. Menghapus *Retweet* dan data yang duplikat, adalah proses penghapusan data yang sudah pernah di*tweet* oleh user lain yang dicopy dan disimbolkan dengan tulisan RT. Hal ini untuk mengurangi penambahan *weight* dari *tweet* yang dari user yang berbeda tetapi sebenarnya memiliki isi yang sama.
4. Menghapus *stopword*, adalah proses untuk menghapus kata-kata yang tidak memiliki makna yang sering muncul dalam kalimat (misal yang, dari, dsb)
5. *Case Folding*, adalah proses menjadikan teks menjadi satu *case*, misal menjadi *lowercase* atau *uppercase* semua.

Sedangkan dalam teknik *Social Network Analysis*, *dataset* dari *twitter* dan *instagram* akan melalui tahap *preprocessing* antara lain :

1. Menghapus *tweet* dan posting *instagram* yang tidak berhubungan dengan Pemprov Jatim atau Provinsi Jatim.
2. *Case folding*, proses untuk mengubah teks *tweet* dan *posting instagram* menjadi salah satu *case*, *lowercase* atau *uppercase*
3. Menghapus *vertex* yang duplikat, karena relasi antara node dilihat berdasarkan *follow* dan *retweet,* maka jika satu node memiliki lebih dari satu vertex, vertex dengan node yang sama akan dihapus dari node tersebut.

3.3.2 Teknik *SA*

Teknik *SA* yang digunakan pada penelitian kali ini adalah dengan pendekatan *Naïve Bayes* dibandingkan dengan *Support Vector Machine*. Setiap algoritma akan ditraining dan diuji dengan menggunakan 4 dataset yang berbeda yang sudah dijelaskan pada pembahasan sebelumnya. Rasio perbandingan antara data training dengan data uji adalah 70:30 untuk masing-masing algoritma.

3.3.3 Teknik *SA* kombinasi *SNA*

Pada kombinasi teknis *SA* dengan *SNA* ini, *SNA* digunakan untuk menambah bobot dari fitur yang akan diproses melalui *SA*. Semakin besar *degree*

centrality dari sebuah node, maka opini yang dimilikinya akan memberikan dampak yang besar untuk *node* lain yang terhubung dengan *node* tersebut.

3.5 Verifikasi dan Validasi

Verifikasi dan validasi dilakukan untuk menguji apakah masing-masing teknik memberikan output sesuai dengan yang diharapkan untuk proses berikutnya. Output dari teknik *SA* secara mandiri akan dibandingkan dengan output dari *SNA* dikombinasikan dengan *SA,* dengan kombinasi *dataset* yang sama sehingga akan diketahui tingkat validasi dari algoritma. Penilaian untuk masing-masing teknik berdasarkan perhitungan *Accuracy, Precision, Recall* dan *F-Score.* Dengan variable *True Positive (TP), True Negative (TN), False Pofitive (FP),* dan *False Negative (FN)*, Perhitungan performa dengan formula

- Accuracy (A) = (TP+TN)/(TP+TN+FP+FN)
- Precision (P) = TP/(TP+FP)
- Recall (R) = TP/(TP+FN)
- F-Scoro = 2*(P*R)/(P+R)

Pola hubungan interaksi yang dibangun untuk teknik *SNA* dalam *dataset* diambil berdasarkan hubungan *mention* dan *follow.* hal ini digunakan baik di twitter maupun di instagram. Sedangkan khusus di *twitter* akan dilihat pola hubungan *re-tweet* yang sudah dijabarkan pada pembahasan sebelumnya. Untuk memvalidasi *opinion leader* dari masing-masing sentimen, dilakukan pengujian menggunakan *Spearmansch correlation coefficient* yang dapat digunakan untuk menentukan pentingnya relasi antara user dengan jejaring sosial yang dimiliki oleh user.

Selain pengujian akurasi dkk, dilakukan pengujian *sampling* secara manual. Data *twitter* dan *instagram* yang telah diklasifikasi oleh teknik gabungan *SNA* dan *SA*, akan dicek secara manual apakah sentimen yang dinilai sudah sesuai dengan konteks dari posting *twitter* dan *instagram.* Dari pengujian ini akan diketahui tingkat keberhasilan model dalam melakukan klasifikasi sentiment data *twitter* dan *instagram.*

3.6 Kesimpulan dan Saran

Bagian ini menjelaskan hasil kesimpulan dari analisis pada tahap ujicoba dan analisa hasil. Kesimpulan dari bagian ini untuk merangkum jawaban dari permasalahan dan tujuan yang ditetapkan di awal. Juga untuk memberikan saran penelitian yang dapat dikembangkan dari penelitian ini.

(Halaman ini sengaja dikosongkan)

BAB 4
HASIL DAN PEMBAHASAN

4.1 Hasil Penelitian

Pada bagian ini akan dijelaskan mengenai hasil yang diperoleh dari penelitian yang dilakukan untuk melihat pengaruh pembobotan *Social Network Analysis* terhadap *Sentiment Analysis* untuk subyek orang dengan menggunakan *Naïve Bayes* dibandingkan dengan *SVM*.

4.1.1 Pengumpulan Data

Langkah awal yang dilakukan pada penelitian adalah dengan melakukan akuisisi data dari sumber data media sosial dan media online. Dari media sosial data dicrawling dari *twitter* dan *instagram* Pemprov Jatim, juga *twitter* dan *instagram* media *online* yang berhubungan dengan Pemprov Jatim. Pencarian dilakukan dengan keyword yang dijelaskan pada metodologi penelitian.

Dari pengumpulan data sejak tanggal 23 September 2018 hingga 3 Maret 2020 diperoleh data sebanyak 393.265 data. Pengambilan data menggunakan *framework scrapy* dan output yang dihasilkan berupa data di *MongoDb*. Masing-masing data memiliki struktur:

- post_owner : merupakan akun pemilik posting / data
- content : merupakan isi dari *tweet* / *posting* / konten berita
- post_url : *url* dari posting
- date : merupakan tanggal posting
- crawled_date : merupakan tanggal *crawling*
- spider_source : merupakan sumber *crawling*, apakah dari *twitter* atau *Instagram*

Setelah data diperoleh dilakukan proses pelabelan secara manual untuk menentukan apakah data termasuk dalam kelas positif, netral atau negatif. Proses pelabelan data dilakukan secara manual. data yang diasumsikan memiliki maksud pujian, penghargaan, motivasi dan rasa terima kasih akan ditandai sebagai sentimen positif

Sedangkan data yang menunjukkan bencana, keluhan dan ketidaksetujuan akan ditandai sebagai sentiment negativePelabelan dilakukan untuk sebagai data awal training dan untuk validasi data pada akhir penelitian. Untuk penelitian kali ini, hanya akan digunakan 9.900 data yang terbagi menjadi 3.300 data positif, 3.300 data netral, dan 3.300 data negatif. Contoh data yang sudah diberikan label dapat dilihat pada tabel 4.1.

