TUGAS AKHIR – KS184822

ANALISIS RESPONS WARGANET TERHADAP DEBAT CALON PRESIDEN 2019 DI TWITTER DENGAN METODE CLUSTERED SUPPORT VECTOR MACHINES

SHINDI SHELLA MAY WARA
NRP 062115 4000 0101

Dosen Pembimbing
Dr. rer.pol. Dedy Dwi Prastyo, S.Si, M.Si.
Dr. Dra. Kartika Fithriasari, M.Si

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA. KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2019

TUGAS AKHIR – KS184822

ANALISIS RESPONS WARGANET TERHADAP DEBAT CALON PRESIDEN 2019 DI TWITTER DENGAN METODE CLUSTERED SUPPORT VECTOR MACHINES

SHINDI SHELLA MAY WARA
NRP 062115 4000 0101

Dosen Pembimbing
Dr. rer.pol. Dedy Dwi Prastyo, S.Si, M.Si.
Dr. Dra. Kartika Fithriasari, M.Si

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA. KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2019

FINAL PROJECT – KS184822

ANALYSIS OF NETIZEN RESPONSE TO DEBATE OF PRESIDENT CANDIDATES 2019 ON TWITTER USING CLUSTERED SUPPORT VECTOR MACHINES

SHINDI SHELLA MAY WARA
NRP 062115 4000 0101

Supervisors
Dr. rer.pol. Dedy Dwi Prastyo, S.Si, M.Si.
Dr. Dra. Kartika Fithriasari, M.Si

UNDERGRADUATE PROGRAMME
DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICS, COMPUTING, AND DATA SCIENCE
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2019

(Halaman ini sengaja dikosongkan)

LEMBAR PENGESAHAN

ANALISIS RESPONS WARGANET TERHADAP DEBAT CALON PRESIDEN 2019 DI TWITTER DENGAN METODE *CLUSTERED SUPPORT VECTOR MACHINES*

TUGAS AKHIR

Diajukan untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Statistika
pada
Program Studi Sarjana Departemen Statistika
Fakultas Matematika, Komputasi, dan Sains Data
Institut Teknologi Sepuluh Nopember

Oleh :
Shindi Shella May Wara
NRP. 062115 4000 0101

Disetujui oleh Pembimbing:
Dr. rer.pol Dedy Dwi Prastyo, S.Si, M.Si. (1)
NIP. 19831204 200812 1 002
Dr. Dra. Kartika Fithriasari M.Si. (2)
NIP. 19691212 199303 2 002

Mengetahui,
Kepala Departemen Statistika

Dr. Suhartono
NIP. 19710929 199512 1 001

SURABAYA, JULI 2019

ANALISIS RESPONS WARGANET TERHADAP DEBAT CALON PRESIDEN 2019 DI TWITTER DENGAN METODE *CLUSTERED SUPPORT VECTOR MACHINES*

Nama Mahasiswa	**:**	**Shindi Shella May Wara**
NRP	**:**	**062115 4000 0101**
Departemen	**:**	**Statistika**
Dosen Pembimbing	**:**	**Dr. rer.pol Dedy Dwi Prastyo, S.Si, M.Si.** **Dr. Dra. Kartika Fithriasari M.Si.**

Abstrak

Debat calon presiden 2019 merupakan salah satu sarana kampanye yang disediakan oleh Komisi Pemilihan Umum (KPU) dimana pasangan calon presiden saling berhadapan untuk menyampaikan visi dan misi serta tujuan yang dibawa untuk masa depan Indonesia. Debat calon presiden 2019 disiarkan langsung secara nasional sehingga masyarakat dapat secara langsung menanggapi dan lebih mengenal calon presidennya. Bagi warganet dapat secara langsung menyampaikan pendapat terhadap debat presiden 2019 pada media sosial, salah satunya melalui Twitter. Tanggapan yang diberikan bisa berupa sentimen positif, negatif, dan netral. Menyikapi keadaan tersebut, dilakukan penelitian tentang analisis klasifikasi sentimen dengan metode Clustered Support Vector Machines. Data yang digunakan dalam penelitian ini diambil dari tweets dari pengguna Twitter Indonesia yang dipublikasikan pasca debat I-V dengan menyebutkan salah satu nama akun twitter pasangan calon presiden Indonesia 2019. Hasil klasifikasi menunjukkan bahwa pasangan calon 1 lebih banyak mendapatkan tweets sentimen positif dan negatif. Penelitian ini membandingkan metode CSVM berdasarkan nilai akurasi dan AUC pada kernel Linear, Polinomial, dan RBF. Hasil terbaik akurasi dan AUC menggunakan Kernel Polinomial pada Pasca Debat I-IV dengan Akurasi sebesar 0,87; 0,85; 0,88; dan 0,87 dan AUC sebesar 0,90; 0,88; 0,90; dan 0,90. Pada pasca debat V, kernel RBF menghasilkan hasil terbaik dengan Akurasi dan AUC masing-masing sebesar 0,87 dan 0,90.

Kata kunci : Calon Presiden, Debat, K-Means, Sentimen, SVM

(Halaman ini sengaja dikosongkan)

ANALYSIS OF NETIZEN RESPONSE TO DEBATE OF PRESIDENT CANDIDATES 2019 ON TWITTER USING CLUSTERED SUPPORT VECTOR MACHINES

Name	**:**	**Shindi Shella May Wara**
Student Number	**:**	**062115 4000 0101**
Department	**:**	**Statistika**
Supervisors	**:**	**Dr. rer.pol Dedy Dwi Prastyo, S.Si, M.Si.**
		Dr. Dra. Kartika Fithriasari M.Si.

Abstract

Debate of president candidates in 2019 is one of campaign endorsed by "Komisi Pemilihan Umum" (KPU) where president candidates are facing each other to convey their vision, mission and the purpose they brought for the future of Indonesia. Debate of president candidates in 2019 was live broadcasted nationally so that people could respond and have an understanding on their president candidates. In social media, people could directly convey their opinions by online towards presiden debate 2019, one of them is via Twitter. Responses given by people through social media could be positive sentiment, negative or neutral. Given these situations, this research about analysis of sentiment using Clustered Support Vector Machines. The data used in this research were obtained from tweets from Indonesian users that publicated on post-debate I-V that mentioned one of president candidates twitter account. The result of classification show that candidate number 1 have more positive and negative sentiment tweets. The purpose of this research is to compare the accuracy and AUC value in Linear, Polinomial and RBF Kernel. Thus, the highest accuracy and AUC was obtained by using Polinomial Kernel on post-debate I-IV with the accuracy of 0,87; 0,85; 0,88; and 0,87 and the AUC value of 0,90; 0,88; 0,90 and 0,90. In post-debate V, RBF Kernel produce the highest accuracy of 0,87 and AUC value of 0,90.

Keywords : Debate, K-Means, President Candidates, Sentiment, SVM

(Halaman ini sengaja dikosongkan)

KATA PENGANTAR

Puji syukur kehadirat Allah SWT atas berkat rahmat, taufik, dan hidayah-Nya penulis dapat menyelesaikan laporan Tugas Akhir yang berjudul "**Analisis Respons Warganet terhadap Debat Calon Presiden 2019 di Twitter dengan Metode *Clustered Suport Vector Machines***".

Penulis menyadari bahwa Tugas Akhir ini dapat terselesaikan tidak terlepas dari bantuan dan dukungan berbagai pihak. Oleh karena itu, penulis menyampaikan terima kasih kepada:

1. Kedua orang tua, Bapak Sutoyo dan Ibu Yus Laeli serta keluarga, tante Anjar Mudaini dan Adhiesya Pradanadinda Cahya yang selalu memberikan do'a, nasihat, kasih sayang, dan dukungan yang diberikan kepada penulis.
2. Dr. Suhartono selaku Ketua Departemen Statistika dan Dr. Santi Wulan Purnami, S.Si, M.Si selaku Ketua Program Studi Sarjana Statistika yang telah memberikan ijin atas fasilitas, sarana, dan prasarana.
3. Bapak Dr. rer. pol. Dedy Dwi Prastyo, S.Si., M.Si. dan Ibu Dr. Dra. Kartika Fithriasari, M.Si selaku dosen pembimbing yang selalu memberikan bimbingan, masukan, dukungan serta motivasi selama penyusunan Tugas Akhir.
4. Bapak Dr. rer. pol. Heri Kuswanto, S.Si., M.Si. dan Bapak Imam Safawi Ahmad, S.Si., M.Si selaku dosen penguji yang memberikan kritik dan saran yang membangun dalam penyelesaian Tugas Akhir.
5. Seluruh dosen Departemen Statistika ITS yang telah memberikan ilmu dan, serta segenap civitas akademika Departemen Statistika ITS.
6. *My partner in crime* : Yoga Samudra, Febri Indah, Ulfa Siti, dan Rike Andriyani.
7. *Vitamin Squad* : Henidar Islami, Imas Ayu, Fitria Nurul, Ulfa Siti.

8. Semangat Belajar : Devita Prima, Moch. Trianto, Arlandio Nur, Taufiq Azmi, Ulfa Siti, Nurun Nahdliyah, Ihsan Ananto, Ainun Umami, Yusuf Herman
9. Kakak Motivator S2 : Siti Qomariyah dan Elly Pusporani
10. Grup Cecan : Helenna Asa, Ilma Zuhrotul, Esti Wulandari
11. Teman berbagi cerita : Dian Rizky, Desintya Rachma, Aprilia Andririani, Imroatus Sholikhah, dan Rahayu Prihatini,
12. Teman SMA : Nimas Lutfiana, Shafira Thrisnadia, Wiandra Alif, Novia Nurhaslinda, Rodia Amanata
13. Teman SMP aka D'Pink : Prata Pramesti, Pratita Kirana, Dwi Niken, Miranda Amami, Aprinia Fajar, Lola Windy.
14. Teman-Teman SCC-HIMASTA ITS periode kepengurusan 2017-2018 yang memberi dukungan kepada penulis.
15. Teman-teman Statistika ITS ∑26 angkatan 2015, serta ∑27 angkatan 2016, yang selalu memberikan dukungan kepada penulis selama ini.
16. Semua teman, relasi dan berbagai pihak yang tidak bisa penulis sebutkan namanya satu persatu yang telah membantu dalam penulisan laporan ini.

Dengan selesainya laporan Tugas Akhir ini, penulis menyadari bahwa penelitian Tugas Akhir ini masih belum sempurna, jika masih ada kekurangan diharapkan saran dan kritik agar dapat mengembangkan penelitian ini.

Surabaya, Jusni 2019

Penulis

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

(Halaman ini sengaja dikosongkan)

DAFTAR TABEL

Halaman

DAFTAR LAMPIRAN

Halaman

(Halaman ini sengaja dikosongkan)

BAB I
PENDAHULUAN

1.1 Latar Belakang

Demokrasi merupakan salah satu bentuk pemerintahan. Menurut Abraham Lincoln sebagai Bapak Demokrasi saat berpidato di Gettysburg, mengatakan bahwa demokrasi merupakan pemerintahan dari rakyat, oleh rakyat, dan untuk rakyat (Hum, 2014). Indonesia merupakan salah satu negara yang memiliki bentuk pemerintahan dengan sistem demokrasi. Hal tersebut terlihat pada Undang-Undang Dasar 1945 Pasal 1 Ayat 2 yang berisi bahwa kedaulatan ada di tangan rakyat dan dilaksanakan menurut Undang-Undang Dasar. Salah satu bentuk aspirasi rakyat terhadap pemerintahan dapat terlihat saat dilaksanakannya pemilihan umum. Karim (1991) mengatakan bahwa pemilihan umum merupakan salah satu saran utama untuk menegakkan tatanan demokrasi (kedaulatan rakyat), yang berfungsi sebagai alat menyehatkan dan menyempurnakan demokrasi.

Pemilihan Umum (PEMILU) sebagai pesta demokrasi di Indonesia pertama kali di Indonesia dilaksanakan pada tahun 2004 sesuai dengan Undang-Undang Dasar 1945 setelah amandemen ke-IV ditetapkan. PEMILU pertama dilakukan untuk memilih presiden dan wakil presiden, serta anggota parlemen, yaitu DPR, DPRD dan DPD. PEMILU selanjutnya dilakukan pada tahun 2009 dan 2014. Di tahun 2019 dilakukan pemilihan presiden dan wakil presiden dengan kandidat calon nomer 1 adalah pasangan Joko Widodo dan Ma'ruf Amin dan kandidat nomer 2 adalah pasangan calon Prabowo dan Sandiaga Salahudin Uno. Pada pemilihan presiden di tahun 2014, Joko Widodo dan Prabowo telah berhadapan sebagai kandidat calon presiden, tetapi dengan pasangan yang berbeda, yaitu Joko Widodo dengan Jusuf Kalla dan Prabowo dengan Hatta Rajasa. Pemilihan presiden 2014 dimenangkan pasangan Joko Widodo-Jusuf Kalla. Dari total jumlah pemlih di Indonesia sebanyak 193.944.150, sebanyak 134.953.967 menggunakan hak pilihnya. Dari daftar orang yang menggunakan hak pilihnya, terdapat suara yang memilih Joko

Widodo-Jusuf Kalla sebanyak 70.997.833, memilih Prabowo-Hatta Rajasa sebanyak 62.576.444 suara, dan suara yang tidak sah sebanyak 1.379.690 (KPU, 2014). Sementara sisa dari jumlah pemilih yang tidak menggunakan hak pilihnya terkenal sebagai pemilih golongan putih (golput) yang berarti sekelompok masyarakat yang menolak untuk memilih.

Pemilihan presiden dan wakil presiden akan dilaksanakan pada tanggal 17 April 2019. Masing-masing kandidat memiliki tujuan serta visi-misi yang dibawa untuk masa depan Indonesia. Maka dari itu, Komisi Pemilihan Umum (KPU) memberikan waktu mulai 23 September 2018 sampai 13 April 2019 untuk melakukan kampanye. Aplikasi strategi kampanye dikategorikan dalam tiga bentuk saluran media yaitu media lini atas (surat kabar, televisi, dan radio), media lini bawah (poster, plamfet, dan spanduk), dan media baru berupa sosial media (Heryanto & Rumaru, 2013). Pada Undang-Undang nomor 7 tahun 2017 tentang pemilihan umum pasal 267 menjelaskan bahwa kampanye pada PEMILU merupakan bagian dari pendidikan politik untuk masyarakat dan dilaksanakan secara bertanggung jawab. Pada undang-undang yang sama dengan pasal 275 menjelaskan bahwa salah satu sarana kampanye adalah melalui debat pasangan calon presiden 2019.

Debat merupakan proses menyusun argumen dari pernyataan yang masuk akal untuk meyakinkan lawan bicara agar menerima pendapat yang dilontarkan. Debat telah dipraktikkan pada sistem ketatanegaraan sebagai ajang untuk mendiskusikan isu-isu kemasyarakatan dan membuat resolusi dari permasalahan tersebut (Pratama, dkk., 2016). Selama periode debat presiden 2019 dilaksanakan oleh KPU sebanyak lima kali dan disiarkan langsung secara nasional oleh media elektronik melalui lembaga penyiaran publik. Dalam setiap debat mengusung tema yang berbeda-beda tetapi tetap berlandaskan pada pembukaan Undang-Undang Dasar 1945.

Debat memiliki pengaruh yang besar bagi elektabilitas pasangan calon presiden. Menurut Holbrook (1999), perdebatan akan memperkuat keyakinan pemilih dalam menilai kandidat,

terutama pada debat perama karena masyarakat kurang memiliki informasi dan masyarakat cenderung belum dimemutuskan. Pada studi kasus pemilihan presiden di Amerika Serikat tahun 2008, dua pertiga dari pengguna hak pilih merasa terbantu dalam menentukan calon yang dipilih ketika melihat perdebatan antara Barack Obama dan John McCain (Heimlich, 2012). Sedangkan menurut Jamieson & Gottfried (2010) mengkonfirmasi kekuatan debat presiden untuk meningkatkan pengetahuan pemilih. Bentrokan dua sisi dari debat mengenai ide-ide yang saling bersaing, yang tidak dimediasi oleh interpretasi dari wartawan, meningkatkan pengetahuan pemilih. Para calon presiden dan wakil presiden menerima tipu daya yang dilakukan oleh pihak lain.

Saat dalam periode kampanye, masyarakat ramai menanggapi sesi Debat Calon Presiden 2019 dengan berbagai sentimen, diantaranya positif, negatif, atau tidak memihak kubu manapun (netral). Bagi masyarakat yang aktif di dunia maya, atau biasa disebut warganet kerap menyampaikan pendapat terhadap debat presiden 2019 pada media sosial, salah satunya melalui Twitter. Pengguna Twitter dapat mengemukakan pendapatnya terhadap debat presiden dan dapat saling mengomentari melalui *tweets* yang dibatasi hingga 280 karakter. Pengguna juga dapat mengunggah foto atau video pendek sebagai sisipan informasi (Statista, 2018). Pengguna Twitter bisa terdiri dari berbagai macam kalangan dan dapat saling berinteraksi dengan merujuk satu sama lain dalam pesan menggunakan simbol @, atau yang biasa disebut dengan *mention* (Namaan, Becker, & Gravano, 2011). Tak jarang para warganet menyebutkan akun twitter pasangan Calon Presiden dengan menyebutkannya melalui *mention*. *Tweets* pada setiap pengguna Twitter dapat berpengaruh dalam pembentukan citra baik maupun citra buruk terhadap pasangan Calon Presiden. Hal tersebut dikarenakan semakin banyak suatu topik diulas dalam *tweets* para pengguna maka topik tersebut dapat menjadi *trending topic* di Twitter dan topik menjadi perbincangan hangat yang layak untuk dibahas. Dari permasalahan tersebut dapat dilakukan analisis

respons warganet terhadap Debat Calon Presiden 2019 yang dikelompokkan dengan sentimen positif, negatif, dan netral.

Metode *Clustered Support Vector Machine* (CSVM) merupakan metode klasifikasi yang berasal dari pengembangan metode Support Vector Machines (SVM). Gu & Han (2013) menjelaskan bahwa CSVM cocok untuk digunakan pada data berskala besar, dimana data dibagi menjadi beberapa klaster yang kemudian melakukan klasifikasi pada masing masing klaster dengan metode SVM. Metode CSVM pernah dilakukan oleh Rofiq (2018) yang menganalisis *profiling tweets* pada pemilihan gubernur di Provinsi Jawa Timur, Jawa Tengah, dan Jawa Barat pada tahun 2018 menggunakan metode CSVM. Data yang digunakan merupakan *tweets* yang menyebutkan nama calon pasangan gubernur dan wakil gubernur. Hasil dari penelitian ini menunjukkan bahwa pada ketiga provinsi untuk metode CSVM menggunakan kernel linear memiliki ketepakatan klasifikasi lebih baik dibandingkan dengan menggunakan kernel RBF dengan nilai akurasi dan *time process* sebesar 100%; 97,8%;98,7%; serta 7,157 detik; 1,731 detik; dan 2,329 detik.

