ANALISIS SENTIMEN PADA *REVIEW* APLIKASI BERITA *ONLINE* MENGGUNAKAN METODE *MAXIMUM ENTROPY*

(Studi Kasus: *Review* Detikcom pada *Google Play* 2019)

TUGAS AKHIR

Annisa Alwasi'a

14 611 247

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2020

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul	:	Analisis Sentimen pada Aplikasi Berita *Online* Menggunakan Metode *Maximum Entropy* (Studi Kasus: *Review* Detikcom pada *Google Play* Tahun 2019)
Nama Mahasiswa	:	Annisa Alwasi'a
NIM	:	14 611 247

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta, 13 Mei 2020

Pembimbing

(Jaka Nugraha, S.Si., M.Si., Dr.)

HALAMAN PENGESAHAN

TUGAS AKHIR

ANALISIS SENTIMEN PADA REVIEW APLIKASI BERITA *ONLINE* MENGGUNAKAN METODE *MAXIMUM ENTROPY*

(Studi Kasus: *Review* Detikcom pada *Google Play* tahun 2019)

Nama Mahasiswa : Annisa Alwasi'a

NIM : 14 611 247

TUGAS AKHIR INI TELAH DIUJIKAN
PADA TANGGAL: 13 Mei 2020

Nama Penguji:	**Tanda Tangan**
1. Rohmatul Fajiyah, S.Si., M.Si., Dr.techn	……..
2. Ayundyah Kesumawati, S.Si., M.Si.	……..
3. Jaka Nugraha, S.Si., M.Si., Dr.	……..

Mengetahui,
Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Prof. Riyanto, S.Pd., M.Si., Ph.D.

KATA PENGANTAR

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

Assalamu'alaikum wa rahmatullahi wa barakaatuh

Alhamdulillah puji syukur penulis panjatkan kehadirat Allah SWT atas segala rahmat dan hidayah yang telah diberikan oleh-Nya. Kemudian shalawat dan salam juga dicurahkan tercurah kepada Nabi Muhammad SAW atas petunjuk untuk selalu berada di jalan yang diridhoi-Nya sehingga penulis dapat menyelesaikan laporan ini. Laporan Tugas Akhir / Skripsi yang berjudul "Analisis Sentimen pada *Review* Aplikasi Berita *Online* Menggunakan Metode *Maximum Entropy* (Studi Kasus: *Review* Detikcom pada *Google Play* Tahun 2019)" ini adalah sebagai salah satu persyaratan dalam menempuh gelar Sarjana Statistika (S.Stat) di Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia.

Adapun terselesaikannya Laporan Tugas Akhir / Skripsi ini tidak terlepas dari bimbingan dan dukungan berbagai pihak. Oleh karenanya, pada kesempatan ini dengan segala rasa syukur penulis ingin mengucapkan terima kasih yang sebesar-besarnya kepada,

1. Bapak Fathul Wahid, S.T., M.Sc., Ph.D. selaku Rektor Universitas Islam Indonesia, Yogyakarta.
2. Bapak Prof. Riyanto, S.Pd., M.Si., Ph.D. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam, beserta seluruh jajarannya.
3. Bapak Dr. Edy Widodo, S.Si., M.Si. selaku Ketua Prodi Statistika yang telah memberikan dukungan dan motivasi kepada penulis.
4. Bapak Dr. Jaka Nugraha, S.Si., M.Si. selaku dosen pembimbing yang selalu sabar dalam membimbing dan memberikan arahan serta saran selama penulisan Tugas Akhir ini.
5. Seluruh Dosen dan Staff Administasi Universitas Islam Indonesia yang telah banyak mengajarkan ilmu dan senantiasa membantu penulis.

6. Kedua orang tua tersayang, Bapak Heri dan Ibu Titin yang telah memberikan kepercayaan kepada peneliti untuk kuliah dan selalu memberikan semangat, dukungan serta doa hari hingga bisa lulus jenjang sarjana ini.
7. Keluarga Besar yang senantiasa memberikan doa yang terbaik untuk peneliti.
8. Sahabat-sahabat seperjuangan Lia, Nur, Boki, Ayu, Rabi, Reny, Dhea, Khusnul, Sari, Edwika, Maulida, Zarina, dan Ditia yang selalu memberikan semangat dan nasihat.
9. Sahabat dan teman-teman penulis yang telah memberikan dukungan dan saran kepada penulis selama menyelesaikan Laporan Tugas Akhir.

Terimakasih kepada semua pihak yang telah membantu, semoga Allah SWT tiada henti selalu memberikan rahmat, hidayah dan anugerah-Nya kepada mereka semua.

Demikian laporan Tugas Akhir ini, penulis menyadari bahwa masih banyak kekurangan karena keterbatasan pengetahuan dan kemampuan dalam penyusunan laporan ini. Oleh karena itu penulis mengharapkan kritik dan saran dari pembaca untuk menyempurnakan penulisan laporan ini. Semoga laporan ini dapat memberikan manfaat kepada penulis dan semua pihak yang membaca laporan ini. Amin.

Wassalamu alaikum wa rahmatullahi wa barakaatuh.