Tabel 4.3. Contoh data hasil pelabelan

#	Source	Text	Sentiment
1	Twitter	Malang banjir, mohon untuk menyelesaikan masalah banjir @KemensosRI @jokowi @KemenPU @OmbudsmanRI137 @DPR_RI @PDemokrat @Pak_JK @JatimPemprov @PemkotMalang @PemkabMalang	Negatif
2	Twitter	Min update terkini banjir yg melanda madiun, ngawi & magetan dong	Negatif
3	Twitter	Selamat berkarya melayani masyarakat SDA prov jatim , semoga banjir segera redah	Positif
4	Twitter	Gubernur Jawa Timur, Khofifah Indar Parawansa menghimbau seluruh masyarakat mewaspadai potensi fenomena hidrometeorologi hingga pertengahan Maret. Hidrometeorologi dimaksud meliputi banjir, rob, banjir bandang, tanah longsor, dan angin puting beliung. @KhofifahIP @bpbd_jatim	Positif

Selanjutkan dilakukan pengumpulan data yang terkait dengan Social Network Analysis. Data yang diambil adalah atribut dari akun pemilik posting. Data yang diambil dari *Twitter* berupa data *follower* dan *friend count*. Sedangkan untuk Instagram data yang diambil berupa data *follower* dan *following*. Contoh data yang didapatkan dapat dilihat pada tabel 4.2

Tabel 4.2. Contoh hasil penarikan data

#	Source	Username	follower	following
1	Twitter	@Detikcom	16.072.382	29
2		@RTMCJatim	215.282	69
3		@JatimPemprov	57.431	291
4	Instagram	nurmajdi	875	840
5		hanastyr	855	311
6		FadilLating	832	754

selain data *follower* dan *following*. Pengumpulan data yang dilakukan dengan framework lint juga menyimpan data post *tweet, likes* dan *join_date*. Sedangkan untuk data *Instagram* ditarik dengan menggunakan *python* dengan tambahan data post.

4.1.2 Preprocessing

Pada tahap ini, data yang telah dikumpulkan akan dibersihkan. Proses yang dilakukan adalah menghapus dan mentransformasi kata yang dianggap memilki makna ambigu atau tidak memiliki arti. Beberapa proses yang dilakukan :

1. Menghapus *URL* dan *tag html*, adalah proses penghapusan *URL* dan *tag html* yang mungkin diikutkan dalam *tweet* atau komen instagram. Contoh perubahan dapat dilihat pada tabel 4.3

Tabel 4.3. Preprocessing penghapusan URL

1	Sebelum	Dewan & @JatimPemprov Siapkan Jatim Hadapi Industri 4.0 **http://www. surabayapagi.com/read/180845/2018/12/10/dewan--pemprov-siapkan-jatim-hadapi-industri-40.html**
	Sesudah	Dewan & @JatimPemprov Siapkan Jatim Hadapi Industri 4.0
2	Sebelum	@JatimPemprov kembali meraih penghargaan di bidang pelayanan publik. **http://p3tki-jatim.go.id/portal/content/news/gubernur-jatim-terima-penghargaan-pemprov-terinovatif-iga-2018-dalam-pelayanan-masyarakat**

	Sesudah	@JatimPemprov kembali meraih penghargaan di bidang pelayanan publik

2. Menghapus User akun, adalah proses penghapusan akun lain yang dimention oleh pembuat posting atau komentar. Biasanya diawali tanda '@' diikuti *username*.

Tabel 4.4. Preprocessing penghapusan user tag

1	Sebelum	Dewan & **@**JatimPemprov Siapkan Jatim Hadapi Industri 4.0
	Sesudah	Dewan & JatimPemprov Siapkan Jatim Hadapi Industri 4.0
2	Sebelum	**@**JatimPemprov kembali meraih penghargaan di bidang pelayanan publik
	Sesudah	JatimPemprov kembali meraih penghargaan di bidang pelayanan publik

3. Menghapus *Retweet* dan data yang duplikat, adalah proses penghapusan data yang sudah pernah di*tweet* oleh user lain yang dicopy dan disimbolkan dengan tulisan RT. Hal ini untuk mengurangi penambahan *weight* dari *tweet* yang dari user yang berbeda tetapi sebenarnya memiliki isi yang sama.
4. Menghapus *stopword*, adalah proses untuk menghapus kata-kata yang tidak memiliki makna yang sering muncul dalam kalimat (misal yang, dari, dsb).

Tabel 4.5. Preprocessing penghapusan stopword

1	Sebelum	Dewan **&** JatimPemprov Siapkan Jatim Hadapi Industri 4.0
	Sesudah	dewan jatimpemprov siapkan jatim hadapi industri 4.0
2	Sebelum	JatimPemprov **kembali** meraih penghargaan **di** bidang pelayanan publik
	Sesudah	jatimpemprov meraih penghargaan bidang pelayanan public

5. *Case Folding*, adalah proses menjadikan teks menjadi satu *case*, misal menjadi *lowercase* atau *uppercase* semua.

Tabel 4.6. Preprocessing case folding

1	Sebelum	Dewan & JatimPemprov Siapkan Jatim Hadapi Industri 4.0
	Sesudah	dewan & jatimpemprov siapkan jatim hadapi industri 4.0
2	Sebelum	JatimPemprov meraih penghargaan bidang pelayanan publik
	Sesudah	jatimpemprov meraih penghargaan bidang pelayanan publik

Untuk *SNA* dilakukan penghapusan data untuk akun yang tidak memiliki keterkaitan dengan *Sentiment Analysis*. Seperti contoh data iklan, informasi cuaca dan akun yang hanya berisi link.

4.1.3 Social Netwok Analysis

Pada proses Social Network Analysis, data akun yang sebelumnya dilakukan proses pengumpulan data *follower*, *friend* dan *following* akan dilakukan perhitungan *follower* dibandingkan dengan total *follower* + *following*. Matriks perhitungan ini disebut *reputation* yang merupakah salah satu fitur yang digunakan Tiwari and Velayutham (2017) dalam penelitian *bot detection*. Analisa ini dilakukan dengan mengukur koneksi dari sebuah node dalam sebuah jejaring. Dalam hal ini node diwakili oleh akun pada jejaring social. Dan matriks *reputation* dapat dihitung dari *jumlah follower / (jumlah follower + jumlah following)*. contoh hasil perhitungan *SNA* dapat dilihat pada tabel 4.7.

Tabel 4.7. Contoh data hasil perhitungan SNA

#	Source	Username	Follower rate
1	Twitter	@ akuanakayamm	0.00278
2		@ JeffKieff	0.00290
3		@JatimPemprov	0.99495
4	Instagram	AmbrosiusAdvent	0.571428
5		dipendajatim	1
6		Astidhaniar	0.58870

4.1.4 Sentiment Analysis

Pada proses sentiment analysis, dilakukan klasifikasi terhadap data yang telah dilakukan preprocessing. Akan dilakukan empat percobaan yang menggunakan metode yang berbeda dan penggunaan bobot *SNA* seperti dirangkum dalam tabel 4.8. Percobaan pertama menggunakan metode *Naïve Bayes* dengan fitur vector kata dari dataset posting Twitter dan Instagram. Percobaan kedua menggunakan metode *Support Vector Machine (SVM)* dengan fitur yang sama yaitu vector kata dengan dataset yang sama juga. Percobaan ketiga menggunakan metode Naïve Bayes dengan fitur vector kata dari data posting Twitter dan Instagram dengan tambahan fitur berasal dari *Social Network Analysis (SNA)* sebagai pembobotan dalam *Sentiment Analysis*. Sedangkan untuk percobaan keempat sama dengan percobaan ketiga yang menambahkan pembobotan *SNA* untuk fitur datanya, tetapi menggunakan metode *Support Vector Machine* dalam proses Analisa sentimennya.