Terkait perbandingan akurasi fungsi kernel pada metode CSVM pernah dilakukan oleh Jiang, Xie, & Zhang (2009). Penelitian tersebut membahas hasil analisis klasifikasi empat jenis data tidak berlabel menggunakan metode CSVM dengan tiga jenis kernel untuk klasifikasi SVM yaitu SVM Linear, Polinomial dan RBF. Penelitian tersebut memberikan hasil waktu komputasi sangat efisien dengan rata-rata nilai akurasi untuk keempat jenis data tersebut adalah 95,63%; 98,97%; dan 97,89%. Pada penelitian tersebut menunjukkan bahwa CSVM dengan kernel polinomial menghasilkan klasifikasi terbaik. Mayasari (2018) pernah membandingkan metode Regresi Logistik, SVM, dan Naïve Bayes Classifier (NBC) pada studi kasus *text mining* pada akun resmi pemerintah Kota Surabaya. Pada penelitian tersebut didapatkan bahwa metode SVM dengan kernel RBF menghasilkan akurasi terbaik sebesar 82%. Diani, Wisesty, & Aditsania (2017) menunjukkan bahwa pemakaian kernel pada metode SVM

bergantung pada karakteristik data yang dikenai klasifikasi. Untuk algoritma *cluster* pernah diteliti oleh Iriawan, Fithriasari, & Pravitasari (2018) dengan membandingkan metode K-*Means*, FCM, GMM dan FSSN untuk Segmentasi Tumor Otak data noise menunjukkan bahwa metode K-*Means* menunjukkan hasil cluster yang terbaik.

Pada penelitian ini data yang digunakan bersumber dari *tweets* yang menyebutkan nama akun calon pasangan Calon Presiden yang dikelompokkan berdasarkan periode pasca Debat Presiden 2019. Penelitian ini bertujuan untuk melakukan analisis terhadap respon warganet di Twitter pasca Debat Calon Presiden pada masa kampanye presiden 2019 serta sebagai bahan evaluasi dari tim sukses pendukung calon presiden dalam melaksanakan kampanye. Melalui penelitian ini diharapkan dapat diketahui karakteristik respon warganet di Twitter ke dalam sentimen positif, negatif atau netral dengan metode *Clustered Support Vector Machines* dengan algoritma *cluster* menggunakan metode K-*Means* dan membandingkan tiga fungsi kernel yaitu, *linear*, Polinomial dan RBF.

1.2 Rumusan Masalah

Mengetahui fenomena respon warganet yang ramai di Twitter terkait Debat Calon Presiden 2019 merupakan hal yang menarik untuk diulas, sehingga dilakukan analisis terkait karakteristik respon warganet berupa sentimen pasca Debat Calon Presiden 2019. Dari hasil karakteristik tersebut selanjutnya dilakukan perbandingan ketepatan hasil pengelompokan *tweets* terhadap Debat Calon Presiden 2019 dengan metode *Clustered Support Vector Machines* dengan melihat nilai akurasi, AUC dan *running time programe*. Pengelompokan berdasarkan pada sentimen positif, negatif dan netral dengan bantuan kamus *lexicon*. Selain itu dilakukan analisis terkait peran debat untuk mencerdaskan masyarakat Indonesia dengan cara membandingkan kata-kata yang sering muncul pada debat dengan tanggapan warganet di Twitter.

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah tersebut, tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut.

1. Memperoleh karakteristik berupa sentimen respons warganet terhadap Debat Presiden 2019 di Twitter.
2. Menganalisis peran Debat Calon Presiden 2019 untuk mencerdaskan masyarakat Indonesia dengan membandingkan kata-kata yang sering muncul pada debat dengan tanggapan warganet di Twitter.
3. Mendapatkan perbandingan ketepatan hasil pengelompokan respons warganet terhadap Debat Presiden 2019 dengan metode *Clustered Support Vector Machines.*

1.4 Manfaat Penelitian

Hasil penelitian diharapkan dapat bermanfaat sebagai informasi tentang respons berupa sentimen masyarakat Indonesia terkait Debat Calon Presiden 2019 di Twitter berdasarkan tema yang diusung pada masing-masing sesi debat serta mengetahui seberapa jauh pengaruh debat terhadap hasil PEMILU 2019. Dari Debat Calon Presiden 2019 dilakukan analisis terhadap fungsi debat untuk mencerdaskan masyarakat Indonesia dengan membandingkan Debat Presiden 2019 dengan tanggapan warganet via Twitter. Selain itu hasil perbandingan metode dalam penelitian ini dapat dijadikan referensi untuk penelitian dengan menggunakan metode *Clustered Support Vector Machines.*

1.5 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut.

1. Data yang digunakan merupakan t*weet* yang dipublikasikan tiga hari pasca debat Presiden 2019.
2. Penelitian hanya melakukan analisis terhadap *tweet* menggunakan Bahasa Indonesia.
3. Data *tweet* yang digunakan bersumber pada *tweet* yang menyebutkan akun Twitter pasangan Calon Presiden 2019.
4. Sentimen menggunakan bantuan kamus *lexicon sentiment* yang didapatkan dari https://github.com/masdevid/ID-

OpinionWords dan https://github.com/riochr17/Analisis-Sentimen-ID/tree/master/data

5. Model yang terbentuk pada *Clustered Support Vector Machines* adalah dengan memaksimalkan parameter C dan gamma pada tiap kernelnya.

(Halaman ini sengaja dikosongkan)

BAB II
TINJAUAN PUSTAKA

2.1 *Text Mining*

Text mining adalah sebagai suatu proses menggali informasi dimana seorang pengguna berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis (Feldman & Sanger, 2007). Definisi lain yang berkaitan dengan *text mining* merupakan suatu proses ekstraksi pola tertentu dari *database* dokumen teks yang besar yang bertujuan untuk menemukan pengetahuan (Maimon & Rokach, 2010).

2.2 *Text Preprocessing*

Text Preprocessing merupakan tahapan awal dalam *text mining*, dimana file teks mentah akan dirubah menjadi rangkai unit bahasa yang sangat jelas (Herbrich & Graepel, 2010). Berikut merupakan tahapan dari *text preprocessing*.

a. *Data Cleaning*

Data Cleaning merupakan tahapan dalam membersihkan data teks dari kata yang tidak diperlukan untuk mengurangi *noise.* Kata yang dihilangkan dalam dokumen teks antara lain karakter HTML, *emoticon, hastag (#)* dan URL.

b. *Case Folding*

Case Folding merupakan perubahan data teks menjadi teks dengan huruf kecil. Karakter selain huruf dihilangkan dan dianggap sebagai *delimiter*.

c. *Tokenizing*

Tokenizing adalah tahapan pemotongan *string input* berdasarkan pada tiap kata sebagai penyusunnya.

d. *Filetering*

Tahapan pengambilan kata-kata yang penting dari hasil token dengan membuang kata yang berada di dalam *stopword* dan menyimpan kata yang penting (*wordlist*).

e. *Stemming*

Tahapan *stemming* adalah tahap mencari akar kata dari hasil *filtering*. Pada tahap ini dilakukan proses pengambilan

berbagai bentuk kata ke dalam suatu representasi yang sama. (Feldman & Sanger.2007).

Selain itu, pada teks dari Twitter perlu dilakukan normalisasi kata, artinya penggantian kata-kata singkatan dan tidak baku menjadi kata-kata yang baku sesuai Ejaan Bahasa Indonesia (EBI).

2.3 Pembobotan Kata

Metode pembobotan kata terhadap suatu *term* menggunakan jumlah banyaknya kemunculan kata pada suatu kalimat. Pembobotan kata yang lazim digunakan adalah TF-IDF (*Term Frequency-Inverse Document Frequency*). TF-IDF merupakan pembobotan dengan membangun sebuah model vektor. TF berperan penting dalam menghitung jumlah sebuah kata dalam suatu berkas. Sedangkan IDF mengurangi kata dengan melihat frekuensi kemunculan kata pada suatu dokumen. Berikut merupakan rumus dari TF-IDF (Susanto, 2014)

$$TF_{i,j} = \frac{tf_{i,j}}{\sum_{k} tf_{i,j}} \tag{2.1}$$

$$IDF = \log \frac{|\mathrm{D}|}{|\{d : \mathrm{w}_{\mathrm{i}} \in d\}|} \tag{2.2}$$

$$W_{dt} = TF_{i,j} \times IDF \tag{2.3}$$

Keterangan :

$TF_{i,j}$ = *Term Frequency* ke-*i* pada dokumen ke-*j*

$tf_{i,j}$ = Jumlah muncul kata ke-*i* pada dokumen ke-*j*

$\sum_{k} tf_{i,j}$ = Total kata pada dokumen

$|\mathrm{D}|$ = Total dokumen

$|\{d : \mathrm{w}_{\mathrm{i}} \in d\}|$ = Jumlah dokumen (*termword*) w_i yang muncul.

2.4 Analisis Sentimen

Analisis sentimen adalah salah satu cabang penelitian dari *text mining* yang berguna untuk mengklasifikasi dokumen teks

berupa opini berdasarkan sentimen. Analisis sentimen atau opinion mining adalah studi komputasional dari opini-opini orang, sentimen, dan emosi melalui entitas dan atribut yang dimiliki dan diekspresikan dalam bentuk teks (Liu, 2012). Analisis sentimen ini dapat mengelompokkan polaritas dari teks dalam kalimat atau dokumen untuk mengetahui apakah opini pada kalimat atau dokumen tersebut apakah termasuk positif atau negatif. Pengecekan dokumen dilakukan berdasarkan pada *lexicon sentiment dictionaries* dan dihitung berapa banyak kemunculannya pada dokumen teks (Rofiqoh, 2017). Berikut merupakan contoh pengambilan kata-kata yang menunjukkan sentimen.

Tabel 2.1 Pengambilan Kata-Kata Sentimen

Kata-Kata	Sentimen Positif	Sentimen Negatif
Apa berani bang hanya berbicara saja supaya dapat simpati rakyat kalau berani itu namanya bunuh diri penyandang dana kabur	berani(2), berbicara(1), simpati(1)	bunuh(1), kabur(1)
aku tidak heran kalau orang bodoh bela orang bodoh sudah biasa	bela(1), biasa(1)	tidak(1), bodoh(2)

2.5 *Feature Selection*

Feature Selection dalam *machine learning* disebut juga variable selection (seleksi variabel) adalah proses pemilihan variabel yang relevan untuk digunakan dalam pembentukan model. Salah satu metode *feature selection* menggunakan metode pengujian *Chi-Square*, dimana *Chi-Square* sangat efektif mengurangi jumlah term tanpa mengurangi akurasi dari klasifikasi. Cara kerja pengujian *Chi-Square* dengan mengukur nilai ketergantungan antara variabel-variabel stokastik sehingga dapat menghilangkan term yang dianggap independen dari kelas yang ditentukan sehingga tidak relevan untuk metode klasifikasi (Yang & Pedersen, 1997). Berikut merupakan hipotesis pada pengujian Chi-Square.

H_0 : Tidak ada hubungan *term*/kata terhadap sentimen
H_1 : Adanya hubungan *term*/kata terhadap sentimen
Statistik uji yang digunakan sebagai berikut.

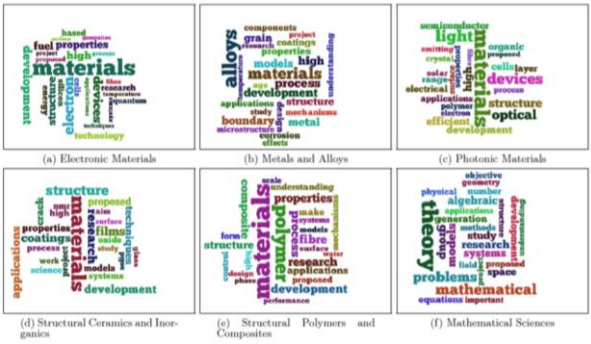
$$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - \mu_i)^2}{\mu_i} \tag{2.4}$$

$$X_i = \sum_{j=1}^{n} \mathbf{Y} W_j \tag{2.5}$$

$$\mu_i = P(Y_i) \times \sum_{j=1}^{i} W_j \tag{2.6}$$

Keterangan :
X_i = Nilai observasi kata pada sentimen ke-*i*
μ_i = Nilai harapan kata pada sentimen ke-*i*
n = Jumlah sentimen
Y = sentimen dalam bentuk dummy
W_j = bobot suatu kata di setiap dokumen.
$P(Y_i)$ = peluang kemunculan sentimen ke-*i* pada suatu dokumen.

2.6 *Wordcloud*



(sumber : Castella & Sutton, 2014)
Gambar 2.1 *Wordcloud*

Wordcloud merupakan salah satu metode yang sering digunakan untuk penggambaran data teks. *Wordcloud* merupakan

representasi grafis dari sebuah dokumen yang dilakukan dengan *plotting* kata-kata yang sering muncul pada sebuah dokumen pada ruang dua dimensi. Frekuensi dari kata yang sering muncul ditunjukkan melalui ukuran huruf kata tersebut. Semakin besar ukuran kata menunjukkan semakin besar frekuensi kata tersebut muncul dalam dokumen (Castella & Sutton, 2014).

2.7 K-*Means Clustering*

Salah satu metode analisis cluster non-hierarki adalah K-*Means Clustering*, yaitu metode untuk mengklasifikasikan atau mengelompokkan objek-objek (data) ke dalam K-grup (*cluster*) berdasarkan atribut tertentu. Data dikelompokkan dengan memperhitungkan jarak terdekat antara data-data (objek observasi) dengan pusat *cluster* (*centroid*). Prinsip utama dari metode ini adalah menyusun K buah centroid atau rata-rata (*mean*) dari sekumpulan data berdimensi N, dimana metode ini mensyaratkan nilai K sudah diketahui sebelumnya (apriori). Kedekatan suatu observasi atau variabel dengan observasi tertentu daripada dengan observasi yang lain dinyatakan dengan fungsi yang disebut jarak eqluidean (*d*). Berikut merupakan rumus menghitung jarak *(d)*.

$$d(i,k) = \sqrt{\sum_{j=1}^{p} | x_{ij} - x_{kj} |^2} \qquad (2.7)$$

dimana *d(i,k)* adalah jarak antara observasi *i* dengan *k*.

Algoritma pada metode K-*Means* Clustering adalah sebagai berikut.

1. Partisi data ke dalam K-*Cluster* sebagai inisiasi
2. Lanjutkan melalui daftar data melalui data yang terdekat dengan mean. Hitung kembali mean untuk tiap kluster yang menerima maupun kehilangan data.
3. Ulangi langkah kedua hingga tidak ada penugasan yang terjadi (Johnson, 2007).

Sebuah metode klaster membagi sebuah data dengan $X = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_n\}$ ke dalam beberapa klaster yang didefinisikan $C = \{C_1, C_2, ..., C_k\}$ dengan n merupakan jumlah data yang diklasterkan dan *k* merupakan jumlah klaster. Untuk mendapatkan klaster yang optimal dengan meminimalkan nilai *within-cluster*

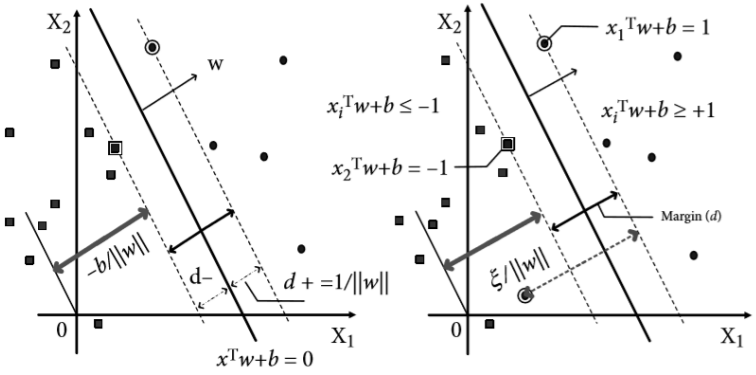
sum of square distances. Berikut merupakan rumus mendapatkan nilai *within-cluster sum of square distances*.

$$withinss = \arg\min \sum_{j=1}^{k} \sum_{x \in C} \| \bar{x}_i - \bar{\mu}_j \|^2 \qquad (2.8)$$

Dengan nilai $\bar{\mu}_j = \frac{1}{n_j} \sum_{x \in C} \bar{x}_i$ (Szkaliczki, 2016).

2.8 *Support Vector Machine*

Support Vector Machine (SVM) merupakan algortima yang cepat dan efektif dalam masalah klasifikasi teks (Feldman & Sanger. 2007). Prinsip dasar SVM adalah *linear classifier*, dan selanjutnya dikembangkan agar dapat bekerja pada *problem non-linear* dengan memasukkan konsep *kernel trick* pada ruang kerja berdimensi tinggi (Hardle, Prastyo, & Hafner, 2014). SVM terbagi menjadi SVM *Linear Separable* dan SVM *Non Linear Separable*. Berikut merupakan visualisasi dari keduanya.



(sumber : Hardle, Prastyo, & Hafner, 2014)

Gambar 2.2 Ilustrasi Metode SVM *linear Separable* (kiri) dan SVM *Non Linear Separable* (kanan)

Definisi lain menurut Yunliang, dkk (2010) mengatakan bahwa *Support Vector Machine (SVM)* adalah metode klasifikasi yang bekerja dengan cara mencari *hyperplane* dengan margin

terbesar. *Hyperplane* adalah garis batas pemisah data antar kelas. *Margin* adalah jarak antara *hyperplane* dengan data terdekat pada masing-masing kelas. Adapun data terdekat dengan *hyperplane* pada masing-masing kelas inilah yang disebut *support vector*. Dalam istilah geometri, pemisah biner SVM dapat dilihat sebagai *hyperlane* dalam ruang pemisah yang memisahkan ruang hal-hal positif dengan ruang hal-hal negatif.

2.9 *Support Vector Machine Linear Separable*

Linear separable merupakan data yang dapat dipisahkan secara linear. Misal diberikan data set $\mathbf{D} = \{x_i, y_i\}_{i=1}^{n}$ setiap observasi terdiri dari sepasang *p* prediktor $x_i = \{x_{i1}, x_{i2}, x_{i3}, \ldots x_{ip}\}$, $x_i \in R^p$ untuk $i = 1, 2, 3, \ldots, n$, dimana *n* merupakan banyak data dan label kelas dari data x_i dinotasikan $y_i \in y = \{-1, +1\}$. Jika x_i adalah anggota kelas (+1) maka x_i diberi label (target) $y_i = +1$ dan jika tidak maka diberi label (target) $y_i = -1$

Merujuk pada Gambar 2.2 (kiri) memperlihatkan beberapa titik yang merupakan anggota dari dua buah hasil klasifikasi yaitu: -1 dan +1. Berdasarkan tersebut dapat dilihat bahwa data dapat dipisahkan secara *linear*, karena garis lurus dapat ditarik untuk memisahkan semua anggota *class* -1 yang disimbolkan kotak dari *class* +1 dengan simbol lingkaran. Konsep utama SVM pada kasus *Linear Separable* adalah menetapkan pemisah *linear* (*hyperplane*). Persamaan *hiperplane* dituliskan sebagai berikut.

$$f(x) = \mathbf{x}'w + \mathrm{b} = 0 \tag{2.9}$$

dimana,

bidang pembatas kelas pertama : $\mathbf{x_i}'w + \mathrm{b} \geq +1$ untuk $\mathbf{y_i} = +1$

bidang pembatas kelas kedua : $\mathbf{x_i}'w + \mathrm{b} \leq -1$ untuk $\mathbf{y_i} = -1$.