Yogyakarta, 13 Mei 2020

Annisa Alwasi'a

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Laporan Tugas Akhir ini tidak terdapat karya yang telah diajukan sebagai persyaratan mendapatkan gelar kesarjanaan pada Perguruan Tinggi lain dan sejauh sepengetahuan saya juga tidak terdapat karya yang pernah dituliskan maupun diterbitkan pihak lain, terkecuali karya yang telah diacukan dalam naskah ini dan disebutkan dalam Daftar Pustaka.

Yogyakarta, 13 Mei 2020

Annisa Alwasi'a

ABSTRAK

ANALISIS SENTIMEN PADA REVIEW APLIKASI BERITA *ONLINE* MENGGUNAKAN METODE *MAXIMUM ENTROPY*

(Studi Kasus: *Review* Detikcom pada *Google Play* Tahun 2019)

Annisa Alwasi'a

Program Studi Statistika, Fakultas MIPA

Universitas Islam Indonesia

Kemajuan teknologi mendorong berbagai perkembangan di seluruh aspek masyarakat. Keberadaan internet menjadi faktor pendorong utama perkembangan teknologi informasi dan komunikasi, terutama pada teknologi informasi digital yang memicu bermunculannya media baru seperti portal berita *online*. Detikcom merupakan salah satu portal berita *online* yang paling terkemuka di Indonesia. Dalam rangka terus menjaga dan memperbaiki peforma Detikcom, penilaian publik terhadap layanan dan berita yang disajikan menjadi sangat penting. Adapun penilaian publik dapat dilihat dari situs *Google Play* pada kolom ulasan pengguna. Analisis sentimen dapat digunakan untuk menganalisa ulasan tersebut dengan cara pengklasifikasian antara setimen positif dan negatif. Data ulasan pengguna Detikcom selama tahun 2019 kemudian dilakukan pelabelan dan dianalisis menggunakan metode algoritma *Maximum Entropy*. Hasil klasifikasi sentimen diperoleh tingkat akurasi yang cukup tinggi sebesar 95,69% dengan kinerja sistem dalam pengklasifikasian kelas positif sebesar 97,45% dan kelas negatif sebesar 86,17%. Selanjutnya, dari asosiasi teks diperoleh informasi yang berkaitan dengan topik / kata yang sering dibicarakan oleh pengguna aplikasi Detikcom yakni bagus, berita, oke, *update*, dan mantap untuk kelas positif dan kata tidak, berita, iklan, Detik, dan aplikasi untuk kelas negatif. Informasi berguna untuk melihat keunggulan dan kelemahan aplikasi dimana hasil ulasan negatif tersebut dibuat ke dalam diagram *fishbone* sebagai proses pemecahan masalah.

Kata Kunci: Aplikasi Detikcom, Analisis Sentimen, Klasifikasi, *Maximum Entropy* (Maxent), Asosiasi Teks, Diagram *Fishbone.*

ABSTRACT

SENTIMENTS ANALYSIS OF ONLINE NEWS APPLICATIONS REVIEW USING THE MAXIMUM ENTROPY METHOD

(*Case Study: Detikcom Review on Google Play in* 2019)

Annisa Alwasi'a

Department of Statistics, Faculty of Mathematics and Natural Sience

Islamic University of Indonesia

Technological advancements encourage various developments in all aspects of society. The existence internet is a major driving factor in the development of information and communication technology, especially in digital information technology which has triggered the emergence of new media such as online news. Detikcom is one of the best online news in Indonesia. In order to maintain and improve the performance of Detikcom, public opinion of the services and news presented becomes very important. Public opinion can be seen from Google Play site in review column. Sentiment analysis can be used to analyze these reviews by classifying between positive and negative class sentiment. Detikcom review data during 2019 then be labeled and analyzed using the Maximum Entropy algorithm method. Sentiment classification results obtained a fairly high level of accuracy at 95.69% with the system performance in the classification of positive classes at 97.45% and negative classes at 86.17%. Furthermore, from the text association can be known information relating to topics / words that are often discussed, i.e. good, news, ok, updated, and great for positive classes and words no, news, advertisements, seconds, and applications for negative classes. Then information can be used to look at the strengths and weaknesses of applications where the results of negative reviews be made fishbone diagrams as problem solving.

Keywords: *Detikcom Application, Sentiment Analysis, Classification, Maximum Entropy (Maxent), Association, Fishbone Diagram.*