Tabel 4.8. Daftar percobaan

Klasifikasi	Metode	Kombinasi	Dataset
Percobaan pertama	Naïve Bayes	SA	Twitter + Instagram
Percobaan kedua	SVM	SA	Twitter + Instagram
Percobaan ketiga	Naïve Bayes	SNA + SA	Twitter + Instagram
Percobaan keempat	SVM	SNA + SA	Twitter + Instagram

Total data yang digunakan dalam penelitian sebanyak 3.300 data netral, 3.300 data positif dan 3.300 data negatif. Data berasal dari posting *Twitter* dan *Instagram* yang sebelumnya telah dilakukan proses *preprocessing*. Dari dataset tersebut dibagi menjadi dua bagian, yaitu 70% digunakan untuk data training classifier dan 30% untuk data testing classifier. Klasifikasi *Naïve Bayes* dan *SVM* dilakukan dengan menggunakan module sklearn dari Bahasa *Python*. Dan fungsi *train_test_split* digunakan untuk membagi besar data training dan testing. Untuk ekstraksi fitur dari dataset, digunakan fungsi *feature_extraction* dari *sklearn*. Fungsi ini akan

mengekstrak fitur vector kata dari data posting yang dimasukkan. Secara detail fungsi *feature_extraction* bekerja melalui tahap:

1) *Tokenisasi* kata: proses pembuatan kamus kata yang ada dari seluruh data training. Kamus kata pada tahap ini adalah daftar seluruh kata yang ada pada model, termasuk data training. Pada kamus ini seluruh data tweet akan dipecah menjadi daftar kata unik dengan informasi frekuensi kata yang muncul dari seluruh data. Data yang dihasilkan dapat dilihat pada tabel 4.9.

Tabel 4.9. Hasil Tokenisasi Kata

Kata	Frekuensi
jawatimurpemprov	3724
meraih	212
penghargaan	1782
bidang	1341
pelayanan	265
publik	587

2) Membuat fitur vector kata dari kamus kata: Pada tahap ini, kata yang telah dilakukan *tokensisasi*, dijadikan menjadi vektor kata sesuai dengan index kata pada daftar dataset, hasil vektor kata dapat dilihat pada tabel 4.10.

Tabel 4.10. Contoh Vektor Kata

Kata	Frekuensi	Vektor Kata
jawatimurpemprov	3724	(1,1) 3724
meraih	212	(1,2) 212
penghargaan	1782	(1,3) 1782
bidang	1341	(1,4) 1341
pelayanan	265	(1,5) 265
publik	587	(1,6) 587

Seluruh tahap yang dijelaskan pada tabel 4.9 dan 4.10 dilakukan dalam ekstraksi fitur yang dimiliki *sklearn* dengan menggunakan fungsi

CountVectorizer. Setelah fitur diekstraksi, dilakukan proses splitting data. Data dipecah menjadi 70% data training dan 30% data testing dengan menggunakan fungsi *sklearn* yaitu *data_train_test_split*. Pada data split yang ada, vektor kata yang sebelumnya telah diberikan label sesuai dengan kelas yang telah diberikan pada data training. Contoh data training yang telah diberikan label seperti pada tabel 4.11. Data yang telah dipecah dilakukan proses training yang dilakukan pada masing-masing percobaan.

Tabel 4.11. Contoh Training Data dengan label kelas

Dokumen	W1	W2	W3	W4	sentimen
Tweet1	1	1	2	1	Pos
Tweet2	3	1	6	1	Pos
Tweet3	1	1	3	3	Neg
Tweet4	8	7	1	5	Pos
….	…..	…..	…..	…..	…..

a. Percobaan Pertama

Pada percobaan pertama, data diproses dengan klasifikasi *Naïve Bayes*. Data yang sudah diekstrak fitur vektor katanya diproses dengan menggunakan salah satu fungsi dari sklearn yaitu *naïve_bayes classifier*. Setting *classifier* menggunakan nilai default dari fungsi *naïve_bayes sklearn*, dengan besar data *training* sebesar 70% atau sebesar 6.993 data. Setelah data selesai dilakukan *training*, data dilakukan testing dengan 30% atau 2.970 data yang telah *displit* sebelumnya. Dengan jumlah kelas yang digunakan adalah tiga kelas (positif, negatif dan netral) proses klasifikasi berdasarkan teroema bayes :

$$P (b \mid a) = [p (a \mid b) * p (b)] / [p (a)]$$

Dengan a adalah set dari fitur yang ada, dan p (b) adalah total *probability* dari sebuah kelas dari frekuensi kelas b. Untuk contoh tweet "*jawatimurpemprov meraih penghargaan bidang pelayanan publik*" dengan memasukkan fitur dan kelas yang ada pada teorema bayes, didapatkan perhitungan untuk masing-masing kelas untuk masing-masing fitur kata :

P(dok) = P(jawatimurpemprov) * P(meraih) * P(penghargaan) … * P(publik). Setiap kalimat memiliki perhitungan dari masing-masing kelas. Contoh untuk kelas positif: P(dok|positif) = P(jawatimrupemprov|positif) * P(meraih|positif) ……* P(public|positif). Kelas yang memiliki nilai tertinggi akan menjadi menjadi hasil akhir prediksi. Untuk contoh kalimat sebelumnya didapatkan hasil seperti pada tabel 4.12.

Tabel 4.12. Hasil akhir klasifikasi

Data	Nilai
Tweet	jawatimurpemprov meraih penghargaan bidang pelayanan publik
Training Positif	**2.72123074**
Training Negatif	0.34286629
Training Netral	1.83328752
Klasifikasi	Positif

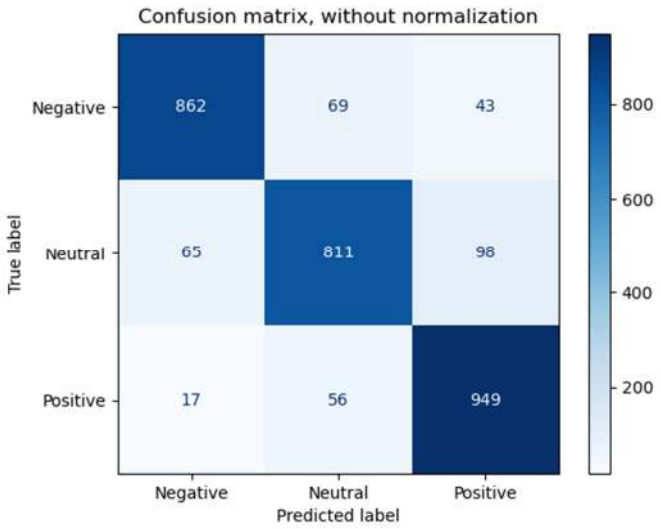
Setelah seluruh dokumen dilakukan training dan prediksi, hasil prediksi data dilakukan pengukuran *performance* dengan fungsi *classification_report* dari modul sklearn. Dan menghasilkan performa seperti pada table 4.13

Tabel 4.13. Hasil percobaan pertama dengan Naïve Bayes

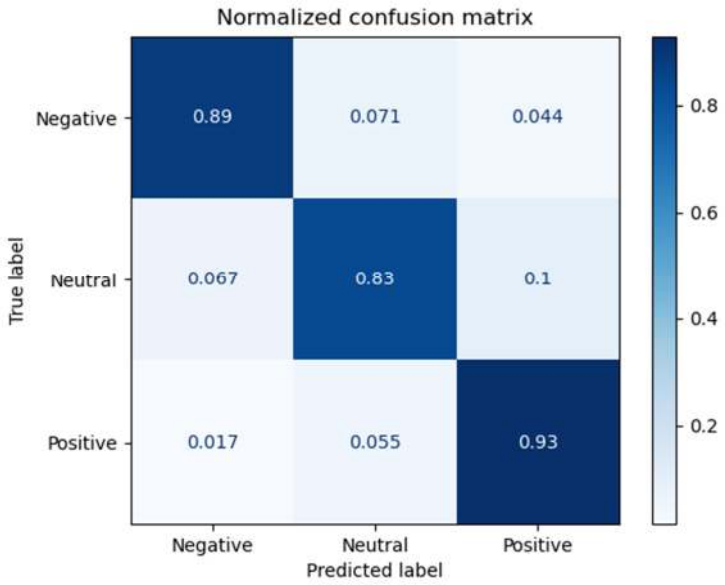
Nilai	Kelas		
	Positif	Netral	Negatif
Precision	0.870642202	0.866452991	0.913135593
Recall	0.928571429	0.832648871	0.885010267
F-Measure	0.898674242	0.84921466	0.898852972

pada percobaan pertama ini, didapatkan akurasi sebesar 88,28% dengan presisi paling besar ditunjukkan oleh kelas negatif sebesar 0,91. Untuk *Recall* terbesar ditunjukkan oleh kelas positif sebesar 0,928. Untuk nilai negatif menunjukkan nilai terendah untuk nilai *recall*. Waktu yang dibutuhkan untuk proses training klasifikasi *Naïve Bayes* dengan data sebesar 9.900 adalah 0,0139 detik dan prediksi waktu sebesar 0,0078 detik.