Ddengan mengkombinasikan dua persamaan sebelumnya didapatkan persamaan baru.

$$y_i(\mathbf{x_i}'w + \mathrm{b}) - 1 \geq 0 \tag{2.10}$$

Nilai *w* merupakan vektor bobot yang berukuran ($p \times 1$), *b* adalah nilai bias yang bernilai skalar. Pada Gambar 2.2 (kiri) menunjukkan $\frac{|\mathrm{b}|}{||w||}$ merupakan jarak bidang pemisah yang tegak

lurus dari titik pusat koordinat dan $\| w \|$ adalah jarak *euclidean* dari *w* (margin). Secara matematis, optimasi SVM *Linear Separable* dengan memaksimalkan ||**w**|| dengan cara sebagai berikut.

$$\min_{w} \frac{1}{2} \| w \|^2 \tag{2.11}$$

Merujuk persamaan 2.10 dan untuk mendapatkan nilai optimasi pada persamaan 2.11 lebih mudah dilakukan dengan *Lagrange Multiplier* sebagai berikut.

$$\min_{w,b} L_p(w,b) = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{n} \alpha_i [y_i (\mathbf{x_i}' w + \mathrm{b}) - 1] \tag{2.12}$$

Dengan nilai $\alpha_i \geq 0$. Untuk meminimalkan *L* terhadap **w** dan *b* dilakukan sehingga didapatkan persamaan berikut.

$$\min_{\alpha} L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{2.13}$$

Titik ke-*i* dimana nilai $y_i (\mathbf{x_i}' w + \mathrm{b}) = 1$ berlaku dinamakan *support vector*. Selanjutnya, dan didapatkan aturan klasifikasi sebagai berikut.

$$g(x) = sign(\mathbf{x}' w + \mathrm{b}) \tag{2.14}$$

dimana $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ (Hardle, Prastyo, & Hafner, 2014).

2.10 *Support Vector Machine Non-Linear Separable* dengan Metode Kernel

Non Linear Separable merupakan data yang tidak dapat dipisahkan secara linear. Berdasarkan Gambar 2.2 (kanan) untuk kasus data yang tidak bisa dipisahkan secara *linear*, perlu ditambahkan batasan berupa variabel *slack* ξ_i yang menunjukkan pelanggaran dari kesalahan klasifikasi terkait dengan jarak titik kesalahan dari *hiperplane*. Dengan penambahan nilai $\xi_i \geq 0$ pada fungsi SVM dinamakan sebagai fungsi *soft margin* dan mengacu pada persamaan 2.10 berubah menjadi sebagai berikut.

$$y_i(\mathbf{x_i}'w + b) \geq 1 - \xi_i \tag{2.15}$$

Untuk mengoptimalkan nilai margin dengan persamaan sebagai berikut.

$$\min_{w,\xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i \tag{2.16}$$

Dimana C merupakan paremeter yang untuk mengontrol *trade off* antara margin dan error klasifikasi. Nilai C yang besar akan memberikan pinalti yang lebih besar terhadap *error* klasifikasi. Merujuk pada persamaan 2.13 dan 2.16 didapatkan fungsi *Lagrange Multiplier* sebagai berikut.

$$\begin{aligned} \min_{w,b,\xi} L_p(w,b,\xi) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \mu_\partial \xi_i \\ - \sum_{i=1}^{n} \alpha_i [y_i(\mathbf{x_i}'w + b) - 1 + \xi_i] \end{aligned} \tag{2.17}$$

Dengan nilai $\alpha_i \geq 0$ dan $\mu_i \geq 0$ Setelah meminimalkan *L* terhadap *w***,** *b* dan variabel *slack*, maka didapatkan persamaan *dual problem* sebagai berikut.

$$\min_{\alpha} L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{2.18}$$

dimana nilai $0 \leq \alpha_i \leq C$ dan $\sum_{i=1}^{n} \alpha_i y_i = 0$.

Pada kasus non linear, SVM perlu dimodifikasi agar dapat menemukan solusi dengan metode kernel. Penggunaan kernel bertujuan untuk mengimplementasikan suatu model pada ruang dimensi yang lebih tinggi (*feature space*) sehingga kasus yang *non linear separable* pada ruang input bisa ditransformasi menjadi *linear separable* pada *feature space*. Suatu data *x* di *input space* ke *feature space* dengan menggunakan fungsi transformasi $x_i \rightarrow \phi(x_i)$. Sehingga fungsi klasifikasi yang dihasilkan ditujukkan sebagai berikut.

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i)' \phi(x_j) + \mathrm{b} \qquad (2.19)$$

Selain itu, sulit mengetahui fungsi transformasi yang tepat, sehingga digunakan solusi *Kernel Trick*. Pada persamaan (2.19) dapat dilihat *dot product* $\phi(x_i)'\phi(x_j)$. Jika terdapat fungsi kernel K sehingga $K(x_i, x_j) = \phi(x_i)'\phi(x_j)$, maka fungsi transfomasi $\phi(x_i)$ tidak perlu diketahui secara persis sehingga dihasilkan persamaan sebagai berikut.

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x_j) + \mathrm{b} \qquad (2.20)$$

Syarat sebuah fungsi menjadi fungsi kernel adalah memenuhi teorema Mercer yang menyatakan bahwa matriks Kernel yang dihasilkan harus bersifat *positive semi definite*. Fungsi kernel yang umum digunakan pada metode SVM ditunjukkan pada tabel berikut.

Tabel 2.2 Fungsi Kernel pada SVM

Fungsi Kernel	Tema
Linear	$K(x_i, x_j) = x_i' x_j + C$
Radial Basis Function	$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \gamma > 0$
Polinomial	$K(x_i, x_j) = \left(\gamma x_i' x_j + r\right)^p, \gamma > 0$

2.11 *Clustered Support Vector Machines*

Data Algoritma *Clustered Support Vector Machiness* merupakan pengembangan dari metode SVM untuk meangani *data training* dalam skala besar. Dalam pendekatan satu level CSVM, *data training* pertama kali dibagi ke dalam *k* klaster dengan algoritma K-*Means*, dimisalkan $\{C_1,\ldots,C_k\}$. Metode SVM diperlakukan pada masing-masing klaster dengan variabel (x_i^l, y_i^l) dengan $i = 1,...,n_l$ sebagai contoh indeks cluster ke-*l*, dimana

$l = 1,...,k$ merupakan banyaknya pengamatan pada klaster ke-*l* dan digunakan untuk memodelkan hubungan *non linear* dalam satu klaster. Sehingga model klasifikasi akhir didefinisikan sebagai berikut.

$$f(x) = \sum_{i=1}^{k} x' w_i \ \mathbf{1}\ (x \in C_l) \qquad (2.21)$$

Dimana **1**(.) merupakan fungsi indikator. Pada kasus ini tidak menggunakakn bias *b*, sehingga algoritma CSVM ditulis pada persamaan 2.21.

$$\min_{w,w_l,\xi_i^l \geq 0} \frac{\lambda}{2} \| w \|^2 + \frac{1}{2} \sum_{i=1}^{k} \| w_i - w \|^2 + C \sum_{i=1}^{k} \sum_{i=1}^{n_l} \xi_i \qquad (2.22)$$

Dimana fungsi kendala : $y_i^l w_l^{'} x_i^l \geq 1 - \xi_i^l, i = 1,...,n_l, \forall l$ dengan *w* merupakan vector bobot referensi global, $\frac{1}{2} \sum_{i=1}^{k} \| w_i - w \|^2$ merupakan global regularization, dimana membutuhkan membutuhkan bobot linier lokal SVM (w_l) sejalan dengan bobot referensi global. *w* menjembatani di antara kelompok yang berbeda, sehingga informasi dari satu klaster dapat dimanfaatkan ke yang lain. Oleh karena itu, hal tersebut dapat menghindari *overfitting* di setiap klaster lokal. Secara khusus, jika kita menetapkan $\boldsymbol{w} = 0$, maka CSVM akan menjadi *k* SVM independen yang dilatih di setiap *cluster* secara terpisah.

Ketika $v_i = w_i - w$ dengan $w_i = v_i + w$, sehingga dari persamaan 2.22 dapat ditulis sebagai berikut.

$$\min_{w,w_l,\xi_i^l \geq 0} \frac{\lambda}{2} \| w \|^2 + \frac{1}{2} \sum_{i=1}^{k} \| v_i \|^2 + C \sum_{i=1}^{k} \sum_{i=1}^{n_l} \xi_i \qquad (2.23)$$

Dimana fungsi kendala : $y_i^l w_l^{'} x_i^l \geq 1 - \xi_i^l, i = 1,...,n_l, \forall l$.

Untuk menyederhanakan optimasi tersebut, didefinisikan $w=[\sqrt{\lambda}w', \mathrm{v}_1', ... \mathrm{v}_k']'$ dan $x_i^l = [\frac{1}{\sqrt{\lambda}} x_i^l', 0', ... \mathrm{v}\, x_i^l', 0']'$ dimana setiap nilai ke-l + 1 dari x_i^l adalah x_i^l . Sehingga permasalahan optimasi pada persamaan 2.23 ditulis sebagai berikut.

$$\min_{w, w_l, \xi_i^l \geq 0} \frac{\lambda}{2} \| w \|^2 + C \sum_{i=1}^{k} \sum_{i=1}^{n_l} \xi_i \qquad (2.24)$$

Dimana fungsi kendala : $y_i^l w_l' x_i^l \geq 1 - \xi_i^l, i = 1, ..., n_l, \forall l$.

Sehingga untuk solusi *dual problem* mengacu pada persamaan 2.18 (Gu & Han, 2013).

2.12 Ketepatan Klasifikasi

Model klasifikasi yang telah diperoleh, selanjutnya akan diukur performa ketepatannya dalam melakukan klasifikasi. *Confusion matrix* merupakan salah satu teknik yang berguna untuk mengukur performa dari sebuah algoritma klasifikasi (Han & Pei, 2012). Berikut ini adalah tabel *confusion matrix* untuk tiga kelas klasifikasi dengan kelas dengan *tweet* yang berisi sentimen positif dan Non Positif menggambarkan sentimen negatif dan netral.

Tabel 2.3 *Confusion Matrix*

Kelas Aktual	**Kelas Prediksi**	
	Positif	**Non Positif**
Positif	TP	FN
Non Positif	FP	TN

TP adalah *True Positive*, FP adalah *False Positive*, TN adalah *True Negative*, dan FN adalah *False Negative*. Terdapat dua jenis pengukuran yang sering digunakan dalam menghitung ketepatan klasifikasi yaitu *accuracy*, *sensitivity*, dan *specisifity*. Akurasi adalah proporsi jumlah total prediksi yang benar. *Accuracy* digunakan untuk menghitung ketepatan klasifikasi sebuah dokumen yang mempunyai data yang *balanced* pada tiap

kategorinya. *Sensitivity* proporsi kasus positif yang diidentifikasi dengan benar. Sedangkan *Specificity* adalah proporsi kasus yang relevan dari semua prediksi berdasarkan kelas positif. Berikut rumus untuk ketepaatan klasifikasi.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.25}$$

$$sensitivity = \frac{TP}{TP + FN}$$

$$Specisifity = \frac{TN}{TN + FP}$$

Selain itu untuk melihat ketepatan klasifikasi dapat menggunakan AUC (*Area Under Curve*). AUC digunakan untuk mengukur kualitas *classifier* probabilistik. AUC yang ditetapkan dengan threshold 1 dapat dihitung dengan rumus sebagai berikut.

$$AUC = \max(0,5; \frac{1}{2}(sensitifity + specisifity) \tag{2.26}$$

Untuk kasus *multiclass* dengan *c* adalah jumlah *multiclass*, ,maka AUC *multiclass* didapatkan dengan menghitung setiap AUC pada masing-masing jenis klasifikasi dan didapatkan AUC akhir sebagai berikut (Hand & Till, 2001).

$$AUC_{multi} = \frac{2}{c(c-1)} \sum_{i=1}^{c} AUC_i \tag{2.27}$$

Level pengukuran kualitas *classifier* AUC menggunakan nilai ROC dilihat berdasarkan akurasi dengan rentang yang diperlihatkan pada tabel berikut (Gorunescu, 2011).

Tabel 2.4 Nilai Kualitas ROC

Rentang Nilai	**Kualitas *Classfier***
1,00-0,90	*Excelent*
0,90-0,80	*Good*
0,80-0,70	*Fair*
0,70-0,60	*Poor*
0,60-0,50	*Failure*

2.13 Pemilihan Presiden 2019

Pemilihan Presiden 2019 merupakan ajang demokrasi sebagai rakyat dalam memilih pemimpin yang tepat selama lima tahun periode mendatang. Pemilihan presiden pertama kali dilakukan pada tahun 2004 yang kemudian kembali dilaksanakan di tahun 2009 dan 2014. Pada pemilihan presiden 2019 terdapat dua calon kandidat presiden-wakil presiden. Kandidat nomor 1 merupakan pasangan Joko Widodo-Ma'ruf Amin sedangkan kandidat calon prsiden nomor 2 adalah pasangan Prabowo Subianto-Sandiga Uno. Kandidat nomor 1 didukung oleh berbagai partai politik, diantaranya PDIP, Golongan Karya, Nasional Demokrasi, PKB, PPP, Hanura, PKPI, PSI, dan Perindo. Sedangkan kandidat nomor 2 didukung oleh partai politik Gerindra, Demokrat, PAN, PKS, dan Bekarya.

2.14 Debat Calon Presiden 2019

Debat merupakan proses menyusun argumen dari pernyataan yang masuk akal untuk meyakinkan lawan bicara dengan tujuan untuk menerima pendapat yang dilontarkan. Debat telah dipraktikkan pada sistem ketatanegaraan sebagai ajang untuk mendiskusikan isu-isu kemasyarakatan dan membuat resolusi dari permasalahan tersebut. Terdapat berbagai macam jenis debat dalam dunia professional, yakni sebagai berikut (Pratama, dkk., 2016).

a. Debat Parlemen
   Dalam debat parlemen, anggota debat mengusulkan rancangan undang-undang dan membuat resolusi yang akan menjadi hukum atau undang-undang.
b. Debat Kompetitif
   Dalam debat yang bersifat kompetitif, terdapat tim yang bersaing dan pemenangnya dinilai berdasarkan kriteria tertentu dari juri.
c. Debat Antar Kandidat
   Dalam sistem kenegaraan yang mengutamakan pemilihan langsung, para kandidat untuk jabatan politik tinggi seperti

presiden atau perdana menteri akan melakukan debat publik selama masa kampanye berlangsung.

Pada Undang-Undang Nomor 7 tahun 2017, di pasal 277 menyatakan bahwa selama periode debat presiden 2019 dilaksanakan oleh KPU sebanyak lima kali dan disiarkan langsung secara nasional oleh media elektronik melalui lembaga penyiaran publik. Debat dipimpin oleh moderator yang tidak memihak kepada salah satu pasangan calon. Dalam setiap debat mengusung tema yang berbeda-beda tetapi tetap berlandaskan pada pembukaan Undang-Undang Dasar 1945 pada alinea keempat:

a. Melindungi segenap bangsa Indonesia dan seluruh tumpah darah Indonesia;
b. Memajukan kesejahteraan umum;
c. Mencerdaskan kehidupan bangsa;
d. Ikut melaksanakan ketertiban dunia yang berdasarkan kemerdekaan, perdamaian abadi dan keadilan sosial.

Pada Debat Presiden 2019 memiliki lima kali sesi debat yang memiliki topik yang berbeda-beda pada tiap sesinya. Berikut merupakan tema pada setiap debat presiden 2019.

Tabel 2.5 Tema Debat Presiden 2019

Debat	Tema
I	Hukum, HAM, Korupsi, dan Terorisme
II	Energi, Pangan, Infrastruktur, Sumber Daya Alam, dan Lingkungan Hidup
III	Pendidikan, Kesehatan, Ketenagakerjaan, Sosial dan Budaya
IV	Ideologi, Pemerintahan, Keamanan, dan Hubungan Internasional
V	Ekonomi dan Kesejahteraan Sosial, Keuangan, Investasi, dan Industri

*sumber : https://nasional.kompas.com/read/2018/12/19/17590871/ini-jadwal-debat-pilpres-2019-dari-tanggal-hingga-tema

2.15 Twitter

Twitter adalah jejaring sosial dan layanan *microblogging*, yang memungkinkan pengguna terdaftar untuk membaca dan

mengirim pesan singkat, yang disebut *tweet*. Pesan Twitter dibatasi hingga 280 karakter dan pengguna juga dapat mengunggah foto atau video pendek. (Statista, 2018). Twitter menawarkan dua alat untuk membuat saling terkoneksi satu sama lain antar penggguna: *mention*(@) dan *hastags*(#). *Mention*(@) memungkinkan pengguna untuk menandai pengguna tertentu dalam *tweets*. *Hashtag* digunakan untuk memulai, dan berpartisipasi dalam *platform* kelompok percakapan. Twitter memberikan fasilitas berupa *trending topic* apabila suatu topik sering dibicarakan oleh warganet pada Twitter bisa terlihat apa topiknya.

2.16 Akun Twitter Calon Presiden

Calon presiden dan wakil presiden 2019 masing-masing memiliki akun Twitter sebagai media sosial sebagai sarana kampanye. Kandidat nomor 1 yaitu Joko Widodo dan Ma'ruf Amin memiliki akun Twitter dengan nama @Jokowi dan @KH_MarufAmin. @Jokowi tergabung pada Twitter semenjak September 2011. Tercatat pada bulan Februari 2019 @Jokowi memiliki *followers* sebanyak 11 juta dan memposting *tweets* sebanyak 1.509. Sementara @KH_MarufAmin baru bergabung di twitter September 2018 dan memiliki *followers* sebanyak 15,1 ribu dan *tweets* sebanyak 988. Kandidat nomer 2 yaitu Prabowo dan Sandiaga Uno memiliki akun Twitter dengan nama @Prabowo dan @Sandiuno. @Prabowo telah bergabung semenjak Mei 2009, memiliki *followers* sebanyak 3,65 juta dan mempublikasikan *tweets* sebanyak 8.835. sedangkan akun @Sandiuno bergabung pada Twitter sejak April 2010, memiliki *followers* sebanyak 1,23 juta dan telah mempublikasikan *tweets* sebanyak 31,7 ribu.