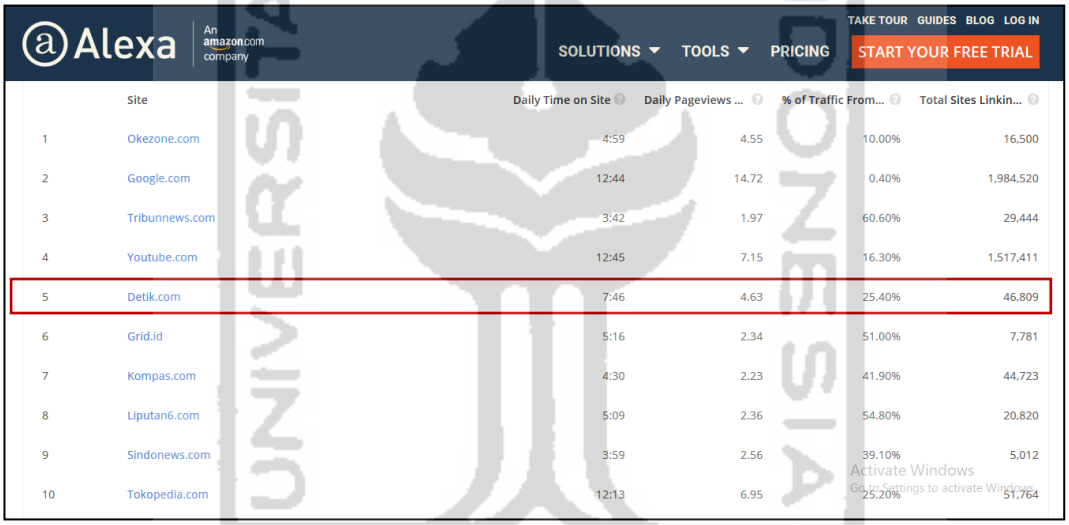
BAB 1 PENDAHULUAN

1.1 Latar Belakang Masalah

Seiring dengan berkembangnya zaman, perkembangan teknologi informasi dan komunikasi berkembang sangat pesat. Terobosan-terobosan teknologi terbaru dibuat untuk mempermudah kehidupan sehari-hari, baik dalam menyelesaikan pekerjaan ataupun memperoleh informasi. Keberadaan internet menjadi salah satu faktor pendorong utama perkembangan teknologi informasi dan komunikasi terutama pada teknologi informasi digital. Penggabungan antara informasi dan teknologi inilah yang melahirkan jurnalisme baru yaitu jurnalisme *online*. Berita sebagai salah satu media penyalur informasi mengenai sesuatu yang sedang terjadi tidak luput dari perkembangan teknologi tersebut. Hal tersebut terlihat dari bermunculannya media baru yang memanfaatkan teknologi informasi dengan membuat portal berita *online*.

Dewasa ini, berita *online* telah menjadi salah satu media massa yang paling sering dikonsumsi oleh masyarakat. Media ini pun mampu mengalahkan media-media generasi sebelumnya seperti media elektronik dan media cetak. Menurut Yunus (2010), laporan berita merupakan tugas profesi wartawan, saat berita dilaporkan oleh wartawan, laporan tersebut menjadi fakta dan ide terkini yang dipilih secara sengaja oleh redaksi pemberitaan atau media untuk disiarkan dengan anggapan bahwa berita tersebut dapat menarik masyarakat umum. Maka dari itu, laporan berita secara tidak langsung menjadi sebuah cerminan dari pendapat masyarakat saat itu. Keungulan media *online* dibandingkan media cetak pada umumnya terletak pada sifatnya yang *up to date, real time* dan praktis. Media *online* bersifat *up to date* karena dapat melakukan pembaharuan informasi dari waktu ke waktu. Media *online* bersifat *real time* karena dapat menyajikan berita atau informasi seiring dengan peristiwa yang ditemukan. Media *online* bersifat praktis karena media *online* dapat diakses di mana dan kapan saja sejauh didukung oleh teknologi internet (Yunus, 2010).

Di Indonesia sendiri menurut Ketua Komisi Penelitian, Pendataan dan Ratifikasi Dewan Pers Indonesia, Ratna Komala memaparkan bahwa terdapat sekitar 43 ribu portal media berita *online* yang terdata. Namun, dari jumlah tersebut hanya 500 yang resmi terdaftar oleh Dewan Pers. Kemudian dari 500 media *online* tersebut sudah tercatat ada 78 media yang terverifikasi faktual dan administrasi dan hanya ada tujuh media *online* yang lolos semua verifikasi tersebut. Media *online* yang lolos tersebut antara lain, Okezone.com, Detikcom, Kompas.com, Metrotvnews.com, Viva.co.id, RMOL.com, dan Arah.com (Detik, 2018). Saat ini sudah terdapat 1.301 portal berita yang resmi terdaftar oleh Dewan Pers dan hanya 529 media yang terverifikasi administrasi dan faktual (Dewan Pers, 2020).



Gambar 2.1. *Website* yang paling sering dikunjungi di Indonesia versi Alexa.

Detikcom merupakan salah satu media pertama yang lolos semua verifikasi tersebut dan masuk kedalam 10 besar website yang paling sering dikunjungi di Indonesia berdasarkan data Alexa Maret 2020. Seiring meningkatnya pengguna internet di Indonesia meningkat pula permintaan masyarakat akan berita. Hal ini, berdampak juga pada persaingan terhadap media berita *online* lainnya. Seperti yang terjadi pada Mei 2019, media Okezone berhasil menggeser peringkat Detikcom yang sebelumnya menduduki peringkat 2 sebagai portal media *online* di Indonesia (Okezone, 2019).

Aplikasi Detikcom yang terdapat disitus *Google Play* masih tercatat sebagai salah satu top posisi sebagai aplikasi berita. *Google Play* sendiri merupakan layanan distribusi digital yang diperkasai oleh *Google* yang berfungsi sebagai toko apikasi resmi dari sistem operasi *android* (Wikipedia, 2020). Salah satu fitur dari *Google Play* adalah keberadaan kolom komentar bagi pengguna untuk memberikan penilaian kepada aplikasi berupa *rating* dan ulasan atau *review*. Hal ini dimanfaatkan untuk melihat penilaian publik berdasarkan ulasan dari pengguna aplikasi Detikcom. Ulasan dari pengguna aplikasi umumnya dapat berupa ulasan positif seperti apresiasi dan saran maupun ulasan negatif seperti keluhan. Hal ini dapat mencerminkan sedikit banyak mengenai kinerja Detikcom di mata para pengguna.