Confusion matriks dapat dilihat pada gambar 4.1 untuk jumlah data, dan gambar 4.2 untuk data yang sudah dinormalisasi



Gambar 4.1. ***Confusion matrix*** **tanpa normalisasi percobaan pertama**



Gambar 4.2. ***Confusion matrix*** **normalisasi percobaan pertama**

b. Percobaan Kedua

Pada percobaan kedua, data diproses dengan klasifikasi *Support Vector Machine (SVM)*. Data yang sudah diekstrak fitur vektor katanya diproses dengan menggunakan salah satu fungsi dari *sklearn* yaitu *SVM* classifier. *Setting classifier* menggunakan nilai *C* sebesar 100.0 dan nilai *gamma* sebesar 0.01. Kernel yang digunakan adalah kernel *linear* dengan besar data *training* sebesar 70% atau sebesar 6.993 data. Setelah data selesai dilakukan *training*, data dilakukan testing dengan 30% atau 2.970 data yang telah *displit* sebelumnya. Dengan jumlah kelas yang digunakan adalah tiga kelas (positif, negatif dan netral). Proses klasifikasi dilakukan dengan memasukkan vektor kata pada model ruang dengan data indeks fitur kata dan bobot sebagai koordinat. Metode *SVM* bertujuan untuk menemukan garis terbaik yang dapat membagi vector-vektor tersebut ke dalam masing-masing kelas.

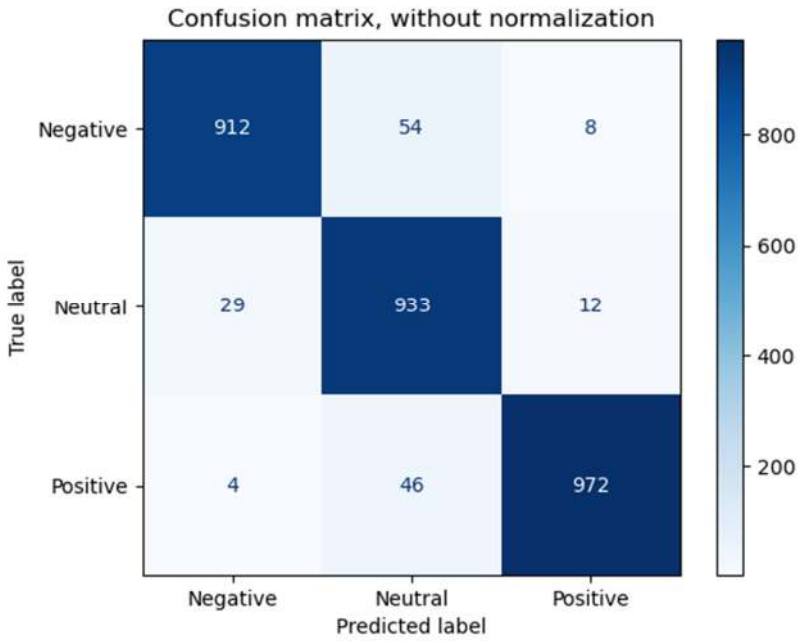
Hasil prediksi data dilakukan pengukuran *performance* dengan fungsi *classification_report* dari modul *sklearn*. Dan menghasilkan performa seperti pada table 4.14

Tabel 4.14. Hasil percobaan kedua dengan SVM

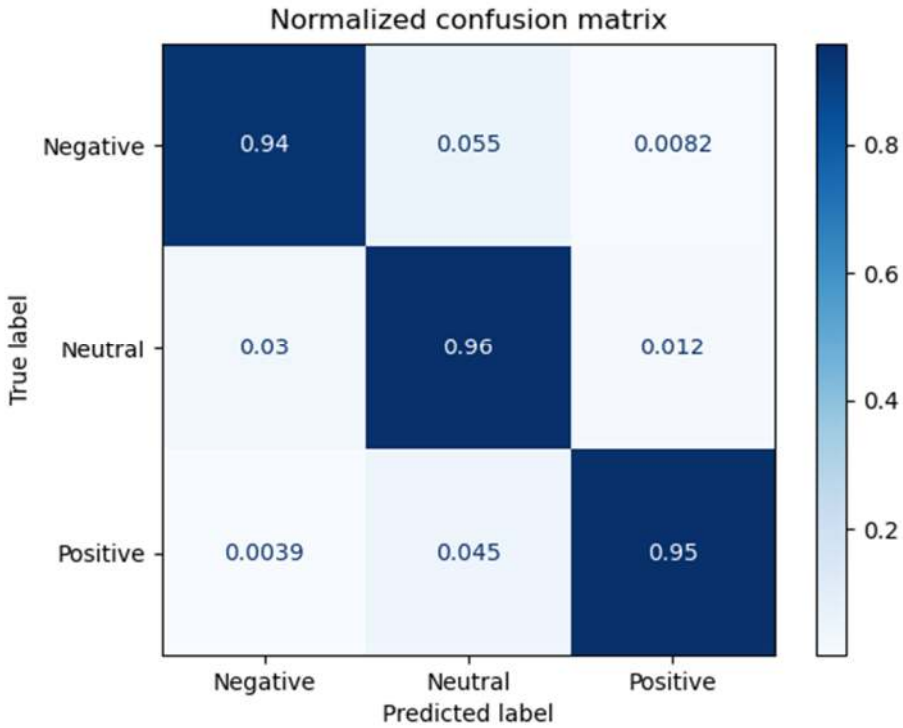
Nilai	Kelas		
	Positif	Netral	Negatif
Precision	0.979838710	0.903194579	0.965079365
Recall	0.951076321	0.957905544	0.936344969
F-Measure	0.965243297	0.929745889	0.950495050

pada percobaan kedua yaitu dengan *SVM* ini, didapatkan akurasi sebesar 94,84% dengan presisi paling besar ditunjukkan oleh kelas positif sebesar 0,97. Untuk *Recall* terbesar ditunjukkan oleh kelas netral sebesar 0,957. Untuk nilai negatif menunjukkan nilai terendah untuk performa *recall*. Waktu yang dibutuhkan untuk proses training klasifikasi *SVM* dengan data sebesar 9.900 lebih lama dibandingkan dengan klasifikasi dengan *Naïve Bayes* yaitu sebesar 57,20 detik dan prediksi waktu sebesar 2,43 detik.

Confusion matriks dapat dilihat pada gambar 4.3 untuk jumlah data, dan gambar 4.4 untuk data yang sudah dinormalisasi



Gambar 4.3. ***Confusion matrix*** **normalisasi percobaan kedua**



Gambar 4.4. ***Confusion matrix*** **normalisasi percobaan kedua**

c. Percobaan Ketiga

Pada percobaan ketiga, data diproses dengan klasifikasi *Naïve Bayes* ditambahkan dengan nilai *SNA* sebagai salah satu fitur. Data yang sudah diekstrak fitur vektor katanya (seperti pada percobaan pertama) dikombinasikan dengan nilai *SNA*. Masing-masing vektor kata akan dikalikan dengan bobot *SNA* dari masing-masing indeks dokumen. Bobot fitur akhir bisa dilihat pada tabel 4.15

Tabel 4.15. Contoh Kombinasi fitur Kata dan SNA

Dokumen	SNA	W1	frek	bobot
Tweet1	0.827	jawatimurpemprov	1	0.827
Tweet1	0.827	penghargaan	2	1.654
…..	……			
Tweet9900	0.334	surabaya	3	1,002
Tweet9900	0.334	macet	1	0.334
….		…..	…..	…..