BAB III
METODOLOGI PENELITIAN

3.1 Sumber Data

Data yang digunakan dalam penelitian ini diambil dengan bantuan Twitter API (*Application Programming Interface*). Data merupakan *tweets* dari warganet, yaitu pengguna Twitter di Indonesia dengan *tweets* yang menyebutkan nama akun Calon Presiden Indonesia 2019, yakni pasangan Calon Presiden nomor 1 (@Jokowi dan @KH_MarufAmin, dan pasangan Calon Presiden nomor 2 (@Prabowo, dan @Sandiuno). Data diambil selama tiga hari pasca Debat Calon Presiden 2019. Data yang diambil pada periode tersebut dibagi berdasarkan pasca dilaksanakannya Debat Calon Presiden 2019. Berikut merupakan periode pengambilan data.

Tabel 3.1 Data Berdasarkan Pasca Debat Calon Presiden 2019

Debat	Periode Data
I	18 Januari - 20 Februari
II	17 Februari – 19 Maret
III	17 Maret – 19 Maret
IV	30 Maret – 1 April
V	13 Maret – 15 Maret

3.2 Variabel Penelitian

Data yang akan dikelompokkan merupakan banyaknya kata-kata yang telah dibobotkan dengan TF-IDF dan telah diambil kata-kata yang mengandung sentimen, baik sentimen positif maupun negatif. Data tersebut dikelompokkan menjadi tiga sentimen, yaitu sentimen positif, senitimen negatif dan tidak memihak manapun (netral).

3.3 Struktur Data

Berikut merupakan struktur data pada penelitian berdasarkan *keyword*-nya sebelum dilakukan *pre-processing.*

Tabel 3.2 Struktur Data Sebelum Pre-Processing

No	Mention Akun Calon	Tanggal	Akun Warganet	*Tweets*
1	@Jokowi	D_1	A_1	T_1
2		D_2	A_2	T_2
…		…	…	…
n_1		D_{n_1}	A_{n_1}	T_{n_1}
1	@KH_MarufAmin	D_1	A_1	T_1
2		D_2	A_2	T_2
…		…	…	…
n_2		D_{n_2}	A_{n_2}	T_{n_2}
1	@Prabowo	D_1	A_1	T_1
2		D_2	A_2	T_2
…		…	…	…
n_3		D_{n_3}	A_{n_3}	T_{n_3}
1	@Sandiuno	D_1	A_1	T_1
2		D_2	A_2	T_2
…		…	…	…
n_4		D_{n_4}	A_{n_4}	T_{n_4}

Pada Tabel 3.2, menunjukkan jumlah data *tweets* warganet yang menyebutkan akun calon presiden dan wakil presiden. Kolom warganet merupakan akun yang membuat *tweets*. Kolom *tweets* menunjukkan isi *tweets* yang dipublikasikan.

Tabel 3.3 Struktur Data Setelah Pre-Processing

Periode	No	Paslon 1	Paslon 2	Kata *1*	Kata *2*	…	Kata *p*
I	1	f_{111}	f_{112}	f_{115}	f_{116}	...	f_{11p}
	2	f_{121}	f_{122}	f_{125}	f_{126}	...	f_{12p}
	…	…	…	…	…	…	…

Tabel 3.4 Struktur Data Setelah *Pre-Processing* (Lanjutan)

Periode	**No**	**Paslon 1**	**Paslon 2**	**Kata *1***	**Kata *2***	**…**	**Kata *p***
	n_1	$f_{1n_1 1}$	$f_{1n_1 2}$	$f_{1n_1 5}$	$f_{1n_1 6}$	…	$f_{1n_1 p}$
II	1	f_{211}	f_{212}	f_{215}	f_{216}	…	f_{21p}
	2	f_{221}	f_{222}	f_{225}	f_{226}	…	f_{22p}
	…	…	…	…	…	…	…
	n_2	$f_{1n_2 1}$	$f_{1n_2 2}$	$f_{1n_2 5}$	$f_{1n_2 6}$	…	$f_{1n_2 p}$
…	…	…	…	…	…	…	…
V	1	f_{511}	f_{512}	f_{515}	f_{516}	…	f_{51p}
	2	f_{521}	f_{522}	f_{525}	f_{526}	…	f_{52p}
	…	…	…	…	…	…	…
	n_2	$f_{1n_5 1}$	$f_{1n_5 2}$	$f_{1n_5 5}$	$f_{1n_5 6}$	…	$f_{1n_5 p}$

Struktur data pada Tabel 3.3 akan digunakan untuk analisis menggunakan metode *Clustered Support Vector Machines.*

3.4 Langkah Analisis

Langkah analisis digunakan untuk menggambarkan langkah-langkah penelitian yang akan dilakukan secara urut. Langkah analisis yang digunakan adalah sebagai berikut.

a. Mengambil data *tweet* dengan bantuan Twitter API berdasarkan keyword yang ditentukan.
   i. *Crawling* data dengan memasukkan *keyword searching* nama akun twitter calon presiden dan wakil presiden, yaitu @Jokowi, @KH_MarufAmin, @Prabowo, dan @Sandiuno.
   ii. Menyimpan hasil *crawling* data dari keyword yang ditentukan dalam format csv.
b. Menyiapkan data *tweets*. dan data *stopwords*
c. *Pre-processing* Data
   i. Melakukan *cleaning data*, yakni membersihkan tweet dari kata yang tidak diperlukan untuk mengurangi

noise. Kata yang dihilangkan dalam *tweets* adalah karakter HTML, *emoticons*, hastag(#), *username* (@username), simbol *retweet* (response *tweets*) "RT", dan *link* URL.

ii. Melakukan *case folding*, yaitu mengubah semua teks dengan huruf kecil (non kapital) serta menghilangkan tanda baca.

iii. Melakukan normalisasi kata-kata pada tweets sesuai dengan Ejaan Baku Indonesia (EBI)

iv. Melakukan *tokenizing* untuk memecah *tweet* menjadi kata per kata.

v. Melakukan *filtering* data dengan membuang kata pada data *stopwords*.

vi. Melakukan *stemming* dengan Bahasa Indonesia untuk mendapatkan kata dasar.

d. Membobotkan kata-kata dengan metode TF-IDF

e. Menentukan label *tweets* dengan bantuan kamus *lexicon* dan melakukan visualisasi data *tweets* berdasarkan respons warganet dengan *wordcloud.*

f. Melakukan *feature selection* pada kata-kata hasil TF-IDF berdasarkan pengujian *chi square*.

g. Membangun model *Clustered Support Vector Machines.*

i. Melakukan partisi data menjadi *data training* dan *data testing*.

ii. Mengelompokkan kata-kata dengan algoritma K-*Means.*

iii. Menentukan pembobot parameter pada SVM tiap jenis kernel.

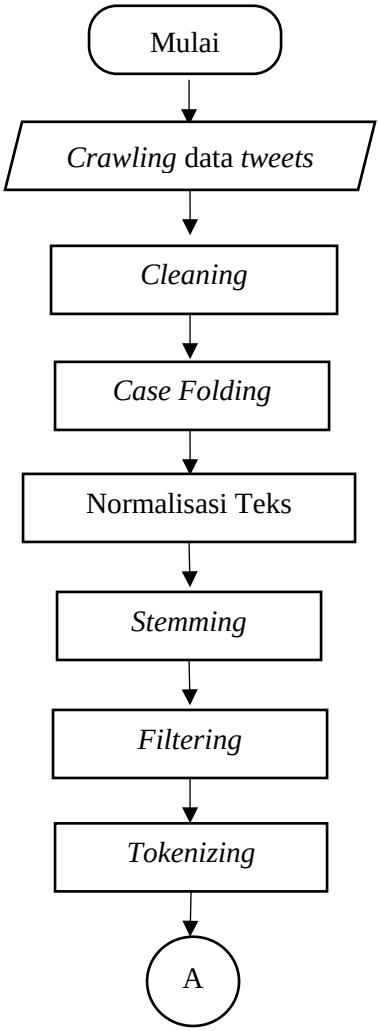
iv. Membangun model SVM dengan fungsi kernel linear, polinomial, dan RBF.

v. Membandingkan ketepatan klasifikasi dan *running time* antar fungsi kernel linear, polinomial, dan RBF.

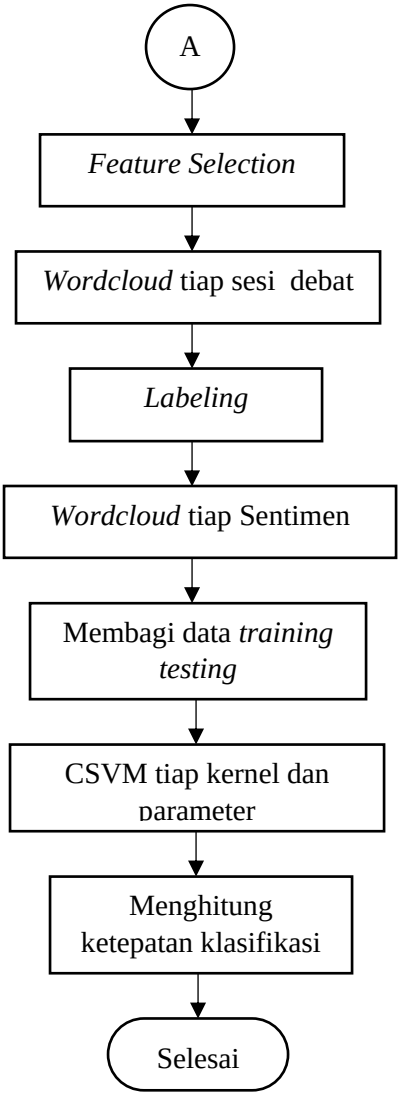
h. Membandingkan kata-kata yang sering muncul pada debat presiden 2019 dengan tanggapan warganet di Twitter.

3.5 Diagram Alir

Dengan langkah analisis yang dijabarkan sebelumnya, disusun diagram alir yang disajikan pada Gambar 3.1



Gambar 3.1 Diagram Alir



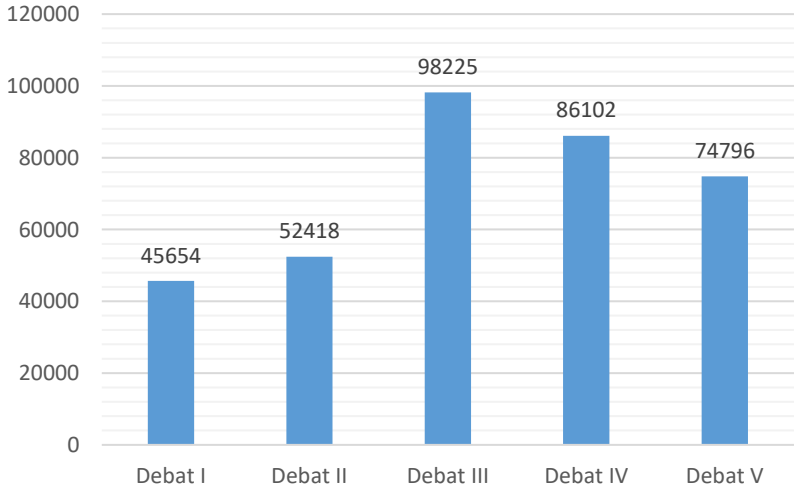
Gambar 3.2 Diagram Alir (Lanjutan)

BAB IV
ANALISIS DAN PEMBAHASAN

Analisis yang dilakukan dalam bab ini meliputi karakteristik *tweets* warganet terkait respon terhadap debat Presiden 2019 dan mengelompokkan *tweets* tersebut berdasasarkan sentimen positif, negatif dan netral dan dilakukan klasifikasi menggunakan metode *Clustered Support Vector Machines.*

4.1 Statistika Deskriptif Data Tweets

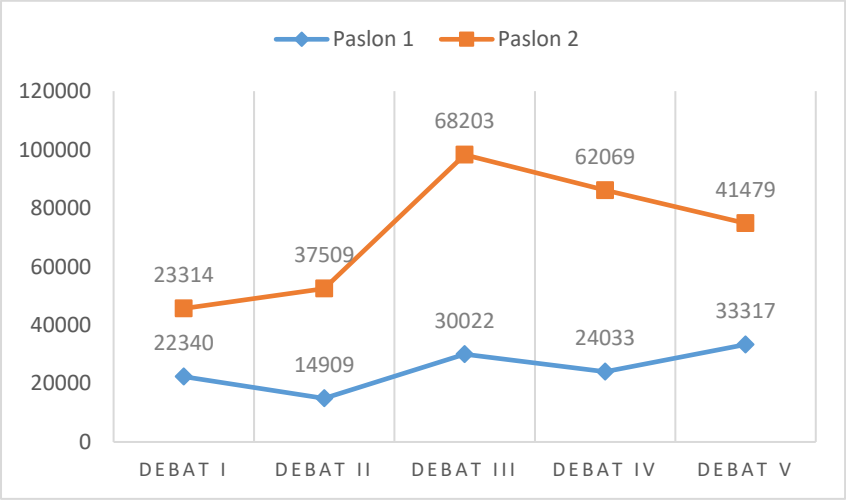
Hasil *crawling* data twiter diambil pada pasca debat Presiden 2019 digunakan untuk mengetahui respon warganet terhadap debat Presiden 2019. Debat dilakukan sebanyak lima kali. Tweets yang diambil merupakan *tweets* yang memention salah satu akun pasangan calon presiden 2019. Berikut merupakan banyak *tweets* yang terkumpul selama pasca debat presiden 2019.



Gambar 4.1 Banyak *Tweets* Pasca Debat Presiden 2019

Gambar 4.1 menunjukkan banyak *tweets* yang dikumpulkan pasca Debat Presiden 2019. Tweets terbanyak dikumpulkan pada periode debat III sebanyak 98.225. Hal tersebut menunjukkan

antusiasme warganet terhadap debat III antara calon wakil presiden Sandiaga Uno dan Ma'ruf Amin dengan tema Pendidikan, Kesehatan, Ketenagakerjaan, Sosial dan Budaya lebih lebih tinggi daripada debat yang lainnya. Berikut merupakan jumlah tweets yang memention pasangan calon presiden 2019.



Gambar 4.2 Banyak *Tweets* yang Memention Paslon Presiden 2019

Berdasarkan Gambar 4.2, terlihat bahwa warganet yang memention pasangan calon presiden nomor 2, yaitu Prabowo-Sandiaga Uno lebih banyak daripada pasangan calon nomer 1, yaitu Jokowi-Ma'ruf Amin pada semua sesi debat presiden 2019. Perbedaan jumlah *tweets* terjauh pada pasca debat III dimana selisih *tweets* antara pasangan calon 2 dan pasangan calon 1 sebanyak 38.181 buah. *Tweets* tersebut dianalisis berdasarkan sentimen positif, negatif, dan netral pada tiap sesi debat.

3.2 *Preprocessing Tweets*

Pada Data *tweets* mengenai respon warganet terhadap debat presiden 2019 yang terkumpul dilakukan preprocessing *tweets* yang meliputi *cleaning*, *case folding*, normalisasi teks, *stemming*, *filtering* dari *stopword*, dan *tokenizing*. Penjelasan mengenai hasil *prepocessing tweets* dari setiap tahapan akan dijabarkan pada simulasi *praproses teks*

pada sebuah data. Data berupa *tweets* yang terdapat pada periode debat II yang memention kedua kubu pasangan calon. Berikut merupakan tahapan proses *preprocessing tweets.*

Tabel 4.1 Tahapan *Preprocessing Tweets Cleaning*

Sebelum	Sesudah
@pascabowo @budimandjatmiko @AmaliaMalikrid1 @addiems @prabowo @sandiuno Berarti visi misi yg lama hoax. Laporkan #02GagapUnicorn #DebatPintarJokowi #JokowiMenangDebat #JokowiOrangnyaBaik	Berarti visi misi yg lama hoax. Laporkan
@saididu @jokowi Kalau ada 5 point diatas gak ada bantahan, ya wajar aja kalo Unicorn gak bakalan ngerti juga. Atau barangkali takut kena sksk mat lagi...protes bagi2 tanah ke rakyat kecil gak tau nya yg protes miliki tanah 3400rb ha.	Kalau ada 5 point diatas gak ada bantahan, ya wajar aja kalo Unicorn gak bakalan ngerti juga. Atau barangkali takut kena sksk mat lagi...protes bagi2 tanah ke rakyat kecil gak tau nya yg protes miliki tanah 3400rb ha.

Tabel 4.1 di sebelah kiri menunjukkan tweet mentah yang merupakan hasil crawling dari Twitter, sedangkan di sebelah kanan menunjukkan *tweets* yang sudah mengalami proses *cleaning.* Proses cleaning merupakan proses pembersihan *tweets* dengan tujuan untuk mengurangi *noise* pada data. Proses *cleaning* terdiri dari penghapusan symbol RT, hastag (#), penghapusan *username*.

Selanjutnya dilakukan proses case folding, dimana tweet yang memiliki huruf kapital akan akan diganti menjadi huruf kecil. Pada proses case folding juga menghilangkan tanda baca dan angka pada tweet karena dianggap noise untuk data. Berikut pada tabel 4.2 merupakan proses *case folding.*

Tabel 4.2 Tahapan *Preprocessing Tweets Case Folding*

Sebelum	Sesudah
Berarti visi misi yg lama hoax. Laporkan	berarti visi misi yg lama hoax laporkan
Kalau ada 5 point diatas gak ada bantahan, ya wajar aja kalo Unicorn gak bakalan ngerti juga. Atau barangkali takut kena sksk mat lagi...protes bagi2 tanah ke rakyat kecil gak tau nya yg protes miliki tanah 3400rb ha.	kalau ada point diatas gak ada bantahan ya wajar aja kalo unicorn gak bakalan ngerti juga atau barangkali takut kena sksk mat lagi protes bagi tanah ke rakyat kecil gak tau nya yg protes miliki tanah rb ha

Langkah selanjutnya pada preprocessing data adalah normalisasi kata, dimana semua kata yang terindikasi sebagai kata slang dan kata-kata yang disingkat akan dirubah menjadi kata-kata baku berbahasa Indonesia. Berikut merupakan hasil proses dari normalisasi kata.

Tabel 4.3 Tahapan *Preprocessing Tweets* Normalisasi

Sebelum	Sesudah
berarti visi misi yg lama hoax laporkan	berarti visi misi yang lama hoax laporkan
kalau ada point diatas gak ada bantahan ya wajar aja kalo unicorn gak bakalan ngerti juga atau barangkali takut kena sksk mat lagi protes bagi tanah ke rakyat kecil gak tau nya yg protes miliki tanah rb ha	kalau ada point diatas tidak ada bantahan ya wajar saja kalau unicorn tidak bakalan mengerti juga atau barangkali takut kena sksk mat lagi protes bagi tanah ke rakyat kecil tidak tahu nya yang protes miliki tanah rb ha

Setelah dilakukan proses normalisasi, tahapan preprocessing data selanjutnya adalah stemming. Tahapan stemming merupakan proses merubah kata-kata menjadi bentukan kata dasar. Stemming yang digunakan menggunakan modul sastrawi yang prosesnya bisa dilihat pada tabel 4.4 berikut.