Dalam proses mengumpulkan maupun mensortir ulasan tersebut, tentunya bukan hal yang mudah dimana jumlah ulasan yang terdapat dalam media sosial umumnya berjumlah sangat banyak. Oleh karenanya, suatu metode ataupun teknik khusus diperlukan untuk mengumpulkan maupun mengolah data informasi dalam jumlah yang besar. Menurut Marres dkk (2013), salah satu teknik yang cocok dalam pengumpulan data informasi tersebut yakni menggunakan teknik *scraping.* Teknik *scraping* merupakan proses pengambilan sebuah dokumen semi-terstuktur dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman di internet. Adapun dalam penelitian ini teknik *scraping* digunakan untuk mengumpulkan data *review* atau ulasan yang diberikan oleh pengguna aplikasi Detikcom.

Pada proses analisis dilakukan analisis sentimen untuk mengidentifikasi penilaian pengguna Detikcom. Ulasan pengguna akan diklasifikasi guna mengidentifikasi mana ulasan yang bersifat positif dan negatif. Salah satu metode yang dapat digunakan adalah metode klasifikasi *machine learning.* Klasifikasi *machine learning* merupakan salah satu bidang *data mining*. *Data mining* sendiri adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar (Davies dkk, 2004). *Text mining* sebagai salah satu cabang *data mining* memegang peran penting dalam

analisis yang bersifat tidak terstruktur terutama data yang berbentuk teks (Xiang dkk, 2015). Menurut Feldman (2007), berdasarkan tujuan penggunaanya *text mining* terbagi atas *text clustering*, *text categorization,* dan *sentiment analysis,* sehingga *text mining* sesuai digunakan untuk penelitian ini.

Menurut Han dkk (2001), klasifikasi merupakan proses menemukan suatu pola dari kumpulan data yang berguna dalam memprediksi data yang belum memiliki kelas data tertentu. Sebagai algoritma klasifikasi yang melibatkan metode statistik, *machine learning* dan manajemen *database*, juga dikategorikan sebagai kunci elemen dalam interprestasi data dan visualisasi data (Berry, 2010). Metode klasifikasi *machine learning* yang sering digunakan antara lain *Naïve Bayes Classifer, Support Vector Machine, Lexicon Based, Neural Network, K-Nearest Neighbors, Decision Tree, Random Forest* dan *Maximum Entropy*. Adapun proses klasifikasi *machine learning* akan dilakukan pada penelitian ini menggunakan metode algoritma *Maximum Entropy.* Metode *Maximum Entropy* adalah metode klasifikasi yang mampu mencari distribusi *p(a|b)* yang akan memberikan nilai *entropy* tertinggi dengan tujuan mendapatkan distribusi probabilitas terbaik yang paling mendekati kenyataan (Pratiwi, 2018).

Setelah melakukan klasifikasi, dilakukan proses ekstraksi dan eksplorasi informasi seluas-luasnya dari masing-masing klasifikasi sentimen positif dan sentimen negatif. Proses analisis yang dilakukan menggunakan statistik deskriptif untuk memperoleh gambaran umum dari aplikasi Detikcom maupun asosiasi antarkata untuk menemukan topik atau bahasan yang umumnya diulas oleh pengguna beserta informasi yang berkaitan dengan topik atau bahasan tersebut. Analisis tersebut berguna untuk melihat keunggulan dan kelemahan aplikasi dimana hasil ulasan sentimen negatif dapat diolah menggunakan diagram *fishbone* untuk menemukan pemecahan masalah. Hasil analisis dari penelitian ini diharapkan dapat mengolah data ulasan atau *review* dengan maksimal sehingga memberikan informasi seluas-luasnya dan dapat berguna bagi berbagai pihak yang membutuhkan.

1.2 Rumusan Masalah

Adapun pada penelitian ini terdapat beberapa rumusan masalah yang diperoleh berdasarkan pemaparan latar belakang tersebut adalah sebagai berikut.

1. Bagaimana gambaran umum mengenai perkembangan aplikasi berita *online* Detikcom?
2. Bagaimana pengklasifikasian ulasan aplikasi berita *online* Detikcom menggunakan metode algoritma *Maximum Entropy*?
3. Bagaimana ketepatan *machine learning* dari metode algoritma *Maximum Entropy* dalam mengkasifikasian teks?
4. Bagaimana performa aplikasi berita *online* Detikcom berdasarkan informasi dari ulasan penguna aplikasi?

1.3 Batasan Masalah

Adapun pada penelitian ini terdapat beberapa batasan masalah yang diberikan adalah sebagai berikut.