Dengan menggunakan tiga kelas (positif, negatif dan netral) fitur kata tersebut dilakukan perhitungan sesuai teorema bayes :

$$P (b \mid a) = [p (a \mid b) * p (b)] / [p (a)]$$

Dengan a adalah set dari fitur yang ada, dan p (b) adalah total *probability* dari sebuah kelas dari frekuensi kelas b. Untuk contoh tweet "jawatimurpemprov meraih penghargaan bidang pelayanan publik" dengan memasukkan fitur dan kelas yang ada pada teorema bayes, didapatkan perhitungan untuk masing-masing kelas untuk masing-masing fitur kata :

P(dok) = P(jawatimurpemprov) * P(meraih) * P(penghargaan) … * P(publik). Setiap kalimat memiliki perhitungan dari masing-masing kelas. Contoh untuk kelas positif: P(dok|positif) = P(jawatimrupemprov|positif) * P(meraih|positif) ……* P(public|positif). Kelas yang memiliki nilai tertinggi akan menjadi menjadi hasil akhir prediksi. Untuk contoh kalimat sebelumnya didapatkan hasil seperti pada tabel 4.16.

Tabel 4.16. Hasil akhir klasifikasi dengan fitur SNA

Data	Nilai
Tweet	jawatimurpemprov meraih penghargaan bidang pelayanan publik
Training Positif	**2.21259231**
Training Negatif	0.29551233
Training Netral	1.53141611
Klasifikasi	Positif

Data kemudian diproses dengan menggunakan salah satu fungsi dari sklearn yaitu *naïve_bayes classifier*. Setting *classifier* menggunakan nilai default dari fungsi *naïve_bayes sklearn*, dengan besar *training* sebesar 70% atau sebesar 6.993 data. Setelah data selesai dilakukan *training*, data dilakukan testing dengan 30% atau 2.970 data yang telah *displit* sebelumnya. Perhitungan selanjutnya seperti pada percobaan 1 hingga menghasilkan model training untuk masing-masing kelas. Data testing diproses pada model gabungan sehingga diperoleh hasil prediksi untuk seluruh data posting. Hasil prediksi data dilakukan pengukuran *performance* dengan fungsi *classification_report* dari modul *sklearn*. Dan menghaislkan performa seperti pada table 4.17

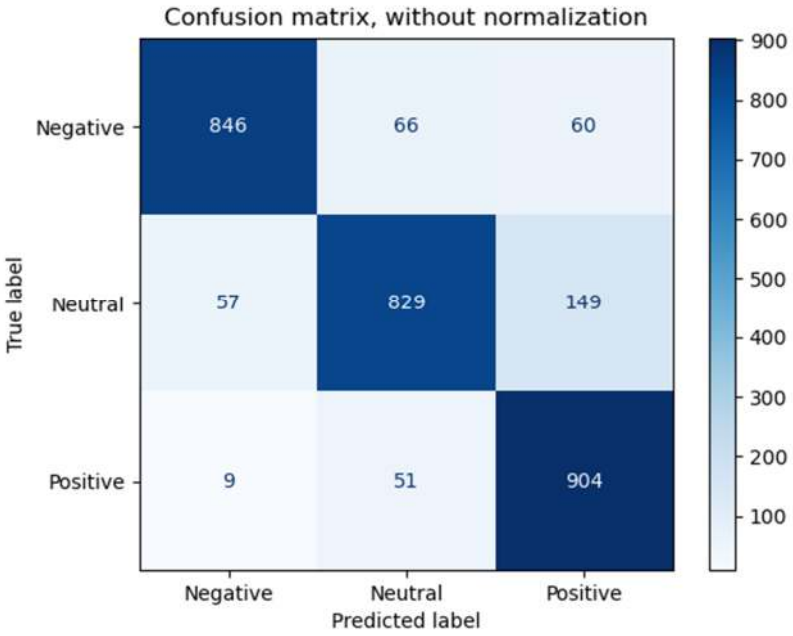
Tabel 4.17. Hasil percobaan ketiga dengan Naïve Bayes + SNA

Nilai	Kelas		
	Positif	Netral	Negatif
Precision	0.812219227	0.876321353	0.927631579
Recall	0.937759336	0.800966184	0.870370370
F-Measure	0.870486278	0.836951035	0.898089172

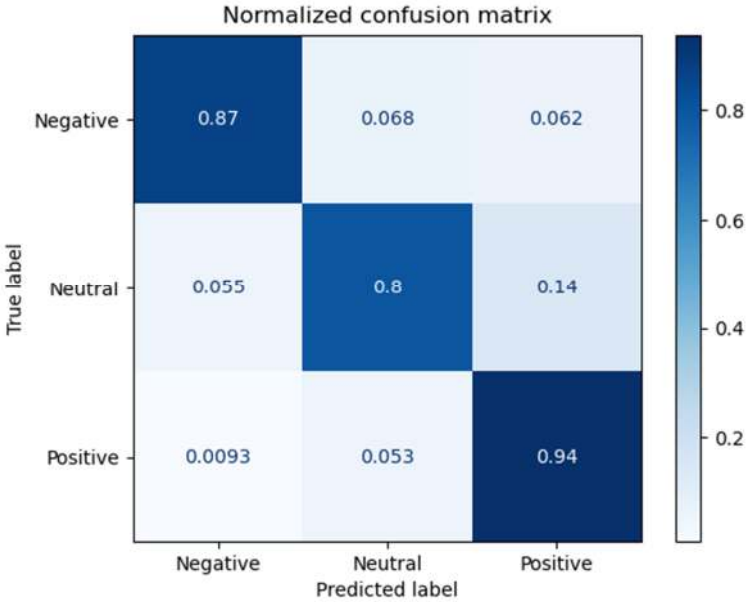
pada percobaan ketiga ini, didapatkan akurasi sebesar 86,80% dengan presisi paling besar ditunjukkan oleh kelas negatif sebesar 0,927. Untuk *Recall* terbesar ditunjukkan oleh kelas positif sebesar 0,937. Nilai *Recall* terkecil dimiliki oleh kelas netral. Waktu yang dibutuhkan untuk proses training klasifikasi *Naïve*

Bayes dengan data sebesar 9.900 adalah 0,0155 detik dan prediksi waktu sebesar 0,0041 detik.