Tabel 4.4 Tahapan *Preprocessing Tweets Stemming*

Sebelum	Sesudah
berarti visi misi yang lama hoax laporkan	berarti visi misi yang lama hoax lapor
kalau ada point diatas tidak ada bantahan ya wajar saja kalau unicorn tidak bakalan mengerti juga atau barangkali takut kena sksk mat lagi protes bagi tanah ke rakyat kecil tidak tahu nya yg protes miliki tanah rb ha	kalau ada point diatas tidak ada bantah ya wajar saja kalu unicorn tidak bakalan mengerti juga atau barangkali takut kena sksk mat lagi protes bagi tanah ke rakyat kecil tidak tahu nya yg protes milik tanah rb ha

Tahapan selanjutnya dari *preprocessing* data adalah filtering. Tahapan filtering dilakukan untuk menghilangkan kata-kata yang dianggap tidak bermakna. Kata-kata tersebut dimasukkan dalam stopwords. Berikut merupakan bentuk perubahan dari tahapan *filtering*.

Tabel 4.5 Tahapan *Preprocessing Tweets* Filtering

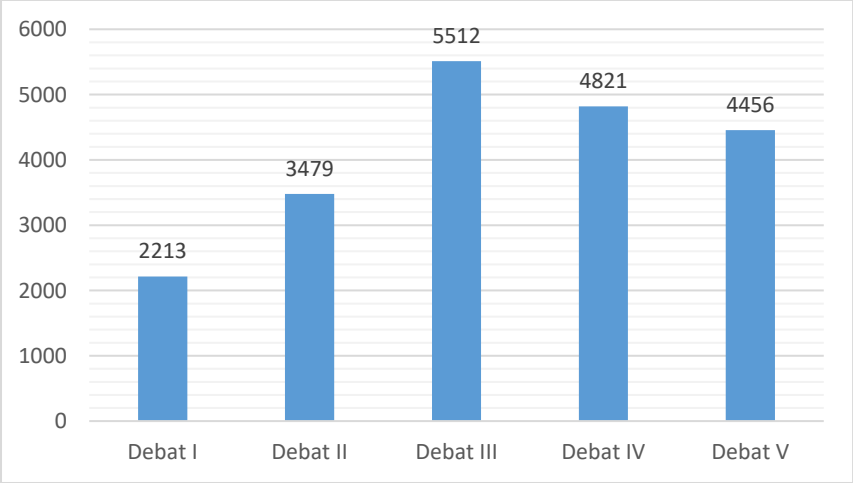
Sebelum	Sesudah
berarti visi misi yang lama hoax lapor	visi misi lama hoax lapor
kalau ada point diatas tidak ada bantah ya wajar saja kalau unicorn tidak bakalan mengerti juga atau barangkali takut kena sksk mat lagi protes bagi tanah ke rakyat kecil tidak tahu nya yg protes milik tanah rb ha	point diatas tidak bantah wajar unicorn tidak bakalan mengerti barangkali takut sksk protes tanah tidak tahu rakyat kecil protes milik tanah

Setelah melalui proses filtering, langkah selanjutnya dari preprocessing data adalah tahap tokenizing, dimana kata-kata yang didapatkan dari proses filtering dipisah-pisah berdasarkan kata-katanyanya. Berikut merupakan perubahan tweet pada tahapan *tokenizing.*

Tabel 4.6 Tahapan *Preprocessing Tweets* Tokenizing

Sebelum	Sesudah
berarti visi misi yang lama hoax lapor	visi', 'misi', 'lama', 'hoax', 'lapor'
point diatas tidak bantah wajar unicorn tidak bakalan mengerti barangkali takut sksk protes tanah tidak tahu rakyat kecil protes milik tanah	'point', 'diatas', 'tidak', 'bantah', 'wajar', 'unicorn', 'tidak', 'bakalan', 'mengerti', 'barangkali', 'takut', 'sksk', 'protes', 'tanah', 'tidak', 'tahu', 'rakyat', 'kecil', 'protes', 'milik', 'tanah'

Dari langkah *preprocessing* yang sudah dilakukan, telah didapatkan kata-kata dari *tweets* pada pada tiap sesi debat. Berikut merupakan jumlah kata yang ditemukan setelah langkah preprocessing.



Gambar 4.3 Jumlah Kata Tiap Debat Setelah Tokenizing

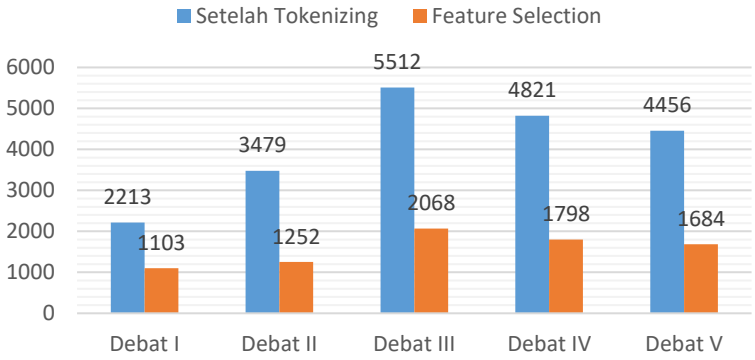
Berdasarkan Gambar 4.3 menunjukkan jumlah pada *tweets* tiap sesi pasca debat. Jika dibandingkan dengan Gambar 4.1 menunjukkan bahwa semakin banyak *tweets* yang dikumpulkan, maka kata-kata yang dikumpulkan semakin banyak. Hal tersebut terlihat pada Debat III memiliki jumlah kata yang paling banyak, yaitu 5512 kata yang berasal dari jumlah *tweets* terbanyak yaitu 98.225. Karena kata-kata pada tiap

debat memiliki jumlah yang besar sehingga tidak mampu dihitung pada komputer, maka dilakukan pemilihan term dengan metode *feature selection* dengan pengujian *Chi-square* sesuai dengan persamaan (2.1). Berikut merupakan simulasi feature selection yang diaplikasikan pada *tweets* pada sesi pasca debat kedua.

Tabel 4.7 *Simulasi Feature Selection*

No	Kata	Nilai Chi-square	P-Value	Keputusan
1	aaamiin	90.9029	0.0000	Tolak H_0
2	Abadi	3.3353	0.1887	Gagal Tolak H_0
⋮	⋮	⋮	⋮	⋮
641	debat	928.01	0.0000	Tolak H_0
642	demo	2.1653	0.3386	Gagal Tolak H_0
⋮	⋮	⋮	⋮	⋮
1075	hoax	334.4071	0.0000	Tolak H_0
1076	hotel	0.72392	0.6461	Gagal Tolak H_0
⋮	⋮	⋮	⋮	⋮
3479	zaman	Belum	0.0090	Tolak H_0

Terlihat pada Tabel 4.7 beberapa kata yang terpilih dan dibuang pada proses *feature selection*. Dengan nilai $\chi^2_{(0,05;1)}$ sebesar 3,841 sehingga dikatakan tolak H_0 apabila $\chi^2_{hitung} > \chi^2_{(0,05;1)}$ atau jika dilihat dari P-Value lebih dari taraf signifikansi sebesar 0,05. Dari Tabel 4.7 terlihat bahwa kata ‘aaamiin’, ‘debat’, ‘hoax’, dan ‘zaman’ memiliki keputusan tolak H_0 yang artinya bahwa kata-kata mempunyai hubungan terhadap nilai klasifikasi. Sementara kata ‘abadi’, ‘demo’, dan ‘hotel’ tidak memiliki hubungan terhadap nilai klasifikasi. Setelah dilakukan *feature selection*, maka jumlah kata-kata tiap debat dapat dilihat pada Gambar 4.4



Gambar 4.4 Jumlah Kata Setelah *Feature Selection*

Berdasarkan Gambar 4.4 terlihat pengurangan kata-kata dari tokenizing sampai tahap *feature selection* lebih dari 50%. Pengurangan kata paling ekstrim terjadi pada kata-kata di debat II yakni dilakukan pengurangan kata sebanyak 64%. Untuk mengetahui kata yang sering muncul pada *tweets* pada tiap debat dapat divisualisasikan menggunakan *wordcloud.* Berikut hasil dari visualisasi *wordcloud* satu kata.

Gambar 4.5 *Wordcloud* Satu Kata

Gambar 4.5 menunjukkan visualisasi kata-kata pada *tweets* yang sering muncul di setiap sesi pasca debat. Besarnya ukuran pada kata-kata di *wordcloud* menunjukkan frekuensi kata tersebut sering muncul pada *tweets* di setiap sesi pasca debat. Pada debat pertama terlihat bahwa kata "debat" memiliki frekuensi terbesar, disusul dengan kata "rakyat", "presiden", dan "pilpres". Pada sesi pasca debat kedua, kata "debat" kembali memiliki frekuensi terbesar lalu terdapat kata "tanah", "rakyat", "Indonesia", "presiden". Dilihat dari kata-kata pada sesi pasca debat pertama dan kedua menunjukkan bahwa warganet di Twitter antusias membahas debat presiden 2019. Pada sesi pasca debat ketiga kata yang memiliki frekuensi kemunculan terbesar adalah kata "Indonesia", "kartu", "pilih", "menang", dan "allah". Sedangkan pada sesi pasca debat keempat, kata "Indonesia" memiliki frekuensi kemunculan terbesar dan disusul oleh kata "rakyat", "menang", "presiden", "dukung", dan "pilih". Pada sesi pasca debat kelima frekuensi kata terbesar ada pada kata "Indonesia", "allah", "presiden", "pilih, dan "semoga". Dari sesi debat ketiga sampai kelima, kata "Indonesia", "pilih", dan "presiden" menjadi kata yang sering muncul. Hal tersebut menjelaskan bahwa tweet pada sesi pasca debat ketiga sampai kelima mayoritas berupa ajakan untuk memilih pasangan calon presiden tertentu. Sementara pada keseluruhan *tweets* sesi pasca debat, terdapat kata-kata "Indonesia", "rakyat" dan "presiden" sehingga dapat dikatakan bahwa warganet di Twitter fokus membicarakan presiden untuk rakyat Indonesia.

3.3 *Labeling Tweets*

Data *tweets* mengenai respon warganet terhadap debat presiden 2019 yang terkumpul dilakukan preprocessing *tweets*, selanjutnya dilakukan *labelling* pada *tweets* yang terbagi menjadi sentimen positif, negatif, dan netral. *Labelling* dilakukan dengan bantuan kamus lexicon yang terdiri dari sentimen positif dan negatif. Pada *tweets* dihitung kemunculan kata-kata yang memiliki sentimen positif dan negatif, yang kemudian dibandingkan antara jumlah sentiment positif dan negatifnya. *Tweets* diberi label positif jika pada *tweets* lebih banyak kata kata yang

mengandung sentimen positif daripada yang negatif, begitu pula sebaliknya. Apabila jumlah *tweets* yang mengandung sentimen positif sama dengan jumlah *tweets* yang mengandung sentimen negatif, maka *tweets* akan dikategorikan sebagai *tweets* yang bersifat netral. Berikut merupakan contoh dari *labelling tweets*.

Tabel 4.8 *Labeling Tweets*

Sebelum	Kata Positif	Kata Negatif	Score	Sentimen
visi', 'misi', 'lama', 'hoax', 'lapor'		‘hoax’	0 - 1 = - 1	Negatif
point', 'diatas', 'tidak', 'bantah', 'wajar', 'unicorn', 'tidak', 'bakalan', 'mengerti', 'barangkali', 'takut', 'sksk', 'protes', 'tanah', 'tidak', 'tahu', 'rakyat', 'kecil', 'protes', 'milik', 'tanah'	‘mengerti', 'tahu'	tidak', 'bantah', 'tidak', 'bakalan', 'takut', 'protes', 'tidak', 'kecil', 'milik'	2 - 9 = -7	Negatif

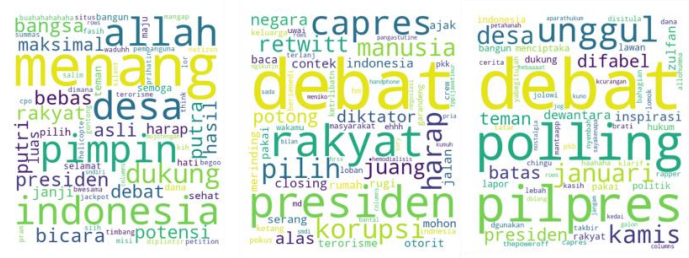
Labeling tweets dilakukan pada setiap *tweets* pada sesi pasca debat Presiden 2019. Berikut merupakan persentase sentimen pada sesi debat berdasarkan pasangan calon presiden yang di *mention*.

Tabel 4.9 Persentase Sentimen pada Tiap Sesi Debat

Pelaksanaan	Paslon 1			Paslon 2		
	Positif	Negatif	Netral	Positif	Negatif	Netral
Debat I	15.5	38.1	46.4	17.0	33.3	49.6
Debat II	18.3	33.7	48.0	18.2	31.1	50.7
Debat III	21.9	31.7	46.4	19.7	25.7	54.6
Debat IV	20.7	26.0	53.3	23.1	28.1	48.8
Debat V	24.7	26.2	49.1	21.6	25.5	52.9

Berdasarkan Tabel 4.9 terlihat bahwa mayoritas warganet yang mention baik pasangan calon 1 maupun pasangan calon 2 menanggapi debat calon presiden di Twitter dengan tanggapan netral,

yang artinya tidak memihak kubu manapun. Jika dibandingkan antara kedua pasangan calon, pada debat I, II, III, dan V *tweets* yang memention pasangan calon 1 lebih banyak bersifat sentimen negatif daripada pasangan calon 2. Sedangkan untuk sentimen positif, warganet lebih banyak memention pasangan calon nomor 1 pula pada debat II, III dan V. Hal tersebut menunjukkan bahwa pasangan calon 1, yaitu Joko Widodo-Mar'ruf Amin lebih banyak mendapatkan *tweets* berisi sentimen, baik sentimen positif maupun negatif. Jika dihubungkan dengan Tabel 4.2 meskipun pasangan calon 2 lebih banyak di mention oleh warganet, tetapi *tweets* yang banyak mengandung sentimen positif dan negatif terdapat pada *tweets* yang memention pasangan calon 1.



Gambar 4.6 *Wordcloud* Tiap Sentimen pada Debat 1

Gambar 4.6 menunjukkan banyaknya frekuensi kemunculan kata pada setiap sentimen, baik sentimen positif, negatif dan netral di periode debat pertama yang divisualisasikan melalui *wordcloud*. Terlihat pada sentimen positif didominasi oleh kata 'menang', 'indonesia', 'allah', 'desa' dan 'pimpin'. Untuk sentimen negatif didominasi oleh kata 'debat' kemudian disusul dengan kata 'rakyat', 'capres', 'rakyat', dan 'presiden'. Sedangkan kubu netral dominan dengan kata 'debat', 'pilpres', 'polling', dan 'unggul'. Terlihat pada kubu netral di *tweets* pasca debat I mayoritas lebih condong melakukan polling untuk pemilihan capres. Hal tersebut terlihat bahwa kata 'polling' mendominasi kubu netral.

Gambar 4.7 *Wordcloud* Tiap Sentimen pada Debat 2

Pada Gambar 4.7 menunjukkan *wordcloud* tiap sentimen pada Debat 2. Pada sentimen positif kata-kata yang dominan, yaitu ‘indonesia’, ‘rakyat’, ‘allah’, ‘menang’, ‘pilih’, dan ‘dukung’. Pada sentimen negatif terlihat kata ‘debat’, ‘tanah’ dan ‘bohong’. Topik ‘tanah’ merupakan topik yang dibahas pada debat 2 yang artinya bahwawarganet menanggapi topik tanah dengan sentimen negatif. pada kubu netral terlihat kata ‘’indonesia’, ‘mantap’, ‘presiden’, dan ‘rakyat merupakan tanggapan netral. Selanjutnya merupakan frekuensi kemunculan kata pada setiap sentimen di periode debat ketiga.

Gambar 4.8 *Wordcloud* Tiap Sentimen pada Debat 3

Berdasarkan Gambar 4.8 pada sentimen positif terdapat banyak kata yang dominan, diantaranya pada kata ‘dukung’, ‘indonesia’, ‘rakyat’, ‘allah’, ‘menang’ dan semoga. Untuk kata ‘pilih’, ‘jalan’, ‘hoax’ dan ‘bodoh’ merupakan kata-kata yang mendominasi sentimen negatif. Sebelumnya telah dibahas bahwa topik jalan merupakan salah

satu topik yang dibahas pada debat 2. Hal tersebut menunjukkan bahwa warganet menanggapi topik jalan dengan sentimen negatif. Sementara pada sentimen yang tidak memihak manapun, kata ‘aamiin, ‘indonesia’, ‘kartu’, ‘dan ‘presiden’ menjadi kata-kata yang mendominasi. Topik kartu merupakan topik yang dibahas pada debat 3 sehingga dapat dikatakan bahwa warganet menanggapi topik kartu dengan tanggapan netral. Berikut merupakan frekuensi kemunculan kata pada tiap sentimen yang divisualisasikan dengan *wordcloud* pada debat keempat.

Gambar 4.9 *Wordcloud* Tiap Sentimen pada Debat 4

Pada Gambar 4.9 di sentimen positif terlihat bahwa kata-kata yang mendominasi diantaranya kata ‘dukung’, ‘indonesia’, ‘rakyat’, ‘allah’ dan ‘menang. Pada sentimen negatif terlihat kata ‘rakyat’, ‘indonesia’, ‘pilih’, dan ‘presiden’. Sedangkan untuk kubu netral terlihat *tweets* dengan kata ‘indonesia’, ‘dukung’, ‘presiden’, dan ‘rakyat’ sebagai kata-kata yang dominan.

Gambar 4.10 *Wordcloud* Tiap Sentimen pada Debat 5

Gambar 4.10 menunjukkan *wordcloud* pada setiap sentimen di sesi pasca debat 5. Pada sentimen positif terilihat bahwa kata ‘allah’, ‘menang’, ‘indonesia’, ‘pilih’, dan presiden merupakan kata yang dominan. Sedangkan pada sentimen negatif terlihat kata-kata ‘pilih’, ‘indonesia’ dan ‘fitnah’ merupakan kata-kata yang dominan. Untuk kubu netral didominasi dengan kata ‘pilih’, ‘indonesia’, ‘aamiin’, ‘presiden’, dan ‘semoga’.

4.2 Pembobotan Kata

Setelah mendapatkan kata-kata pada tahapan *preprocessing tweets*, kata-kata tersebut kemudian dibobotkan dengan melihat frekuensi kemunculan kata pada suatu *tweets* yang kemudian dihitung frekuensi kemunculan pada seluruh *tweets* di setiap sesi debat. Pembobotan kata dilakukan dengan metode TF-IDF. Tabel 4.10 menunjukkan ilustrasi dalam perhitungan TF-IDF yang didapatkan pada *tweets* pada pasca sesi debat kedua.