1. Data yang dianalisis hanya berasal dari ulasan pada *Google Play* untuk aplikasi Detikcom tahun 2019.
2. *Software* yang digunakan dalam analisis adalah R (x64) 3.5.3.
3. Pada data yang tidak dapat diperoses dengan *software*, dilakukan secara manual oleh peneliti dan didasarkan prespektif peneliti.
4. Proses iterasi pada metode evaluasi hanya terbatas sampai iterasi ke 5.

1.4 Tujuan Penelitian

Adapun pada penelitian ini terdapat beberapa tujuan yang diperoleh berdasarkan pemaparan rumusan masalah sebagai berikut.

1. Mengetahui gambaran umum mengenai perkembangan aplikasi berita *online* Detikcom.
2. Mengetahui pengklasifikasian aplikasi berita *online* Detikcom menggunakan metode algoritma *Maximum Entropy*.
3. Mengetahui ketepatan *machine learning* dengan menggunakan metode *Maximum Entropy* dalam mengklasifikasikan teks.

4. Mengetahui performa aplikasi berita *online* Detikcom menurut ulasan penguna aplikasi.

1.5 Manfaat Penelitian

Adapun pada penelitian ini terdapat beberapa manfaat yang diberikan adalah sebagai berikut:

1. Memberikan informasi mengenai gambaran umum dari pola ulasan berita *online* Detikcom.
2. Menambah pengetahuan tentang pengklasifikasian *machine learning* terutama untuk metode *Maximum Entropy*.
3. Mempermudah pihak Detikcom dalam mengetahui tanggapan publik mengenai keunggulan dan kelemahan dari aplikasi.
4. Memberikan beberapa pilihan solusi yang dapat digunakan untuk meningkatkan kualitas dan evaluasi dari pelayanan media Detikcom di masa yang akan datang.

1.6 Sistematika Penulisan

Adapun sistematika yang digunakan pada proses penulisan dalam penelitian ini sebagai berikut.

BAB I PENDAHULUAN

Pada bab ini akan dipaparkan mengenai latar belakang dari penelitian ini, sehingga dapat dirumuskan masalah, batasan-batasan masalah, tujuan dan manfaat penelitan ini, serta sistematika penulisan yang digunakan.

latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Pada bab ini akan dipaparkan penelitian-penelitian terdahulu yang berhubungan dengan permasalahan yang diteliti dan menjadi acuan konseptual

BAB III LANDASAN TEORI

Pada bab ini akan dibahas tentang teori-teori dan konsep yang berhubungan dengan penelitian yang dilakukan dan mendukung dalam

pemecahan masalahnya. Selain itu, bab ini juga memuat teori-teori dalam proses mengumpulkan, mengolah, dan menganalisa data.

BAB IV METODOLOGI PENELITIAN

Pada bab ini akan dipaparkan populasi dan sampel, variabel penelitian, jenis dan sumber data, metode analisis, dan tahapan penelitian.

BAB IV ANALISIS DAN PEMBAHASAN

Pada bab ini akan dibahas mengenai analisa yang dilakukan terhadap hasil pengumpulan, pengolahan dan analisa data yang diperoleh dari hasil penelitian.

BAB IV PENUTUP

Pada bab ini akan dipaparkan kesimpulan yang diperoleh dari hasil penelitian dan analisis data yang telah dilakukan serta saran-saran yang dapat diterapkan dari hasil pengolahan data yang dapat mendi masukan yang berguna kedepannya.

BAB 2 TINJAUAN PUSTAKA

Penulisan skripsi ini adalah suatu pemikiran baru yang menggunakan dasar dari beberapa pemikiran terlahulu sebagai acuan dan dasar penelitian. Hal ini bermanfaat untuk menunjukkan bahwa penelitian ini mempunyai arti penting sehingga dapat diketahui kontribusi penelitian tehadap ilmu pengetahuan.

Adapun beberapa penelitian terdahulu seperti yang dilakukan oleh Putranti dkk (2014) terhadap sentimen pengguna *twitter* pada teks berbahasa Indonesia menggunakan *Maximum Entropy* (Maxent) dan *Support Vector Machine* (SVM). Proses klasifikasi dokumen tekstual dilakukan dalam dua kelas sentimen, yaitu kelas sentimen positif dan negatif. Tujuan penelitian ini untuk membantu riset pasar atas opini publik terutama sentimen mengenai objek tertentu yang disampaikan di *twitter* dalam Bahasa Indonesia. Metode algoritma Maxent menggunakan POS *tagger* SVM sebagai dasar dalam pembangunan model klasifikasi. Berdasarkan hasil implementasi klasifikasi diperoleh akurasi 86,81% untuk pengujian 7-*fold cross validation* untuk tipe sigmoid. Pelabelan kelas secara manual dengan POS *tagger* menghasilkan akurasi 81,67%.