Confusion matriks dapat dilihat pada gambar 4.5 untuk jumlah data, dan gambar 4.6 untuk data yang sudah dinormalisasi



Gambar 4.5. *Confusion matrix* tanpa normalisasi percobaan ketiga



Gambar 4.6. *Confusion matrix* normalisasi percobaan ketiga

Selain itu dilakukan beberapa percobaan dengan perbandingan data training-testing yang beragam, antara lain 60:40 dan 50:50. Dari hasil percobaan didapatkan pada perbandingan data 60:40 menghasilkan data dengan nilai lebih baik daripada menggunakan data training 70:30. Percobaan dengan penambahan pembobotan juga dilakukan dengan kondisi rendah, sedang, tinggi : data dengan *SNA* rendah dengan nilai kurang dari 0.3 diberikan bobot 0.5, untuk *SNA* sedang dengan nilai antara 0.3 dan 0.6 diberikan bobot 0.75. Sedangkan *SNA* tinggi dengan nilai lebih dari 0.6 diberikan bobot 2. Dilakukan juga percobaan dengan perhitungan rasio follower dengan pembobotan seperti pada perhitungan *f-measure* dengan nilai beta 0.5, 1 dan 2. Namun nilai akurasi dan performa dari hasil percobaan belum menunjukkan kenaikan pada kombinasi *Naïve Bayes* dan *SNA*. Hasil percobaan dapat dilihat pada tabel 4.18

Tabel 4.18. Rangkuman Percobaan ketiga dengan variasi perbandingan data

Nilai	Percobaan 3 NB + SNA			
	Precision	**Recall**	**F-Measure**	**Accuration**
Trainingset 70:30	0.872057	0.869699	0.868509	86.80%
Trainingset 60:40	0.872318	0.870758	0.868835	86.92%
Trainingset 50:50	0.872909	0.869747	0.867463	86.87%
Trainingset 70:30 + bobot tinggi, sedang, rendah	0.871036	0.868735	0.867449	86.70%
70:30 dengan pembobotan 0.5* follower*following/ (follower*following)	0.872087	0.869746	0.867915	80.14%
70:30 dengan pembobotan 1* follower*following/ (follower*following)	0.872010	0.869409	0.867609	84.31%

70:30 dengan pembobotan 2* follower*following/ (follower*following)	0.87171	0.869297	0.867658	75.96%

d. Percobaan Keempat

Pada percobaan keempat, data diproses dengan klasifikasi *Support Vector Machine (SVM)* dikombinasikan dengan *SNA* sebagai salah satu fitur. Data yang sudah diekstrak fitur vektor katanya dikombinasikan dengan nilai *SNA* sebagai salah satu fitur. Seperti pada percobaan ketiga, bobot masing-masing fitur dikalikan dengan data *SNA* dari masing-masing indeks dokumen, hasil fitur akhir bisa dilihat pada tabel 4.15. Data kemudian diproses dengan menggunakan salah satu fungsi dari *sklearn* yaitu *SVM* classifier. *Setting classifier* menggunakan nilai *C* sebesar 100.0 dan nilai *gamma* sebesar 0.01. *Kernel* yang digunakan adalah kernel *linear* dengan besar data *training* sebesar 70%. Setelah data selesai dilakukan *training*, data dilakukan *testing* dengan 30% data yang telah displit sebelumnya. Hasil prediksi data dilakukan pengukuran *performance* dengan fungsi *classification_report* dari modul *sklearn*. Dan menghasilkan performa seperti pada table 4.19.

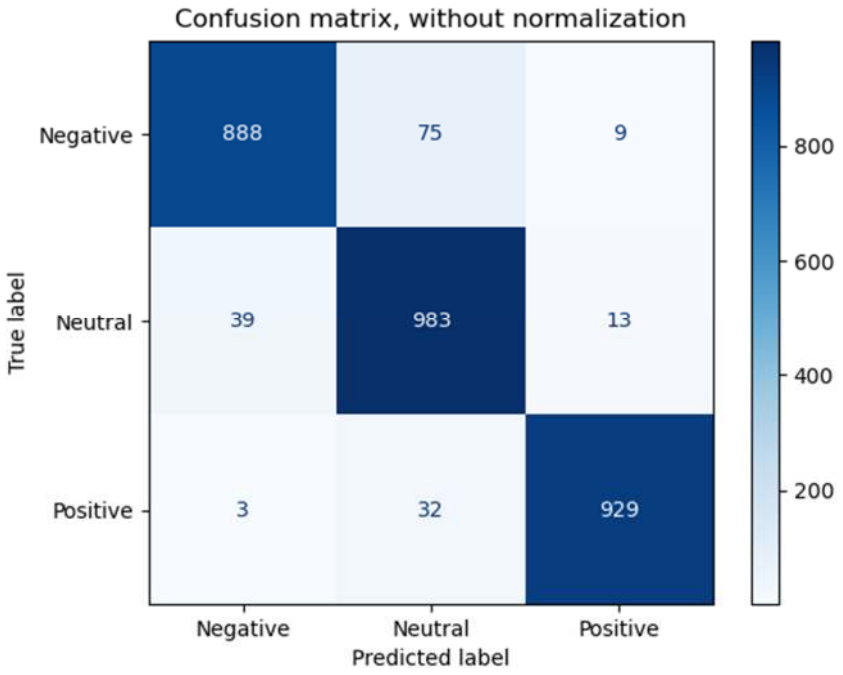
Tabel 4.19. Hasil percobaan keempat dengan SVM + SNA

Nilai	Kelas		
	Positif	Netral	Negatif
Precision	0.976866456	0.901834862	0.954838710
Recall	0.963692946	0.949758454	0.913580247
F-Measure	0.970234987	0.925176471	0.933753943

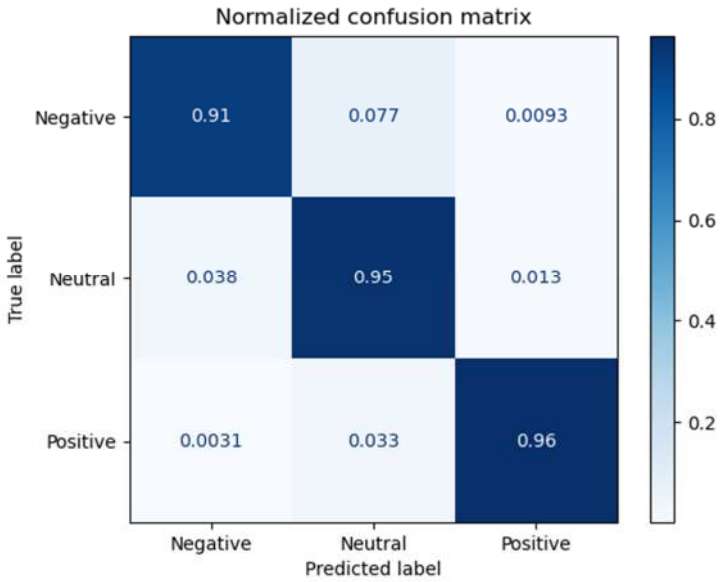
pada percobaan keempat yaitu dengan *SVM +SNA* ini, didapatkan akurasi sebesar 94,24% dengan presisi paling besar ditunjukkan oleh kelas positif sebesar 0,97. Untuk *Recall* terbesar ditunjukkan oleh kelas positif sebesar 0,96. Untuk nilai negatif menunjukkan nilai terendah untuk performa *recall*. Waktu yang dibutuhkan untuk proses training klasifikasi *SVM* dengan data sebesar 9.900 lebih lama

dibandingkan dengan klasifikasi dengan *Naïve Bayes* yaitu sebesar 28,10 detik dan prediksi waktu sebesar 2,28 detik.

Confusion matriks dapat dilihat pada gambar 4.7 untuk jumlah data, dan gambar 4.8 untuk data yang sudah dinormalisasi



Gambar 4.7. *Confusion matrix* normalisasi percobaan keempat



Gambar 4.8. *Confusion matrix* normalisasi percobaan keempat

Selain itu dilakukan beberapa percobaan dengan perbandingan data training-testing yang beragam, antara lain 60:40 dan 50:50. Dari hasil percobaan didapatkan pada perbandingan data 70:30 menghasilkan data dengan nilai lebih baik daripada menggunakan data training dengan perbandingan lainnya. Percobaan dengan penambahan pembobotan juga dilakukan dengan kondisi rendah, sedang, tinggi : data dengan *SNA* rendah dengan nilai kurang dari 0.3 diberikan bobot 0.5, untuk *SNA* sedang dengan nilai antara 0.3 dan 0.6 diberikan bobot 0.75. Sedangkan *SNA* tinggi dengan nilai lebih dari 0.6 diberikan bobot 2. Namun nilai akurasi dan performa dari hasil percobaan belum menunjukkan kenaikan signifikan pada kombinasi *SVM* dan *SNA*. Hasil percobaan dapat dilihat pada tabel 4.20.