Tabel 4.10 Jumlah Kumulatif Kata pada Tiap *Tweets*

Tweets	**Jumlah Kata**	**Frekuensi Kata**						
		aaamiin	**…**	**debat**	**…**	**hoax**	**…**	**zaman**
1	7	0	…	1	…	0	…	0
⋮	⋮	⋮	…	⋮	…	⋮	…	⋮
1037	10	1	…	0	…	0	…	0
⋮	⋮	⋮	…	⋮	…	⋮	…	⋮
22291	4	0	…	0	…	0	…	1
⋮	⋮	⋮	…	⋮	…	⋮	…	⋮
52418	7	0	…	0	…	1	…	0
Jumlah		374		3548		721		88

Berdasarkan Tabel 4.10 menunjukkan jumlah kumulatif tiap kata pada 52418 *tweets*. Tabel diatas juga menunjukkan jumlah kata pada tiap *tweets*. Terlihat bahwa pada *tweet* pertama terdapat 7 kata dan muncul kata ‘debat’ sebanyak satu kali, lalu pada tweet ke-1037 terdapat 10 kata dan muncul kata ‘aaamiin’. Pada tweet ke-22291

terdapat 4 kata dan terlihat muncul kata ‘zaman’ sampai pada tweet ke-52419 terdapat 7 kata dan muncul kata ‘hoax’. Setelah dihitung kumulatif kata-kata, maka dihitung nilai TF sesuai pada persamaan (2.2) yang dijabarkan pada Tabel 4.11.

Tabel 4.11 Perhitungan TF Kata pada Tiap *Tweets*

Tweets	**Jumlah Kata**	**TF**						
		aaamiin	**…**	**debat**	**…**	**hoax**	**…**	**zaman**
1	7	0	…	0.14286	…	0	…	0
⋮	⋮	⋮	…	⋮	…	⋮	…	⋮
1037	10	0.1	…	0	…	0	…	0
⋮	⋮	⋮	…	⋮	…	⋮	…	⋮
22291	4	0	…	0	…	0	…	0.25
⋮	⋮	⋮	…	⋮	…	⋮	…	⋮
52418	7	0	…	0	…	0.14286	…	0
Jumlah		374		3548		721		88

Perhitungan TF didapatkan dari frekuensi kata pada suatu *tweet* dibanding dengan jumlah kata pada *tweet* tersebut. Pada tweet pertama, kata ‘debat’ memiliki nilai 0,14286, kata ‘aaamiin’ pada *tweet* ke-1037 memiliki nilai TF sebesar 0,1. Kata ‘zaman’ pada *tweet* ke-22291 memiliki nilai TF sebesar 0,25 sedangkan kata ‘hoax’ pada *tweet* ke-52418 memiliki nilai TF sebesar 0,14286.

Tabel 4.12 Perhitungan IDF pada Kata Debat II

Kata	**Jumlah Kemunculan**	**Nilai IDF**
aaamiin	374	2.146609
⋮	⋮	⋮
debat	3548	1.169497
⋮	⋮	⋮
hoax	721	1.861545
⋮	⋮	⋮
zaman	88	2.774998

Perhitungan IDF dilakukan sesuai persamaan (2.3) yang didapatkan dari log dikali perbandingan jumlah *tweets* dan jumlah kemunculan data. Dari tabel tersebut terlihat bahwa semakin besar jumlah kemunculan kata pada suatu *tweets*, maka nilai IDF bernilai semakin kecil. Lalu untuk mendapatkan nilai TF-IDF dilakukan dengan mengalikan nilai TF dan nilai IDF sehingga didapatkan hasil sesuai Tabel 4.13.

Tabel 4.13 Perhitungan TF-IDF pada Kata Debat II

Tweets	**Jumlah term**	**TF-IDF**						
		aaamiin	**…**	**debat**	**…**	**hoax**	**…**	**zaman**
1	7	0	…	0.16707	…	0	…	0
⋮	⋮	⋮	…	⋮	…	⋮	…	⋮
1037	10	0.21466	…	0	…	0	…	0
⋮	⋮	⋮	…	⋮	…	⋮	…	⋮
22291	4	0	…	0	…	0	…	0.69375
⋮	⋮	⋮	…	⋮	…	⋮	…	⋮
52418	7	0	…	0	…	0.26594	…	0
Jumlah		374		3548		721		88

4.3 Membandingkan Kata pada Debat dengan *Tweets*

Pada setiap sesi debat presiden 2019, terdapat juru notulensi yang mencatat setiap kalimat yang diucapkan oleh pasangan calon presiden 2019. Kalimat-kalimat tersebut kemudian dianalisis oleh debatcapres.bahasakita.co.id yang bisa diakses secara publik sehingga didapatkan kata-kata penting yang sering diucapkan pasangan calon presiden 2019. Dari kata-kata tersebut peneliti membandingkannya dengan kata-kata yang muncul pada *tweets* di setiap pasca debat presiden 2019 sehingga dapat diketahui apakah warganet membicarakan topik debat selama pasca debat presiden 2019. Berikut merupakan hasil kemunculan kata pada debat di setiap *tweets* pada pasca debat presiden 2019.

Tabel 4.14 Frekuensi Kemunculan Kata Debat pada *Tweets*

Debat	Ya	Tidak	Total
Debat I	244	54	298
Debat II	141	26	167
Debat III	129	20	149
Debat IV	116	47	163
Debat V	180	35	215

Pada tabel diatas menunjukkan bahwa mayoritas kata-kata pada debat ikut dibahas pada *tweets* pasca debat 2019. Persentase terbesar kemunculan kata debat yang dibahas pada *tweets* berada pada debat III yakni sebesar 86% sedangkan persentase terkecil terdapat pada debat IV dengan kemunculan sebesar 71%. Untuk melihat kata yang sering dibahas pada *tweets* bisa dilihat pada visualisasi *wordcloud* berikut.

Gambar 4.11 *Wordcloud* Kata Debat pada *Tweets*

Gambar 4.11 menunjukkan kemunculan kata pada debat yang dibahas di *tweets* pasca debat yang divisualisasikan dengan *wordcloud*. Dari gambar tersebut menunjukkan bahwa frekuensi terbesar kata-kata yang muncul di setiap debat adalah kata "indonesia" dan "presiden". Kata tersebut bersifat umum dan tidak mengacu pada tema yang diusung pada setiap sesi debat. Pada debat I muncul kata "korupsi" yang sesuai dengan tema debat I yaitu Hukum, HAM, Korupsi, dan Terorisme. Pada tweet pasca debat II muncul kata "tanah" yang sesuai dengan tema debat II, yaitu Energi, Pangan, Infrastruktur, Sumber Daya Alam, dan Lingkungan Hidup. *Tweets* pasca debat III muncul kata "kartu" sebagai upaya calon presiden untuk menyampaikan visi-misi sesuai tema debat III. Pada *tweets* pasca debat IV muncul kata "TNI" sesuai dengan bahasan debat IV tentang keamanan Indonesia. Sedangkan pada *tweets* pasca debat V tidak muncul kata-kata sesuai dengan tema pada debat V. Dari hasil tersebut disimpulkan bahwa hanya pada *tweets* pasca debat I, II, III, dan IV yang muncul kata sesuai topik debat calon presiden 2019. hasil kemunculan kata *tweets* pada debat I-V menunjukkan bahwa kata-kata yang sesuai tema debat merupakan bukan kata-kata yang memiliki frekuensi tertinggi yang dibahas pada *tweets*, sehingga warganet tidak banyak mengulas topik debat pada *tweets* di pasca debat presiden 2019.

4.4 *Clustered Support Vector Machines Classification*

Algoritma *Clustered Support Vector Machine Classification* pada penelitian ini akan menggunakan algoritma klaster K-*Means* dan tiga macam kernel sebagai klasifikasi yaitu: kernel linear dengan mengoptimalkan parameter C, polinomial dengan mengoptimalkan parameter C dan gamma dan Radial Basis Function (RBF) dengan mengoptimalkan parameter C dan gamma. Pada algoritma *clustering* jumlah klaster ditentukan sebanyak tiga berdasarkan sentimen *tweets*, yaitu positif, negatif dan netral.

4.4.1 Pengelompokan *Tweets* Menggunakan Metode *K-Means*

Di setiap *tweets* pada pasca debat presiden 2019 dikelompokkan menjadi tiga klaster. Berikut adalah jumlah klaster pada setiap *tweets* di pasca debat presiden 2019.

Tabel 4.15 Jumlah *Tweets* pada Tiap Klaster

Debat	Klaster 1	Klaster 2	Klaster 3
Debat I	244	23.314	22.096
Debat II	355	37.154	14.909
Debat III	4.696	63.507	30.022
Debat IV	4.518	57.550	24.034
Debat V	263	41.479	33.054

Untuk mengetahui kedekatan antar klaster dilihat dari nilai *within cluster sum of square* yang bisa dilihat pada tabel 4.16 berikut.

Tabel 4.16 Nilai *Within Cluster Sum of Square* Tiap Klaster

Debat	Klaster 1	Klaster 2	Klaster 3
Debat I	5,93	15.506,05	13.330,57
Debat II	189,70	17.902,08	7.514,33
Debat III	3.652,81	32.376,95	17.640,83
Debat IV	3.326,53	28.509,60	13.433,89
Debat V	149,67	21.554,29	18.671,14

Dengan melihat antara Tabel 4.15 dan Tabel 4.16 menunjukkan bahwa semakin besar jumlah anggota tiap klaster, maka semakin besar nilai *within cluster sum of square*.

4.4.2 *CSVM Tweets* Pasca Debat Calon Presiden 1

Pembahasan klasifikasi data *tweets* pasca debat calon presiden I dengan mempertimbangkan kernel, serta parameter ‘C’ yang memiliki nilai rentang antara 10^{-2} hingga 10^{2}. Pada kernel polinomial dan Radial Basis Function (RBF) juga dipertimbangkan parameter gamma dengan rentang 10^{-2} hingga 10^{2}. Hasil pengklasifikasian pada tiap kernel dan semua paramater dapat dilihat pada Lampiran. Dari setiap kernel dilihat hasil klasifikasi terbaik dengan mempertimbangkan nilai Akurasi,

Sensitivitas, spesifikasi, ROC dan waktu pengklasifikasian yang dapat dilihat pada Tabel 4.14 berikut.

Tabel 4.17 Ketepatan Klasifikasi *Tweets* Debat Calon Presiden I

Ukuran Ketepatan Klasifikasi	Kernel		
	Linear	**Polinomial**	**RBF**
Cost "C"	1	1	100
Gamma	-	1	0.1
Akurasi	0,847	0,874	0,866
AUC	0,883	0,897	0,890
Waktu	217,28	250,37	1.406,08

Dilihat berdasarkan ketepatan klasifikasi dan waktu pengklasifikasian pada Tabel 4.14 dapat diketahui dengan Kernel Polinomial dengan parameter C sebesar 1 dan gamma sebesar 1 menghasilkan ketepatan klasifikasi terbaik dengan running program tercepat yaitu lama 250,37 detik. AUC menunjukkan nilai sebesar 0,897 menunjukkan klasifikasi dengan metode CSVM dengan parameter dan kernel tersebut sudah baik.

4.4.3 CSVM Tweets Pasca Debat Calon Presiden 1I

Sama seperti CSVM pada *tweets* debat calon presiden I, klasifikasi mempertimbangkan kernel, serta parameter ‘C’ yang memiliki nilai rentang antara 10^{-2} hingga 10^{2}. Khusus untuk kernel polinomial dan Radial Basis Function (RBF) juga dipertimbangkan parameter gamma dengan rentang 10^{-2} hingga 10^{2}. Hasil pengklasifikasian pada tiap kernel dan semua paramater dapat dilihat pada Lampiran. Berikut merupakan hasil ketepatan klasifikasi CSVM *Tweets* Pasca Debat II.

Tabel 4.18 Ketepatan Klasifikasi *Tweets* Debat Calon Presiden II

Ukuran Ketepatan Klasifikasi	Kernel		
	Linear	**Polinomial**	**RBF**
Cost "C"	10	1	10
Gamma	-	1	1

Tabel 4.18 Ketepatan Klasifikasi Tweets Debat Calon Presiden II (Lanjutan)

Ukuran Ketepatan Klasifikasi	Kernel		
	Linear	**Polinomial**	**RBF**
Akurasi	0,839	0,847	0,843
AUC	0,869	0,878	0,874
Waktu	711,09	754,24	1.899,21

Tabel 4.15 menunjukkan ukuran ketepatan klasifikasi dan waktu klasifikasi pada tiap kernel dengan parameter yang optimal pada *tweets* debat calon presiden II. Dilihat dari ukuran ketepatan klasifikasi yaitu akurasi dan AUC disimpulkan bahwa kernel polinomial dengan parameter C dan Gamma sebesar 1 memiliki ukuran ketepatan klasifikasi yang terbaik dibanding dengan kernel lainnya. Untuk waktu pengklasifikasian, kernel linear memiliki waktu yang lebih cepat daripada kernel polinomial. Namun dengan selisih waktu yang tidak terlalu besar, maka kernel polinomial memiliki ukuran ketepatan klasifikasi yang terbaik. Dilihat dari nilai AUC sebesar 0,878 menunjukkan bahwa klasifikasi dikatakan baik.

4.4.4 CSVM Tweets Pasca Debat Calon Presiden 1II

Sama seperti CSVM pada *tweets* debat periode sebelumnya, metode klasifikasi dilakukan dengan mempertimbangkan kernel, parameter ‘C’, dan parameter gamma. Pada Tabel 4.19 menunjukkan hasil ketepatan klasifikasi CSVM *Tweets* Pasca Debat III.

Tabel 4.19 Ketepatan Klasifikasi *Tweets* Debat Calon Presiden III

Ukuran Ketepatan Klasifikasi	Kernel		
	Linear	**Polinomial**	**RBF**
Cost "C"	1	1	10
Gamma	-	1	1
Akurasi	0,878	0,879	0,843
AUC	0,902	0,902	0,898
Waktu	6.868,45	5.530,81	9.121,95

Dengan melihat ketepatan klasifikasi berupa akurasi terlihat bahwa kernel polinomial dengan parameter ‘C’ dan gamma 1 memiliki

akurasi tertinggi yaitu sebesar 0,879. Sedangkan ketika dilihat dari nilai AUC pada masing-masing kernel dengan parameter yang optimal terlihat bahwa kernel linear memiliki nilai tertinggi, yaitu 0,9017. Setelah dipertimbangkan dengan waktu klasifikasi yang lebih singkat dapat disimpulkan bahwa metode CSVM dengan kernel polinomial parameter 'C' dan gamma 1 lebih efektif dalam mengklasifikasikan *tweets* pasca debat calon presiden III tercepat, dengan waktu 5.530,81 detik.

4.4.5 CSVM Tweets Pasca Debat Calon Presiden 1V

Dengan mempertimbangkan kernel, parameter 'C', dan parameter gamma juga, *tweets* pada pasca debat calon presiden keempat didapatkan hasil yang paling optimal yang bisa dilihat pada Tabel 4.20 berikut.

Tabel 4.20 Ketepatan Klasifikasi *Tweets* Debat Calon Presiden IV

Ukuran Ketepatan Klasifikasi	Kernel		
	Linear	**Polinomial**	**RBF**
Cost "C"	10	1	100
Gamma	-	1	0,1
Akurasi	0,849	0,867	0,874
AUC	0,879	0,895	0,894
Waktu	2.677,210	4.145,140	17107,960

Dengan melihat Tabel 4.20 yang membandingkan ketepatan klasifikasi berdasarkan akurasi terlihat bahwa nilai akurasi tertinggi terdapat pada kernel RBF dengan parameter 'C' sebesar 100 dan Gamma sebesar 0.1. Akan tetapi jika dilihat dari nilai AUC-nya menunjukkan bahwa polinomial dengan parameter 'C' dan gamma sebesar 1 memiliki nilai yang tertinggi dibandingkan dengan kernel lainnya. Dengan mempertimbangkan waktu pengklasifikasian, kernel polinomial disimpulkan lebih optimal dikarenakan waktu running yang lebih cepat, yakni sebesar 4145,14 detik dibandingkan dengan RBF sebesar 17107,96 detik.

4.4.6 CSVM Tweets Pasca Debat Calon Presiden V

Sama seperti CSVM pada *tweets* debat calon presiden pasca periode sebelumnya, berikut merupakan hasil pengklasifikasian pada tiap kernel dan semua paramater yang paling optimal.

Tabel 4.21 Ketepatan Klasifikasi *Tweets* Debat Calon Presiden V

Ukuran Ketepatan Klasifikasi	Kernel		
	Linear	**Polinomial**	**RBF**
Cost "C"	10	1	100
Gamma	-	1	0,1
Akurasi	0,854	0,868	0,870
AUC	0,883	0,896	0,896
Waktu	1.723,770	3.049,600	11.257,350

Tabel 4.21 menunjukkan ukuran ketepatan klasifikasi dan waktu klasifikasi pada tiap kernel dengan parameter yang optimal pada *tweets* debat calon presiden V. Dilihat akurasi dan ROC disimpulkan bahwa kernel RBF dengan parameter ‘C’ sebesar 100 dan Gamma sebesar 0.1 memiliki ukuran ketepatan klasifikasi yang terbaik dibanding dengan kernel lainnya.

(Halaman ini sengaja dikosongkan)

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis dan pembahasan yang telah dilakukan pada bab 4, maka diperoleh kesimpulan sebagai berikut.

1. Penentuan sentimen dilakukan dengan bantuan kamus *lexicon* mendapatkan hasil bahwa kata yang mayoritas muncul pada sentimen positif di setiap debat meliputi kata ‘menang’, ‘presiden’, ‘pilih’ dan ‘allah’. Pada Sentimen negatif terdapat beberapa topik pada debat II dan III yaitu topik ‘tanah’, dan ‘jalan’ dan kata-kata lainnya yang dominan diantaranya ‘debat’, ‘bohong’, ‘hoax’, ‘bodoh’ dan ‘fitnah’. T*weets* dengan kubu netral mayoritas muncul kata-kata ‘polling’, ‘dukung’, ‘presiden’ dan ‘semoga’. Jumlah *tweets* yang didapatkan pasangan calon 2 lebih banyak daripada pasangan calon 1, tetapi pasangan calon 1 mendapatkan lebih banyak mendapatkan sentimen positif dan negatif.
2. Kata-kata yang sesuai tema debat bukan kata-kata yang memiliki frekuensi tertinggi. Warganet tidak mengulas topik debat pada *tweets* di pasca debat presiden 2019.
3. Hasil metode *Clustered Support Vector Machines* di setiap debat, yaitu :
    a. Metode terbaik pada Debat I menggunakan kernel polinomial dengan parameter C sebesar 1 dan gamma sebesar 1 dengan akurasi dan AUC sebesar 0,874 dan 0,897 dan *running time* selama 250,37 detik.
    b. Debat II dengan ketepatan klasifikasi terbaik menggunakan kernel polinomial dengan parameter C dan gamma sebesar 1 dengan proses klasifikasi

selama 754,240 detik. Akurasi dan AUC sebesar 0,847 dan 0,878.

c. Pada Debat III CSVM dengan kernel polinomial parameter C dan gamma 1 lebih efektif dengan akurasidan AUC sebesar 0,879 dan 0,902 serta *running time* selama 5530,810 detik.

d. Metode terbaik pada Debat IV menggunakan kernel polinomial dengan parameter C dan gamma sebesar 1 dengan akurasi dan AUC sebesar 0,867 dan 0,896 dan waktu *running* 4145,14 detik.

e. Pada Debat V metode CSVM terbaik menggunakan kernel RBF parameter C sebesar 100 dan gamma sebesar 0,1. Nilai akurasi dan AUC pada metode ini adalah 0,874 dan 0,894 serta *running time* selama 17107,960 detik.