Pada penelitian oleh Ulwan M.N (2016) dilakukan penelitian terhadap *pattern recognition* data teks menggunakan *Support Vector Machine* (SVM) dan *association*. Data yang digunakan adalah laporan masyarakat yang berupa data teks yang tidak terstruktur yang diklasifikasikan menjadi tiga kelas yaitu Aspirasi, Keluhan, dan Peryataan dari Layanan Aspirasi dan Pengaduan Online Rakyat (LAPOR!). Hasil klasifikasi menunjukan tingkat akurasi sebesar 96,7%. Secara umum metode text mining menunjukan hasil ekstaksi informasi pada kelas aspirasi adalah terkait penertiban terhadap PSK, PKL, asap, merokok, *busway*, dan pembagian bantuan masyarakat. Pada kelas keluhan, masyarakat mengeluhkan tentang pembagian BLSM atau KPS yang tidak merata, masalah macet, layanan Telkom yang buruk, serta *busway* yang sering bermasalah. Sedangkan pada kelas pertanyaan yang menjadi hal yang sering ditanyakan adalah masalah BLSM dan

KPS serta seputar informasi mengenai agama, BPJS, beasiswa, setifikasi, dan tunjangan.

Pada penelitian oleh Prakoso dkk (2017) dilakukan penelitian terhadap perbandingan sentimen menggunakan metode *Support Vector Machine* (SVM) dan *Maximum Entropy* (Maxent). Analisis sentimen dilakukan dengan memberikan nilai positif atau negatif pada postingan *twitter* yang membicarakan nama calon dalam pilkada DKI. Kemudian diambilan fitur daripada setiap *tweet* yang dikumpulkan menjadi sebuah list fitur. List fitur ditransformasi menjadi *feature vector* dengan bentuk binary menggunakan metode TF-IDF. Dataset terdiri dari 2 data yaitu data training dan testing, dimana data training diberikan label secara manual. Hasil pengujian keakuratan peforma algoritma dilakukan dengan menggunakan metode *K-Fold Validation* dengan komposisi 90:10. Nilai akurasi yang diperoleh dari metode Maxent adalah rata-rata sebesar 74%, sedangkan metode SVM adalah rata-rata sebesar 75% dengan kernel optimal adalah kernel linier.

Pada penelitian oleh Syah dkk (2017) dilakukan penelitian terhadap perbandingan *review* produk toko *online* metode *Maximum Entropy*. Penelitian ini bertujuan untuk membuat sistem yang dapat melakukan klasifikasi ulasan baik ulasan positif maupaun ulasan negatif. Data yang digunakan yakni ulasan produk Amazon untuk kategori *cell phone* dan *accessories*. Klasifikasi dilakukan dengan menggunakan metode *Maximum Entropy* dan metode TF-IDF untuk mendapatkan fitur dari produk ulasan tersebut. Hasil evaluasi menggunakan nilai presisi, *recall*, f-1 *measure* diperoleh percobaan terbaik pada iterasi ke 1000 dengan nilai akurasi sebesar 83% dan f-1 *measure* sebesar 90,0754%.

Pada penelitian oleh Praptiwi D.Y (2018) dilakukan penelitian terhadap ulasan pada *Google Play* untuk salah satu *e-commerce* yakni Bukalapak menggunakan metode *Support Vector Machine* (SVM) dan *Maximum Entropy* (Maxent). Metode tersebut dapat melakukan pengkategorian terhadap ulasan-ulasan secara otomatis, baik ulasan positif maupun ulasan negatif. Pada metode SVM nilai akurasi yang diperoleh sebesar 91,95%, sedangkan metode Maxent nilai akurasi yang diperoleh sebesar 92,98%. Berdasarkan asosiasi teks diperoleh

analisis dari kata-kata terkait kata barang, transaksi, fitur, pelayanan, pesanan, pengiriman, respon, berbelanja, akulaku, kebutuhan, dan cicilan untuk kelas sentimen positif, dan kata barang, *update*, server, *chat*, email, transaksi, *upload*, promo, voucer, bukadompet, dan upgrade utuk kelas sentimen negatif. Kemudian hasil ulasan negatif digunakan untuk mencari pemecahan masalah menggunakan diagram *fishbone.*

Pada penelitian oleh Gumilang Z.A (2018) dilakukan penelitian terhadap ulasan pada *Google Play* untuk salah satu *e-commerce* yakni Shopee menggunakan metode *Naïve Bayes Clasifer*. Adapun hasil klasifikasi sentimen dengan menggunakan algoritma *Naïve Bayes Clasifer* diperoleh tingkat akurasi sebesar 97,4%. Kemudian, pada proses asosiasi teks didapatkan informasi mengenai topik yang sering dibicarakan oleh pengguna Shopee antara lain aplikasi, ongkir, harga, dan puas untuk kelas sentimen positif dan informasi mengenai topik aplikasi ongkir, buruk, susah, gambar, pengiriman, dan penjual untuk kelas sentimen negatif. Hasil ulasan negatif digunakan untuk mencari pemecahan masalah menggunakan diagram *fishbone.*

Pada penelitian oleh Sabily dkk (2019) dilakukan penelitian terhadap perbandingan sentimen pemilihan Presiden pada *twitter* tahun 2019 dengan menggunakan *Maximum Entropy*. Analisis sentimen dilakukan dengan memberikan nilai positif atau negatif pada postingan yang membicarakan mengenai pemilihan presiden di *twitter*. Metode yang digunakan adalah *Maximum Entropy* dengan metode evaluasi *Confusion Matrix* yang nantinya akan menghitung nilai Macro dan Micro *averaging* dari nilai evaluasi yang dihasilkan. Pengunjian dilakukan dengan menggunakan data traning sebesar 300 *tweet* dan data testing sebanyak 120 *tweet*. Hasil klasifikasi memberikan nilai akurasi Makro yang cukup tinggi yakni sebesar 89.16% dengan tingkat evaluasi untuk nilai presisi sebesar 100%, nilai *recall* sebesar 89,16%, dan nilai F-*measure* sebesar 94,27%.