Tabel 4.20. Rangkuman Percobaan keempat dengan variasi perbandingan data

Nilai	Percobaan 4 SVM+SNA			
	Precision	**Recall**	**F-Measure**	**Accuration**
70:30	0.94451334	0.94234388	0.943055134	94.24%
60:40	0.94012434	0.93645265	0.937227282	93.63%
50:50	0.93234095	0.92852154	0.929261673	93.63%
70:30 + bobot tinggi, sedang, rendah	0.94458698	0.94232585	0.943047322	94.24%

e. Perbandingan Hasil seluruh percobaan

Hasil dari seluruh percobaan dengan jumlah perbandigan dataset yang sama telah dirangkum dalam tabel 4.21. Hasil rangkuman menunjukkan adanya penurunan akurasi yang cukup besar pada metode *Naïve Bayes* yang dikombinasikan dengan perhitungan *SNA* pada penggunaan data training 70% dan testing 30%. Sedangkan pada kombinasi *SVM* dengan *SNA* mengalami penurnunan yang tidak terlalu signifikan. Perhitungan performance rata-rata kelas pada masing-masing percobaan juga tidak menunjukkan kenaikan. Sebaliknya kombinasi menunjukkan penurunan pada kombinasi *SA* dengan *SNA*, meski nilai penurunan tidak terlalu besar untuk masing-masing perhitungan *performance*.

Tabel 4.21. Hasil Rangkuman Percobaan

Nilai	Percobaan			
	NB	SVM	NB + SNA	SVM + SNA
Precision	0.88341026	0.949371	0.872057	0.944513
Recall	0.88207685	0.948442	0.869699	0.942344
F-Measure	0.88224729	0.948495	0.868509	0.943055
Accuration	88.28%	94.84%	86.80%	94.24%

Dari perbandingan tersebut dapat dilihat bahwa akurasi tertinggi dengan menggunakan metode *SVM* tanpa penambahan *SNA* sebesar 94,84 %. Akurasi paling rendah didapatkan dengan metode *NB+SNA* sebesar 86,80%. Hal ini bisa jadi disebabkan karena banyaknya akun yang pada saat di lakukan pencarian data follower dan following sudah dinonaktifkan. Karena pengumpulan data *SA* dan *SNA* tidak dilakukan bersamaan sehingga user yang pada waktu data posting twitter atau Instagram ditarik belum tentu masih aktif pada waktu data *SNA* ditarik.

Untuk mengetahui bagaimana hasil sentiment analysis pada masing-masing algoritma, dilakukan ujicoba pada tiga data yang mewakili pada model gabungan *SA+SNA*. Tiga data tersebut dipilih dengan kondisi nilai *SNA* kurang dari 0.30 , antara 0.30 dan 0.60 dan lebih dari 0.60. Hasil percobaan dapat dilihat pada tabel 4.22.

Tabel 4.22. Hasil Rangkuman Percobaan dengan data posting

Nilai SNA	Data Posting	Manual	NB + SNA	SVM + SNA
0.052	Sip. Semangat dukung tim jawa timur	Postif	Positif	Positif
0.200	Bapak Hari ini benarkah ada Jadwal ke Kab. Jember Jawa Timur	Netral	Netral	Netral

0.0327	Mohon yg wilayah jawa timur @TVRINasional tidak dipakai tvri jawatimur ,, bikin kecewa,,	Negatif	Negatif	Negatif
0.521	Mantap bu @KhofifahIP dan selamat datang di jember @humasprovjatim	Positif	Positif	Positif
0.453	Bibir pantai kenjeran surabaya jawa timur	Netral	Netral	Netral
0.438	Jawa timur menakutkanku padahal kampung sendiri:(	Negatif	**Netral**	**Netral**
0.339	Jln mulai dri utara rel kereta cerme sampai desa pandu banyak jln yg berlubang,klo abis hujan genangan dmn2 serta kondisi jln yg gelap. Sgt brbahaya bagi R2 dn rawan lakalantas. Mohon ditindaklanjuti. @URCBIMAGresik @HumasGresik @RTMCJatim @JatimPemprov @KhofifahIP @infoGRESIK	Negatif	Negatif	Negatif
0.704	Waw selamat jawa timur pemenang PI 2020	Positif	Positif	Positif
0.750	Kawasan Physical Distancing Jawa Timur pic.twitter.com/jJdstYp1Lu	Netral	Netral	Netral

0.999	Penelitian di Jawa Timur mengungkap bahaya residu pembakaran plastik pada produk daging dan telur ayam kampung. https:// tirto.id/el8q	Negatif	Negatif	Negatif

Dari hasil percobaan yang dilakukan dengan masing-masing *sample* data *SNA* baik tinggi, sedang atau rendah didapatkan prediksi terhadap masing-masing kelas terjadi kesalahan pada kelas negatif pada nilai *SNA* sedang. Hal ini bisa dikarenakan, pada data *posting* tersebut vector kata terlalu sedikit dan hanya satu kata yang melambangkan sentiment negatif. Hal tersebut terbukti ketika dilakukan prediksi dengan data *posting* yang lebih panjang, kelas yang diprediksi sesuai dengan label kelas secara manual.

(Halaman ini sengaja dikosongkan)

BAB 5
KESIMPULAN DAN SARAN

Bab ini menjelaskan mengenai kesimpulan berdasarkan hasil dan pembahasan yang telah dilakukan pada bab sebelumnya. Saran yang diperoleh dari kesimpulan juga disampaikan agar berguna sebagai masukan dalam pengembangan penelitian di masa datang.

5.1 Kesimpulan

Berdasarkan tahap yang telah dilakukan sejak awal penelitian dapat ditarik kesimpulan berupa rangkuman dari proses pengumpulan data, *preprocessing*, pembuatan *Klasifikasi Naïve Bayes*, klasifikasi *SVM* dan Analisa *SNA*.

1. Dengan menggunakan Analisa sentiment *SVM* menunjukkan hasil yang lebih baik daripada penggunaan klasifikasi *Naïve Bayes* dengan jumlah data yang sama. Kelas positif menunjukkan performa yang lebih baik dari kelas lain, baik dengan metode *Naïve Bayes* maupun *SVM*
2. Pengaruh pembobotan fitur *SNA* untuk subjek orang pada perhitungan *Sentiment Analysis* tidak sesuai harapan karena mengakibatkan penurunan nilai akurasi baik pada metode *Naïve Bayes* maupun pada *SVM.* Salah satu penyebab hal ini adalah fitur *SNA* yang digunakan menyebabkan perbandingan nilai bobot yang lebih rendah pada fitur kata yang digunakan untuk dataset *training Sentiment Analysis*.
3. Pengumpulan data yang tidak dilakukan secara bersamaan antara pengumpulan data *Posting* dan data *follower/following* menyebabkan beberapa data memiliki bobot *SNA* yang rendah atau bahkan nol.

5.2 Saran

Berdasarkan permasalahan yang muncul pada saat penelitian terdapat saran yang dapat dijadikan pertimbangan untuk penelitian beriktunya. Saran penelitian tersebut antara lain:

1. Perlu dilakukan penyaringan data yang berhubungan dengan Pemprov Jatim atau hanya sekedar melakukan mention.

2. Perlu informasi *SNA* yang lebih detail untuk Analisa yang lebih mendalam, seperti *engagement*, jumlah jumlah post dalam satu bulan terakhir dan informasi lain yang mendukung pembobotan.
3. Pengumpulan data yang dikombinasikan hendaknya bersamaan. Penarikan data pada waktu yang berbeda antara data *posting* dan data *following / follower* dapat menyebabkan kerancuan data.