5.2 Saran

1. Pada penelitian klasifikasi tweet ini sangat membutuhkan komputer yang memiliki kapasitas Random Access Memory (RAM) yang besar agar dapat memangkas waktu running data tweet warganet.
2. Untuk penelitian selanjutnya, penelitian serupa dapat dikembangkan dengan menggunakan API Stream Premium dan dapat dibuat program untuk otomatisasi klasifikasi. Sehingga hasil analisis sentimen dapat diakses secara real time dan lebih akurat.
3. Melengkapi daftar kata pada stopwords dan kamus lexicon dengan daftar kata singkatan dan daftar kata slang dalam bahasa Indonesia.

DAFTAR PUSTAKA

Castella, Q., & Sutton, C. (2014). Word Storm : Multiples of Word Clouds for Visual Comparison of Documents. *Proceedings of the 23rd international conference on World wide web* , 665-676.

Diani, R., Wisesty, U. N., & Aditsania, A. (2017). Analisis Pengaruh Kernel Support Vector Machine (SVM) pada Klasifikasi Data Microarray untuk Deteksi Kanker. *Indonesia Journal on Computing Vol 2*, 109-118.

Feldman, R., & Sanger, J. (2007). *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* New York: Cambridge University Press.

Gorunescu, F. (2011). *Data Mining, Concepts, Models and Techniques.* Berlin: Springer.

Gross, J. H., & Johnson, K. T. (2016). Twitter Taunts And Tirade Negative Campaigning In The Age Of Trump. *American Political Science Association*, 748-754.

Gu, Q., & Han, J. (2013). Clustered Support Vector Machines. *Journal of Machine Learning Research*, 307-315.

Han, J. K., & Pei, J. (2012). *Data Mining Concepts and Techniques.* USA: Morgan Kaufmann Publishers.

Hand, D. J., & Till, R. J. (2001). A Simple Generalisation of The Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning 45*, 171-186.

Hardle, W. K., Prastyo, D. D., & Hafner, C. M. (2014). Support Vector Machines with Evolutionary Model Selection for Default Prediction. In J. Racine, L. Su, & A. Ullah, *The Oxford handbook of applied nonparametric and semiparametriceconometrics and statistics* (pp. 346-376). New York: Oxford University.

Heimlich, R. (2012, September 11). *Most Say Presidential Debates Influence Their Vote*. Diakses dari Factank: http:// www.pewresearch.org/fact-tank/2012/09/11/most-say-presidential-debates-influence-their-vote/

Herbrich, R., & Graepel, T. (2010). *Natural Language Processing.* Unitate State of America: Taylor and Francis Group, LLC.

Heryanto, G. G., & Rumaru, S. (2013). *Komunikasi Poltik Sebuah Pengantar.* Surabaya: Ghalia Indonesia.

Holbrook, T. M. (1999). Political Learning from Presidential Debates. *Political Behavior*, 67-89.

Hum, A. F. (2014). *Abraham Lincoln Bapak Demokrasi Sepanjang Masa.* Yogjakarta: Ircisod.

Iriawan, N., Fithriasari, K., & Pravitasari, A. A. (2018). Comparative study of Brain Tumor Segmentation using Different Segmentation Techniques in Handling Noise . *International Conference on Computer Engineering, Network adn Intelligent Multimedia (CENIM)*, 289-293.

Jamieson, K. H., & Gottfried, J. A. (2010). Are There Lessons for the Future of News from the 2008 Presidential Campaign? *Daedalus*, 18-25.

Jiang, S., Xie, J., & Zhang, Y. (2009). Clustering Support Vector Machines for Unlabeled Data Classification. *Internasional Conference on Test and Measurement*, 34-38.

Johnson, R. (2007). *Applied Multivariate Statistical Analysis.* Madison: Pearson Prentice Hall.

Karim, M. R. (1991). *Pemilu Demokratis Kompetitif.* Yogyakarta: Tiara Wacana.

KPU. (2014, Juli 22). *Hasil Pilpres 2014*. Diakses dari Komisi Pemilihan Umum: https://kpu.go.id/index.php/pages/detail/2014/316

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* California: Morgan and Claypool.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook.* New York: Springer.

Mayasari, R. W. (2018). *Text Mining pada Akun Resmi Pemerintah Kota Surabaya dengan Metode Regresi Logistik, Support Vector Machine (SVM), dan Naive Bayes Classifier (NBC).* Skripsi. Surabaya: Institut Teknologi Sepuluh Nopember.

Namaan, M., Becker, H., & Gravano, L. (2011). Hip and Trendy : Characterizing Emerging Trends on Twitter. *Journal of the American Society for Information Science and Technology*, 902-918.

Rofiq, M. A., Ahmad, I. S., & Prastyo, D. D. (2018). *Analisis Profiling Tweets pada Pemilihan Gubernur di Jawa Tahun 2018 Menggunakan Metode Clustered Support Vector Machines.* Skripsi. Surabaya: Institut Teknologi Sepuluh Nopember.

Rofiqoh, U. (2017). Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexicon Based Features. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 1725-1732.

Statista. (2018, November 22). *Number of Monthly Active Twitter Users*. Diakses dari statista.com: www.statista.com /statistics/282087/number-of-monthly-active-twitter-users/

Susanto, H. (2014). *Visualisasi Data Teks Twitter Berbasis Bahasa Indonesia Menggunakan Teknik Pengklasteran.* Skripsi. Surabaya: Institut Teknologi Sepuluh Nopember.

Szkaliczki, T. (2016). clustering.sc.dp: Optimal Clustering with Sequential Constraint by Using Dynamic Programming. *The R Journal Vol. 8*, 318-327.

Yang, Y., & Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. *ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning*, 412-420.

LAMPIRAN

Lampiran 1. *Ketepatan Klasifikasi Tweets Sesi Pasca Debat Presiden Kernel Linear*

Pasca Debat Presiden I

Cost "C"	Akurasi	AUC	Time
0,01	0,6172569	0,81353762	276,23
0,1	0,7897043	0,85395706	207,95
1	0,8512839	0,88269452	217,28
10	0,8512839	0,8775093	413,56
100	0,8466594	0,8732096	2473,08

Pasca Debat Presiden II

Cost "C"	Akurasi	AUC	Time
0,01	0,576532	0,76653368	408,59
0,1	0,752385	0,82308787	378,85
1	0,831885	0,86528561	339,91
10	0,838881	0,86882547	711,09
100	0,832945	0,86548523	3910,89

Pasca Debat Presiden III

Cost "C"	Akurasi	AUC	Time
0,01	0,770575	0,83656309	4321,92
0,1	0,855591	0,88523716	4233,59
1	0,87867	0,90164976	6868,45
10	0,871769	0,89505403	9573,37
100	0,862605	0,88660225	13293,14

Lampiran 1, Ketepatan Klasifikasi *Tweets* Sesi Pasca Debat Presiden Kernel Linear (Lanjutan)

Pasca Debat Presiden IV

Cost "C"	Akurasi	AUC	Time
0,01	0,584952	0,77969885	1357,41
0,1	0,775634	0,83926726	1442,12
1	0,843131	0,87589129	1661,11
10	0,849455	0,87852327	2677,21
100	0,842679	0,87341259	6599,69

Pasca Debat Presiden V

Cost "C"	Akurasi	ROC	Time
0,01	0,579929	0,75009988	1351,99
0,1	0,767419	0,83192089	1479,55
1	0,845342	0,8768379	1725,27
10	0,853662	0,88263495	1723,77
100	0,845342	0,87647721	4873,5

Lampiran 2. *Ketepatan Klasifikasi Tweets Sesi Pasca Debat Presiden Kernel Polinomial*

Pasca Debat Presiden I

Cost "C"	Gamma	Akurasi	AUC	Time
0,01	0,01	0,479859	0,5	263,22
0,01	0,1	0,479859	0,5	259,65
0,01	1	0,7908	0,871764	164,41
0,01	10	0,86175	0,882912	424,17
0,01	100	0,720701	0,764487	4623,84
0,1	0,01	0,479859	0,5	312,86
0,1	0,1	0,49568	0,5	290,14
0,1	1	0,861385	0,892395	175,53
0,1	10	0,850189	0,872182	1122,83
0,1	100	0,486674	0,610544	3772,53
1	0,01	0,479859	0,5	263,36
1	0,1	0,644152	0,882295	217,56
1	1	0,87392	0,896622	250,37
1	10	0,851284	0,882695	180,7
1	100	0,851284	0,882695	179,11
10	0,01	0,851284	0,5	341,2
10	0,1	0,851284	0,858535	341,5
10	1	0,851284	0,876801	341,65
10	10	0,851284	0,814495	341,28
10	100	0,851284	0,696378	341,73
100	0,01	0,846659	0,5	2232,6
100	0,1	0,846659	0,884982	2234,05
100	1	0,846659	0,86803	2232,19
100	10	0,846659	0,610895	2233,1
100	100	0,846659	0,875047	2241,37

Lampiran 2. Ketepatan Klasifikasi *Tweets* Sesi Pasca Debat Presiden Kernel Polinomial (Lanjutan)

Pasca Debat Presiden II

Cost "C"	Gamma	Akurasi	AUC	Time
0,01	0,01	0,49894	0,5	552,1
0,01	0,1	0,49894	0,5	578,38
0,01	1	0,737651	0,8203354	435,64
0,01	10	0,840153	0,8714955	1447,56
0,01	100	0,713165	0,7730617	7866,11
0,1	0,01	0,49894	0,5	501,25
0,1	0,1	0,49894	0,5	456,96
0,1	1	0,831355	0,8680588	403,18
0,1	10	0,830507	0,8639796	3035,47
0,1	100	0,499788	0,602282	6524,02
1	0,01	0,49894	0,5	497,7
1	0,1	0,558406	0,7655914	463,83
1	1	0,847255	0,877604	754,24
1	10	0,807081	0,8472057	6576,89
1	100	0,653275	0,7730617	13973,06
10	0,01	0,49894	0,5	501,69
10	0,1	0,737545	0,8202578	398,56
10	1	0,840153	0,8714955	1437,55
10	10	0,738605	0,7839356	7944,04
10	100	0,315561	0,4322283	13647,75
100	0,01	0,49894	0,5	461,32
100	0,1	0,831355	0,8680588	380
100	1	0,830401	0,8638514	2906,3
100	10	0,694191	0,7493154	6574,21
100	100	0,315561	0,4329345	13662,59

Lampiran 2. Ketepatan Klasifikasi *Tweets* Sesi Pasca Debat Presiden Kernel Polinomial (Lanjutan)

Pasca Debat Presiden III

Cost "C"	Gamma	Akurasi	AUC	Time
0,01	0,01	0,520928	0,5	2225,98
0,01	0,1	0,520928	0,7978977	3364,7
0,01	1	0,769401	0,9015138	3838,51
0,01	10	0,872172	0,8400632	8664,34
0,01	100	0,567647	0,731522	16180,62
0,1	0,01	0,520928	0,5	6069,96
0,1	0,1	0,520928	0,8370186	6692,76
0,1	1	0,855317	0,8738839	4877,61
0,1	10	0,867534	0,8907696	18189,14
0,1	100	0,65181	0,7232387	25786,35
1	0,01	0,520928	0,5	4170,48
1	0,1	0,581391	0,7978977	2690,2
1	1	0,879129	0,9015138	5520,81
1	10	0,809276	0,8400632	21201,05
1	100	0,68207	0,731522	13038,56
10	0,01	0,520928	0,5	5205,09
10	0,1	0,769457	0,8370186	4054,47
10	1	0,872059	0,6689281	8628,91
10	10	0,579299	0,6689281	22630,29
10	100	0,620645	0,7000056	13882,92
100	0,01	0,520928	0,5	5828,25
100	0,1	0,855373	0,8854074	6295,39
100	1	0,867591	0,8907218	19194,83
100	10	0,443382	0,5599455	22515,84
100	100	0,620645	0,7000056	16991,95

Lampiran 2. Ketepatan Klasifikasi *Tweets* Sesi Pasca Debat Presiden Kernel Polinomial (Lanjutan)

Pasca Debat Presiden IV

Cost "C"	Gamma	Akurasi	AUC	Time
0,01	0,01	0,520552	0,5	3495,14
0,01	0,1	0,520552	0,5	3287,81
0,01	1	0,75834	0,8322577	2952,25
0,01	10	0,859715	0,8890049	6616,97
0,01	100	0,515132	0,6322495	15085,77
0,1	0,01	0,520552	0,5	3190,53
0,1	0,1	0,520552	0,5	3098,69
0,1	1	0,847132	0,8796955	2681,49
0,1	10	0,848358	0,8774373	11585,24
0,1	100	0,549397	0,6492026	14660,92
1	0,01	0,520552	0,5	3190,53
1	0,1	0,520552	0,778841	3098,69
1	1	0,867265	0,8951211	4145,14
1	10	0,83571	0,8724729	14567,63
1	100	0,383945	0,5416719	16878,95
10	0,01	0,520552	0,5	3988,08
10	0,1	0,758534	0,8323785	2952,04
10	1	0,859715	0,8889754	6721,1
10	10	0,559786	0,770777	15105
10	100	0,444989	0,5934875	14930,09
100	0,01	0,520552	0,5	3015,72
100	0,1	0,847003	0,8847551	2870,82
100	1	0,848293	0,8774019	11479,8
100	10	0,528489	0,6608111	14019,08
100	100	0,444989	0,5934875	15151,93

Lampiran 2. Ketepatan Klasifikasi *Tweets* Sesi Pasca Debat Presiden Kernel Polinomial (Lanjutan)

Pasca Debat Presiden V

Cost "C"	Gamma	Akurasi	AUC	Time
0,01	0,01	0,512034	0,5	2029,75
0,01	0,1	0,512034	0,5	2228,95
0,01	1	0,753677	0,8292485	1631,77
0,01	10	0,862948	0,889999	3505,84
0,01	100	0,709776	0,7553555	12875,63
0,1	0,01	0,512034	0,5	2513,38
0,1	0,1	0,512034	0,5	1831,22
0,1	1	0,849577	0,882932	2500,05
0,1	10	0,849874	0,8748206	7828,34
0,1	100	0,450379	0,5556765	10787,59
1	0,01	0,512034	0,5	2925,26
1	0,1	0,565741	0,7436809	2018,03
1	1	0,867999	0,8958935	3049,6
1	10	0,831452	0,8636812	15751,77
1	100	0,396672	0,5897328	13368,58
10	0,01	0,512034	0,5	2902,96
10	0,1	0,753677	0,8292963	2888,79
10	1	0,863022	0,890048	5326,39
10	10	0,716313	0,7626407	13643,46
10	100	0,397712	0,5149287	12900,14
100	0,01	0,512034	0,5	3091,39
100	0,1	0,849502	0,8828842	2543,62
100	1	0,849948	0,8785442	7951,18
100	10	0,454687	0,5645875	11238,47
100	100	0,397712	0,5149287	10809,6

Lampiran 3. *Ketepatan Klasifikasi Tweets Sesi Pasca Debat Presiden Kernel RBF*

Pasca Debat Presiden I

Cost "C"	Gamma	Akurasi	AUC	Time
0,01	0,01	0,479859	0,5	1869,04
0,01	0,1	0,493854	0,5	1860,86
0,01	1	0,622733	0,834113	786,25
0,01	10	0,773032	0,864274	710,69
0,01	100	0,81806	0,856869	558,11
0,1	0,01	0,494219	0,5	899,76
0,1	0,1	0,666058	0,824896	791,42
0,1	1	0,780699	0,859855	445,46
0,1	10	0,816113	0,871068	417,83
0,1	100	0,819521	0,857653	420,52
1	0,01	0,677863	0,819871	849,1
1	0,1	0,811123	0,863568	597,59
1	1	0,851041	0,882531	405,89
1	10	0,83096	0,869661	414,33
1	100	0,818182	0,861951	325,83
10	0,01	0,81733	0,865352	803,3
10	0,1	0,861142	0,889787	620,5
10	1	0,861872	0,886914	690,78
10	10	0,828039	0,865734	418,98
10	100	0,816113	0,854183	332,56
100	0,01	0,858221	0,886478	2295,47
100	0,1	0,865888	0,890085	1406,08
100	1	0,850432	0,875349	1166,36
100	10	0,824389	0,86323	495,63
100	100	0,813922	0,851336	349,48

Lampiran 3. Ketepatan Klasifikasi *Tweets* Sesi Pasca Debat Presiden Kernel RBF (Lanjutan)

Pasca Debat Presiden II

Cost "C"	Gamma	Akurasi	AUC	Time
0,01	0,01	0,49894	0,5	3055,24
0,01	0,1	0,49894	0,5	1434,7
0,01	1	0,548124	0,7637956	1604,87
0,01	10	0,685287	0,8222479	1973,78
0,01	100	0,767225	0,7229065	1611,84
0,1	0,01	0,49894	0,5	1492,31
0,1	0,1	0,608862	0,776657	1379,75
0,1	1	0,735319	0,8149429	792,79
0,1	10	0,758109	0,8357521	789,28
0,1	100	0,767967	0,82218	795,91
1	0,01	0,620416	0,777489	1282,2
1	0,1	0,776553	0,8339788	991
1	1	0,829765	0,8613605	844,63
1	10	0,795103	0,8449074	934,14
1	100	0,769981	0,8222782	1453,06
10	0,01	0,784715	0,8377417	3076,75
10	0,1	0,833051	0,865064	1041,15
10	1	0,843121	0,8738941	1899,21
10	10	0,793195	0,8425718	1033,58
10	100	0,768073	0,8199013	840,5
100	0,01	0,829023	0,8615541	3998,2
100	0,1	0,840153	0,8705049	4421,69
100	1	0,835277	0,8685136	3875,53
100	10	0,789061	0,8396702	1171,14
100	100	0,766059	0,8181147	803,26

Lampiran 3. Ketepatan Klasifikasi *Tweets* Sesi Pasca Debat Presiden Kernel RBF (Lanjutan)

Pasca Debat Presiden III

Cost "C"	Gamma	Akurasi	AUC	Time
0,01	0,01	0,520901	0,5	4182,46
0,01	0,1	0,520901	0,5	4719,91
0,01	1	0,554047	0,7961649	4583,4
0,01	10	0,724928	0,8533114	4752,42
0,01	100	0,794106	0,8460866	4891,67
0,1	0,01	0,520901	0,5	6042,5
0,1	0,1	0,62645	0,7970477	6181,41
0,1	1	0,758357	0,8301229	5351,45
0,1	10	0,781153	0,8630205	5481,45
0,1	100	0,795973	0,8464301	4968,03
1	0,01	0,64619	0,7974833	7044,66
1	0,1	0,807512	0,8529083	6602,8
1	1	0,856723	0,8840665	5998,37
1	10	0,816901	0,8668272	5812,32
1	100	0,799095	0,8487055	4839,94
10	0,01	0,810916	0,8549723	5604,1
10	0,1	0,859672	0,8867087	6033,02
10	1	0,874152	0,8981544	9121,95
10	10	0,817477	0,8647325	3934,43
10	100	0,79802	0,8477082	4091,8
100	0,01	0,856674	0,9023857	14922,89
100	0,1	0,864706	0,8890994	11404,31
100	1	0,864706	0,8890994	11404,31
100	10	0,812613	0,8592312	5455,81
100	100	0,798134	0,847047	4811,49