Adapun penelitian-penelitian tersebut dapat dibandingkan dalam tabel berikut.

Tabel 2.1. Perbandingan penelitian terdahulu

No	Nama Peneliti	Judul Penelitian	Metode	Isi
1.	Putranti dkk (2014)	Analisis sentimen *twitter* untuk teks berbahasa indonesia dengan *maximum entropy* dan *support vector machine*	*Maximum Entropy* (ME) dan *Support Vector Machine* (SVM)	Mengklasifikasikan *tweet* ke dalam kelas positif dan negatif. Hasil menunjukkan ME digunakan untuk POS *tagger* menghasilkan akurasi 81,67% dan SVM menghasilkan akurasi 86,81 % pada pengujian 7-*fold cross* validation untuk tipe *kernel Sigmoid*
2.	Prakoso dkk (2017)	Analisis sentimen menggunakan *support vector machine* dan *maximum entropy*	*Maximum Entropy* (ME) dan *Support Vector Machine* (SVM)	Mengklasifikasikan *tweet* ke dalam kelas positif dan negatif. Hasil pengujian dengan metode *K-Fold Cross Validation* menunjukkan metode SVM memberikan rata-rata tingkat akurasi sebesar 75% sedangkan ME sebesar 74%
3.	Syah dkk (2017)	*Sentiment analysis on online store product reviews with maximum entropy method*	*Maximum Entropy* (ME)	Mengklasifikasikan *review* ke dalam kelas positif dan negatif. Hasil evaluasi menggunakan ME menghasilkan nilai presisi, *recall*, dan *f-measure* yang tinggi. Hasil percobaan terbaik diperoleh akurasi 83% dan *f-measure* 90,70%
4.	Pratiwi DY (2018)	Analisis sentimen *online review* pengguna *e-commerce* menggunakan metode *support vector machine* dan *maximum entropy*	*Maximum Entropy* (ME) dan *Support Vector Machine* (SVM)	Mengklasifikasikan *review* ke dalam kelas positif dan negatif. Hasil menunjukkan metode ME memberikan tingkat akurasi yang lebih tinggi sebesar 92,98% dibanding SVM sebesar 91,95%
5.	Sabily dkk (2019)	Analisis sentiment pemilihan presiden 2019 pada twitter menggunakan metode *maximum entropy*	*Maximum Entropy* (ME)	Mengklasifikasikan *twitter* ke dalam kelas positif dan negatif. Hasil menunjukkan metode ME memberikan tingkat akurasi sebesar 89,16% dengan nilai *precision, recall* dan *f-measure* sebesar 100%, 89,16% dan 94,27%

BAB 3 LANDASAN TEORI

3.1 Berita

Menurut Stephens (2007), berita adalah informasi baru atau informasi mengenai sesuatu yang sedang terjadi, disajikan lewat bentuk cetak, siaran, internet, atau dari mulut ke mulut kepada orang ketiga atau orang banyak. Berita bertujuan untuk menyampaikan kepada masyarakat fakta / ide yang dianggap penting. Namun tidak semua fakta dapat disampaikan begitu saja oleh media. Laporan berita merupakan tugas profesi wartawan, saat berita dilaporkan oleh wartawan laporan tersebut menjadi fakta / ide terkini yang dipilih secara sengaja oleh redaksi pemberitaan / media untuk disiarkan dengan anggapan bahwa berita yang terpilih dapat menarik khalayak banyak karena mengandung unsur-unsur berita. Biasanya berita tidak saja menginformasikan mengenai kejadian dan peristiwa terjadi, namun dapat pula menjadi media yang memberi pengaruh terhadap para para pembaca.

3.1.1 Jenis Berita

Adapun jenis-jenis berita dapat dibagi menjadi sebagai berikut.

1) *Straight news,* yakni berita yang ditulis secara langsung, singkat lugas, dan apa adanya Umumnya sebagain besar bagian halaman depan surat kabar berisi berita seperti ini.
2) *Hard news,* yakni berita dapat bernilai lebih, berkualitas, dan ter*update.* Karena sangat penting maka harus segera disampaikan dan diketauhi oleh masyarakat. Biasanya berisi berita bersifat khusus atau dapat juga mengenai peristiwa yang terjadi secara tiba-tiba.
3) *Soft news*, yakni berita pendukung atau berita yang isi berita hanya membahas mengenai hal-hal ringan dan tidak lebih tinggi dari *hard news*.
4) *Depth news,* berita yang dikembangkan secara mendalam dengan tujuan membahas suatu masalah dengan lebih dalam.
5) *Investigation news,* yakni berita mengenai penelitian ataupun penyelidikan yang dilakukan dari berbagai macam sumber. *Investigation news* serupa

dengan *depth news*, perbedaannya pada *depth news* hanya melaporkan peristiwa yan terjadi secara mendalam saja.