(Halaman ini sengaja dikosongkan)

DAFTAR PUSTAKA

Afshoh, F., Pamungkas, E.W., Kom, S., Kom, M., 2017. Analisa Sentimen Menggunakan Naïve Bayes Untuk Melihat Persepsi Masyarakat Terhadap Kenaikan Harga Jual Rokok Pada Media Sosial Twitter (Undergraduated Thesis). Universitas Muhammadiyah Surakarta.

Automated Fake Account Detection at LinkedIn [WWW Document], n.d. URL https://engineering.linkedin.com/blog/2018/09/automated-fake-account-detection-at-linkedin (accessed 6.17.20).

Bermingham, A., Conway, M., McInerney, L., O'Hare, N., Smeaton, A.F., 2009. Combining social network analysis and sentiment analysis to explore the potential for online radicalisation, in: Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances In. IEEE, pp. 231–236.

Bravo-Marquez, F., 2017. Acquiring and Exploiting Lexical Knowledge for Twitter Sentiment Analysis (PhD Thesis). University of Waikato.

Bravo-Marquez, F., Frank, E., Pfahringer, B., 2016. Annotate-Sample-Average (ASA): A New Distant Supervision Approach for Twitter Sentiment Analysis, in: 22nd European Conference on Artificial Intelligence (ECAI). Presented at the 22nd European Conference on Artificial Intelligence (ECAI), IOS Press, pp. 498–506. https://doi.org/10.3233/978-1-61499-672-9-498

Devi, D.N., Kumar, C.K., Prasad, S., 2016. A feature based approach for sentiment analysis by using support vector machine, in: Advanced Computing (IACC), 2016 IEEE 6th International Conference On. IEEE, pp. 3–8.

El Azab, A., Idrees, A.M., Mahmoud, M.A., Hefny, H., 2016. Fake account detection in twitter based on minimum weighted feature set. Int. Sch. Sci. Res. Innov. 10, 13–18.

Faradhillah, N.Y., Kusumawardani, R.P., Hafidz, I., 2016. Eksperimen Sistem Klasifikasi Analisa Sentimen Twitter pada Akun Resmi Pemerintah Kota Surabaya Berbasis Pembelajaran Mesin. SESINDO 2016 2016.

Fornacciari, P., Mordonini, M., Tomaiuolo, M., 2015. Social Network and Sentiment Analysis on Twitter: Towards a Combined Approach., in: KDWeb. pp. 53–64.

Grunspan, D.Z., Wiggins, B.L., Goodreau, S.M., Dolan, E., 2014. Understanding Classrooms through Social Network Analysis: A Primer for Social Network Analysis in Education Research. CBE—Life Sci. Educ. 13, 167–178. https://doi.org/10.1187/cbe.13-08-0162

Gryc, W., Moilanen, K., 2014. Leveraging textual sentiment analysis with social network modelling. Text Polit. Positions Text Anal. Discip. 55, 47.

Indrayuni, E., 2016. Analisa Sentimen Review Hotel Menggunakan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization. EVOLUSI-J. Sains Dan Manaj. AMIK BSI Purwok. 4.

Kementerian Komunikasi dan Informatika. [WWW Document], 2018. . Jumlah Pengguna Internet 2017 Meningkat Kominfo Terus Lakukan Percepatan Pembang. Broadband. URL https://kominfo.go.id/index.php/content/detail/12640/siaran-pers-no-53hmkominfo022018-tentang-jumlah-pengguna-internet-2017-meningkat-kominfo-terus-lakukan-percepatan-pembangunan-broadband/0/siaran_pers (accessed 11.2.18).

Kusumawati, I., Pamungkas, E.W., Kom, S., Kom, M., 2017. Analisa Sentimen Menggunakan Lexicon Based Untuk Melihat Persepsi Masyarakat Terhadap Kenaikan Harga Rokok Pada Media Sosial Twitter (PhD Thesis). Universitas Muhammadiyah Surakarta.

Medhat, W., Hassan, A., Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Eng. J. 5, 1093–1113.

Nurhuda, F., Sihwi, S.W., Doewes, A., 2016. Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier. ITSMART J. Teknol. Dan Inf. 2, 35–42.

Oktora, R., Alamsyah, A., 2014. POLA INTERAKSI DAN AKTOR YANG PALING BERPERAN PADA EVENT JGTC 2013 MELALUI MEDIA SOSIAL TWITTER (STUDI MENGGUNAKAN METODE SOCIAL NETWORK ANALYSIS). J. Manaj. Indones. 14, 201–209. https://doi.org/10.25124/jmi.v14i3.370

Osimo, D., Mureddu, F., 2012. Research challenge on opinion mining and sentiment analysis. Univ. Paris-Sud Lab. LIMSI-CNRS Bâtim. 508.

OTDA Kemendagri [WWW Document], 2018. . JUMLAH PROVINSI KAB DAN KOTA. URL https://otda.kemendagri.go.id/seputar-otda/jumlah-provinsi-kab-dan-kota/ (accessed 11.2.18).

Pratama, P.G., Rakhmawati, N.A., 2019. Social Bot Detection on 2019 Indonesia President Candidate's Supporter's Tweets. Procedia Comput. Sci., The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia 161, 813–820. https://doi.org/10.1016/j.procs.2019.11.187

Rossi Trisna Lestari, A., Perdana, R., Fauzi, M., 2017. Analisis Sentimen Tentang Opini Pilkada DKI Jakarta 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji. J. Pengemb. Teknlogi Inf. Dan Ilmu Komput. 1, 1718–1724.

Shepelenko, O., 2017. Opinion Mining and Sentiment Analysis using Bayesian and Neural Networks Approaches.

Tiwari, P., Velayutham, T., 2017. Bot identification: Helping analysts for right data in twitter. pp. 1–5. https://doi.org/10.1109/ICACCAF.2017.8344722

Van Der Walt, E., Eloff, J., 2018. Using Machine Learning to Detect Fake Identities: Bots vs Humans. IEEE Access 6, 6540–6549. https://doi.org/10.1109/ACCESS.2018.2796018

Winarko, E., 2017. Sentimen Analisis Tweet Berbahasa Indonesia Dengan Deep Belief Network. IJCCS Indones. J. Comput. Cybern. Syst. 11, 187–198.

Yuliana, I., Santosa, P.I., Setiawan, N.A., 2015. Analisis Jejaring Media Sosial untuk Pemetaan pada Komunitas Online, in: Seminar Nasional Aplikasi Teknologi Informasi (SNATI).

BIODATA PENULIS

Ikhwan Rustanto, lahir di Surabaya pada tanggal 23 Desember 1985. Penulis telah menempuh Pendidikan formal di SDN Wonokusumo 5 Surabaya, SMPN 5 Surabaya, dan SMAN 1 Surabaya. Pada tahun 2004, penulis melanjutkan Pendidikan Strata-1 di Institut Teknologi Sepuluh Nopember, dengan jurusan Teknik Informatika dan kelompok keahlian Rekayasa Perangkat Lunak. Penulis lulus pada tahun 2010 dengan Tugas Akhir berjudul "Pembangunan Perangkat Lunak Pembelajaran Online mandiri dengan berbasis Rich Internet Application (RIA) study kasus pada matakuliah Rekayasa Perangkat Lunak". Pada penelitian ini, penulis mengambil konsentrasi dalam Akuisisi Data dan Diseminasi Informasi (ADDI) dengan topik penelitian mengenai kombinasi *Social Network Analysis* dan *Sentiment Analysis.* Segala kritik dan saran yang membangun dapat disampaikan melalui *ikhwanrustanto@gmail.*