Lampiran 3. Ketepatan Klasifikasi *Tweets* Sesi Pasca Debat Presiden Kernel RBF (Lanjutan)

Pasca Debat Presiden IV

Cost "C"	Gamma	Akurasi	AUC	Time
0,01	0,01	0,520552	0,5	3492,76
0,01	0,1	0,520552	0,5	4178,26
0,01	1	0,562625	0,7879039	3298,15
0,01	10	0,71317	0,8440566	3404,8
0,01	100	0,786604	0,8414151	3278,06
0,1	0,01	0,520746	0,5	4701
0,1	0,1	0,623217	0,7916978	4914,39
0,1	1	0,754856	0,8314296	3042,75
0,1	10	0,768794	0,8544629	5220,98
0,1	100	0,788217	0,8417475	3652,67
1	0,01	0,641866	0,7949021	4867,88
1	0,1	0,800865	0,8503849	4017,91
1	1	0,850874	0,8804177	3856,52
1	10	0,804801	0,8594071	3851,33
1	100	0,788604	0,8403437	3424
10	0,01	0,805962	0,8531853	5132,22
10	0,1	0,854165	0,8837847	5019,37
10	1	0,86649	0,8925817	7109,85
10	10	0,804672	0,8561783	3946,14
10	100	0,787443	0,8389536	3658,55
100	0,01	0,852617	0,8823154	10188,74
100	0,1	0,869201	0,8941464	17107,96
100	1	0,855779	0,8837395	9414,15
100	10	0,799961	0,8515326	4591,46
100	100	0,786152	0,8376898	3553,83

Lampiran 3. Ketepatan Klasifikasi *Tweets* Sesi Pasca Debat Presiden Kernel RBF (Lanjutan)

Pasca Debat Presiden V

Cost "C"	Gamma	Akurasi	AUC	Time
0,01	0,01	0,512034	0,5	2092,27
0,01	0,1	0,512851	0,5	2139,84
0,01	1	0,547244	0,7639395	1760,97
0,01	10	0,708141	0,8410297	1695,75
0,01	100	0,784653	0,840982	1697,39
0,1	0,01	0,514262	0,5	2265,16
0,1	0,1	0,613876	0,7721442	2433
0,1	1	0,740455	0,8198792	1849,4
0,1	10	0,764448	0,8495048	1786,64
0,1	100	0,785916	0,8408508	1986,5
1	0,01	0,630813	0,7759485	2709,9
1	0,1	0,791561	0,8436648	2130,25
1	1	0,850839	0,8821793	2061,61
1	10	0,80419	0,8579742	1937,12
1	100	0,787179	0,8403108	1470,68
10	0,01	0,797801	0,847097	2936,08
10	0,1	0,852697	0,8828159	2703,92
10	1	0,867925	0,894505	3972,31
10	10	0,80575	0,8565641	2511,69
10	100	0,786213	0,8383226	2081,21
100	0,01	0,851211	0,8818823	5881,23
100	0,1	0,86941	0,8964289	11257,35
100	1	0,858119	0,8860554	4875,9
100	10	0,803075	0,8524866	2204,78
100	100	0,785322	0,8370901	1526,31

Lampiran 4. Coding Crawling Data *Menggunakan R*

```
library(rtweet)
setwd('D:/')
create_token(
        app  = "my_twitter_research_app",
        consumer_key = 'consumer_key code ',
        consumer_secret = 'consumer_secret code ',
        access_token = 'access_token code,
        access_secret = 'access_secret code')
tweets <- search_tweets("@prabowo", n = 10000,
                  tweet_mode="extended", include_rts = FALSE)
Prabowo <- data.frame(tweets$created_at, tweets$screen_name,
                  tweets$text, tweets$display_text_width,
                  as.character(tweets$mentions_screen_name))
write.csv(Prabowo,file="1prabowo.csv")

tweets <- search_tweets("@jokowi", n = 10000,
                  tweet_mode="extended", include_rts = FALSE)
Jokowi <- data.frame(tweets$created_at, tweets$screen_name,
                  tweets$text, tweets$display_text_width,
                  as.character(tweets$mentions_screen_name))
write.csv(Jokowi,file="1jokowi.csv")
```

Lampiran 5. Coding Preprocessing Data *Menggunakan Python*

```
import pandas as pd
import pandas as dataframe
import string
import nltk
import matplotlib as mpl
import matplotlib.pyplot as plt
import re
import collections
import sys
import os
import numpy
import seaborn as sns
import csv
from collections import OrderedDict
from nltk.tokenize import word_tokenize
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
from IPython.display import display
from wordcloud import Wordcloud, STOPWORDS,
                    ImageColorGenerator
from pylab import figure, axes, pie, title, show
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer

#menetapkan directory penyimpanan file python
path = r'D:\python1'
os.chdir(path)
os.getcwd()

#memanggil data
data = pd.read_excel ('1prabowo.xlsx', sheet_name='Sheet1')

#menghapus link
datanolink=[]
```

```
for line in data:
    result=re.sub(r"http\S+","", str(line))
    datanolink.append(result)

#menghapus simbol retweet
datanort = []
for line in datanolink:
    result = re.sub(r"RT","",line)
    datanort.append(result)

#menghapus username
datanousername = []
for line in datanort:
    result=re.sub(r"@\S+","",line)
    datanousername.append(result)

#menghapus hastag
dataclearhastag=[]
for line in datanousername:
    result=re.sub(r"#\S+","",line)
    dataclearhastag.append(result)

#menghapus angka
datanonum=[]
for line in dataclearhastag :
    result=re.sub("\d"," ",line)
    datanonum.append(result)

#Menghapus emoticon
datanoemoticon=[]
for line in datanonum :
    result = re.sub(r'<.*?>',"",line)
    datanoemoticon.append(result)
```

```
#Menghapus spasi berlebih
datanodoublespace=[]
for line in datanoemoticon :
    result=re.sub(r'\s+',' ',line)
    datanodoublespace.append(result)

#Menghapus punctuation
datanopunctuation=[]
for line in datanodoublespace :
    result=re.sub(r"[^\w\s]"," ",line)
    datanopunctuation.append(result)

#Menghapus baris baru
datanoline=[]
for line in datanopunctuation :
    result=re.sub("\n"," ",line)
    datanoline.append(result)

len (datanoline)
print (datanoline)

#Case folding : merubah menjadi huruf kecil
data_lower = []
for line in datanoline:
    a = line.lower()
    data_lower.append(a)

#Stemming :
Factory = StemmerFactory ()
Stemmer = factory.create_stemmer ()
Datastemmed = map (lambda x: stemmer.stem(x), data_lower)
Databersih = map (lambda x: x.translate(str.maketrans(", ",
string.punctuation)), datastemmed)
Databersih = list (databersih)
```

```
#Sinonim Kata
kata = {"terimakasih ":"terima kasih ",
"ttg ": "tentang ", "dlm " : "dalam ",
"adlh " : "adalah ", "trims ":"terima kasih ",
"antri ":"antre ", "cepet ":"cepat ",
"dtg ":"datang ", "dmn ":"dimana ",
"diem ":"diam ", "sblm " : "sebelum ",
"gaada ":"tidak ada ", "gabisa ":"tidak bisa ",
"gatau ":"tidak tahu ", "jln ":"jalan ",
"jl ":"jalan ", "jm ":"jam ",
"jurus ":"jurusan ",  "kemaren ":"kemarin ",
"komplen ":"komplain ", "lwt ":"lewat ",
"trimakasih ":"terima kasih ", "makasih ":"terima kasih ",
"makas ":"terima kasih ", "mhn ":"mohon ",
"mksh ":"terima kasih ", "maksiiiii " : "terima kasih ",
"mlm ":"malam ", "malem ":"malam ",
"nungguin ":"tunggu ",  "nunggunya ":"tunggu ",
"nunggu ":"tunggu ",  "nyambung ":"sambung ",
"nyampe ":"sampai ",  "renti ":"berhenti ",
"responnya ":"respon ",  "sampe ":"sampai ",
"slmt ":"selamat ", "pake " : "pakai ",
"smp ":"sampai ", "udh " : "sudah ",
"start ":"mulai ", "blm " : "belum ",
"krj ":"kerja ", "kayak " : "seperti ",
"sm ": "sama ", "negri " : "negeri ",
"bkn " : "bukan ", "smbil " : "sambil ",
"moga" : "semoga", "mb " : "mbak ",
"sy " : "saya ", "yg " : "yang ",  "d " : "di ",
"td " : "tadi ", "tdi " : "tadi ",  "dgn " : "dengan ",
"dg " : "dengan ", "tdk " : "tidak ", "lo " : "kamu ",
"dr " : "dari ", "bang " : "mas ", "sdh " : "sudah ",
"jdi" : "jadi", "org " : "orang ",
"pake " : "pakai ", "banget " : "sangat ",
"biar " : "supaya ","jgn " : "jangan ","jg " : "juga ",
"tp " : "tapi ", "krn " : "karena ", "karna " : "karena ",
```

```
"krna " : "karena ","lg " : "lagi ",
"lgi " : "lagi ","bener " : "benar ","oce " : "oke ",
"oc " : "oke ", "bs " : "bisa ","dungu " : "bodoh ",
"goblok " : "bodoh ", "bpk " : "bapak ","trus " : "terus ",
"udh " : "sudah ","wes " : "sudah ",
"skrg " : "sekarang ","alloh " : "allah ","duit " : "uang ",
"spt " : "seperti ","msh " : "masih ","hrs " : "harus ",
"sok " : "merasa ","lbh " : "lebih ","sbg " : "sebagai ",
"thn " : "tahun ","dpt " : "dapat ","knp " : "kenapa ",
"paslon " : "pasangan calon ","bg " : "bagi ",
"negri " : "negeri","bgt " : "sangat ","klu " : "kalau ",
"kiyai " : "kyai ","cm " : "hanya ","amiin " : "aamiin ",
"sby " : "surabaya ","kiai " : "kyai ","th " : "tahun ",
"sj " : "saja ","gt " : "gitu ","tsb " : "tersebut ",
"oon " : "bodoh ","kek " : "seperti ","for " : "untuk ",
"byk " : "banyak ","bnyk " : "banyak ","tpi " : "tapi ",
"otaknya " : "otak ","info " : "informasi ","dri " : "dari ",
"kyk " : "seperti ","smoga " : "semoga ","smg " : "semoga",
"bego " : "bodoh ","kab " : "kabupaten ","ttp " : "tetap ",
"mmg " : "memang ","kaga " : "tidak ",
"mantab " : "mantap ",
"utang " : "hutang ","rp " : "rupiah ","abis " : "habis ",
"slalu " : "selalu","yes " : "iya ","opo " : "apa ",
"tgl " : "tanggal ","trs " : "terus ","klw " : "kalau ",
"jt " : "juta ","rame " : "ramai ","kudu " : "harus ",
"tv " : "televisi ","all " : "semua ",
"mngharapkn " : "mengharapkan ",
"sub " : "surabaya ","semgt " : "semangat ",
"kecewanya " : "kecewa ", "cnth " : "contoh ",
"mmpin " : "memimpin ","tgjwb " : "tanggung jawab ",
"pilpress " : "pilpres ","pancet " : "tetap ",
"pencitraaaaaaan " : "pencitraan "}

#normalisasi kata
def replace(text, dic):
```

```
    for i, j in dic.items():
        text = text.replace(i, j)
    return text

dic = OrderedDict (kata)
datachange = []
for line in databersih:
    result = replace (line, dic)
    datachange.append (result)

#menghilangkan kata stopword
stopwords = open('stopwords.txt').read()

satudata=[]
datafinal=[]
df =[]
for line in datachange:
    wo = word_tokenize(line)
    wo = [word for word in wo if not word in stopwords and not
word[0].isdigit()]
    datafinal.append(wo)
    df.append(" ".join(wo))

#menyipan file
pd.DataFrame(datafinal).to_csv('datafinalstem1.csv',
index=False)
pd.DataFrame(df).to_csv('df_1stem.csv', index=False)

for l in datafinal:
    satudata+= l
final={v: satudata.count(v) for v in set(satudata)}
with open ('final1_stemming.csv', 'w', newline="") as csv_file:
    writer= csv.writer(csv_file)
    for key, value in final.items():
        writer.writerow([key, value])
```

Lampiran 6. Coding Feature Selection *dengan Python*

```
#Count Vectorizer
from pandas import DataFrame
from sklearn.feature_extraction.text import CountVectorizer
vect = CountVectorizer(min_df=10)
X_dtm = vect.fit_transform(df)
X_dtm = X_dtm.toarray()
cv = DataFrame(X_dtm, columns=vect.get_feature_names())
cv.shape

#tf-idf
from sklearn.feature_extraction.text import TfidfTransformer
tfidf = TfidfTransformer(use_idf=True).fit_transform(cv)
tfidf_train = (tfidf.toarray())
print (tfidf_train)
print (tfidf_train.shape)
tf = DataFrame(tfidf.A,columns=vect.get_feature_names())
print (tf)
tf.shape

#memasukkan label
datay = pd.read_csv('label2.csv')
y = datay['Label']
y.shape

#Feature Selection dengan Chi Square
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
x_new = SelectKBest(chi2, k='jumlah kata')
x_new1 = x_new.fit_transform(tfidf_train, y)
x_new.pvalues_
chi2score = chi2(tfidf_train, y)
p_values = pd.DataFrame({'column': cv.columns, 'p_value':
x_new.pvalues_}).sort_values('p_value')
```

```
terpilih = p_values[p_values['p_value'] < .05]
print(terpilih)
pd.DataFrame(terpilih).to_csv('pval_pilih2.csv', index=True)

x_new = SelectKBest(chi2, k=’jumlah kata yang terpilih’)
x_new1 = x_new.fit_transform(tfidf_train, y)
pd.DataFrame(x_new1).to_csv('x_new2.csv', index=True)

#menyimpan file
pd.DataFrame(cv).to_csv('cv2.csv', index=False)
pd.DataFrame(tf).to_csv('bobot2.csv', index=False)
pd.DataFrame(chi2score).to_csv('chi2.csv', index=True)
pd.DataFrame(p_values).to_csv('pval_all2.csv', index=True)
```

Lampiran 7. *Wordcloud dengan Python*

```
import matplotlib
import matplotlib.pyplot as plt

kata = pd.DataFrame.from_dict(final,orient='index')
hasilsort = kata.sort_values(by=[0], ascending=False)

#mpl.rcParams['figure.figsize']=(8.0,6.0)  #(6.0,4.0)
mpl.rcParams['font.size']=12               #10
mpl.rcParams['savefig.dpi']=100            #72
mpl.rcParams['figure.subplot.bottom']=.1
wordcloud = Wordcloud(collocations = False,
                      background_color='white',
                      max_words=10000,
                      max_font_size=200,
                      width=500, height=560,
                      random_state=43
                     ).generate(str(hasilsort))

print(wordcloud)

fig = plt.figure(1)
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
fig.savefig('wordcloud.png', dpi=500)
```

Lampiran 8. Clustered Support Vector Machines *dengan R*

```
library(caret)
library(SwarmSVM)
library(pROC)
dataa1 <-  read.csv(file="D:/python1/bobot2.csv", header=TRUE,
sep=",")
datay1 <-  read.csv(file="D:/python1/label2.csv", header=TRUE,
sep=",")
data1 <- data.frame(datay1, dataa1)
folds <-  createFolds(factor(data1$Label), k = 10, list =
                  TRUE,returnTrain=TRUE)
nrow(data1)
str(folds)
data <-data1[folds$Fold03,]
nrow(data)
trainRowNumbers <-  createDataPartition(data$Label, p=0.8,
                               list=FALSE)
head(trainRowNumbers)
trainData <- data[trainRowNumbers,]
testData <- data[-trainRowNumbers,]
svmguide1.t = as.matrix(testData)
svmguide1 = as.matrix(trainData)
svmguide1[,1:5]
b=0.01
for(i in 1:5){
  for(j in 1:5){
    if(j==1){
      a=1
    }
    else if(j==2){
      a=2
    }
    else if(j==3){
      a=3
```

```
  }
  else if(j==4){
    a=4
  }
  else {
    a=100
  }
  dcsvm.model = dcSVM(x = svmguide1[,-1],
                      y =svmguide1[,1],verbose = FALSE,
                      k = 3, max.levels = 4, seed = 512, cost = b,
                      gamma = a, kernel = 2,early = 0, m = 800,
                      valid.x = svmguide1.t[,-1], valid.y =
                      svmguide1.t[,1])
  preds = dcsvm.model$valid.pred
  table(preds, svmguide1.t[,1])
  d=dcsvm.model$valid.score #akurasi
  c=dcsvm.model$time$total.time #time process total
  preds = dcsvm.model$valid.pred
  e=table(preds, svmguide1.t[,1]) #confusion matrix
  f=e[1,1]/(e[1,1]+e[1,2]+e[1,3]) #sensityfity kelas positif
  g=e[3,3]/(e[3,3]+e[2,1]+e[3,1]) #specificity kelas positif
  preds1= as.numeric(preds)
  AUC = multiclass.roc(svmguide1.t[,1], preds1,
              levels=c(1, 2, 3)) #Area Under Curve
   print = cat("Akurasi:",d,"\n","C:",b,"\n","Bobot:",a,"\n",
             "Time:",c, "\n", "sensitifity:",f,"\n","spesificity:",g,"\n",
             "tabel confussion :", "\n",e)
  print(AUC)
}
b=b*10
```

BIODATA PENULIS

Penulis dilahirkan di Tulungagung, 18 Mei 1996 dengan nama lengkap Shindi Shella May Wara, biasa dipanggil Shisel. Penulis menempuh pendidikan formal di SDN 2 Plosokandang, SMPN 1 Tulungagung, dan SMAN 1 Boyolangu. Kemudian penulis diterima sebagai mahasiswa Departemen Statistika ITS pada tahun 2015. Selama masa perkuliahan, penulis aktif di divisi Statistics Computer Course (SCC) Himpunan Mahasiswa Statistika ITS (HIMASTA-ITS) sebagai staff *pubic relation* periode 2016-2017 dan ketua divisi pada periode 2017-2018. Selain itu, penulis pernah melakukan magang di PT. Telekomunikasi Selular (Telkomsel) cabang Surabaya. Bagi pembaca yang ingin berdiskusi, memberikan saran, dan kritik mengenai Tugas Akhir ini dapat disampaikan melalui email shindishella@gmail.com atau melalui nomor 082330966581.