6) *Interpretative News*, yakni berita yang dikembangkan dengan pendapat maupun penelitian yang dilakukan oleh penulisnya.
7) *Opinion news*, yakni berita tentang pendapat seseorang. Misalnya pendapat mahasiswa, pejabat, para ahli mengenai suatu peristiwa.

3.1.2 Syarat Berita

Adapun syarat-syarat berita dapat dibagi menjadi sebagai berikut.

1) Fakta, yakni peristiwa atau kejadian yang terjadi benar-benar nyata.
2) Terkini, yakni jarak waktu peristiwa dengan penyiaran beirta tidak telalu jauh.
3) Seimbang, yakni tidak memihak kepada suatu pihak tertentu.
4) Lengkap, yakni memenuhi unsur-unsur berita.
5) Menarik, yakni mampu menarik minat pembaca atau pendengarnya.
6) Sistematis, yakni memiliki urutan yang jelas sehingga tidak membingungkan pembaca dalam mengerti isi berita.

3.1.3 Sifat Berita

Adapun sifat-sifat berita dapat dibagi menjadi sebagai berikut.

1) Baru atau aktual, yakni peristiwa yang baru memiliki nilai lebih untuk dijadikan berita jika dibandingkan dengan peristiwa yang sudah lama terjadi.
2) Penting, yakni peristiwa atau hal-hal tersebut berpengaruh pada kehidupan masyarakat. Jadi initinya suatu berita itu harus yang dianggap penting oleh masyarakat.
3) Akibat, yakni peristiwa menjadi berita karena dapat berakibat atau memiliki dampak.
4) Jarak, yakni peristiwa yang terjadi di sekitar pembaca lebih menarik untuk dijadikan berita daripada peristiwa yang terjadi ditempat jauh.
5) Emosi, yakni peristiwa yang disampaikan dapat memberikan emosi kepada pembaca seperti senang, terharu, kecewa, dst.

3.2 Situs Detikcom

Detikcom merupakan salah satu situs web yang berfokus pada penyampaian isi berita dan artikel *online* di Indonesia. Pembentukan Detikcom didirikan oleh Budiono Darsono, Yayam Sopyan, Abdul Rahman, dan Didi Nurgrahadi. Detikcom lahir pada tanggal 9 Juli 1998 sebagai tanggal pertama kali Detikcom daring dengan fitur lengkap, meskipun server Detikcom sudah ada sejak 30 Mei 1998. Awalmula peliputan utama Detikcom hanya berfokus pada masalah berita politik, ekonomi, dan teknologi informasi saja. Detik baru mulai melampirkan berita hiburan dan olahraga setelah situasi politik mereda dan ekonomi di Indonesia mulai membaik. Kemudian pada tanggal 3 Agustus 2011, Detikcom menjadi salah satu anak perusahaan CT Crop yakni baguan dari PT Trans Coporation.

Tidak seperti situs-situs berita berbahasa Indonesia lainnya yang umumnya merupakan perkembangan dari edisi cetak, Detikcom hanya memiliki edisi *online* hanya mempunyai edisi daring sehingga pendapatannya bergantung dari sektor periklanan. Keunggulan Detikcom terletak pada kecepatan *update* dalam hal berita-berita terbaru (*breaking news*). Detikcom sebagai media daring tidak lagi terbatas dalam karateristik yang terdapat dalam media cetak seperti terbatas waktu dalam harian, mingguan, dan bulanan. Hal ini memungkinkan Detikcom menjadi selalu yang terdepan dalam dalam penyiaran *breaking news*. Konsep ini membawa Detikcom maju sebagai salah satu situs penyedia berita informasi digital paling popular di Indonesia.

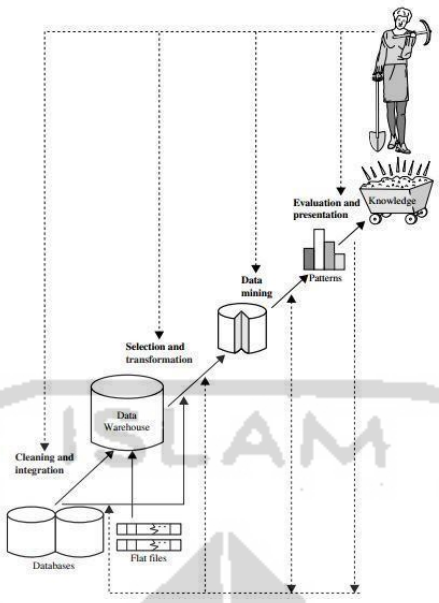
Adapun situs-situs yang termasuk ke dalam jajaran portal Detikcom adalah sebagai berikut.

1) Detiknews (news.detik.com) berisi informasi berita politik-peristiwa
2) Detikfinance (finance.detik.com) berisi berita ekonomi dan keuangan
3) Detikfood (food.detik.com) berisi informasi tentang resep makanan dan kuliner
4) Detikhot (hot.detik.com) berisi info gosip artis/selebriti dan infotainment
5) Detiknet (inet.detik.com) berisi informasi teknologi informasi
6) Detiksport (sport.detik.com) berisi info olahraga termasuk sepak bola

7) Detikhealth (health.detik.com) berisi info dan artikel kesehatan
8) 20detik (tv.detik.com/20detik/) berisi original konten video mulai dari berita sampai gaya hidup
9) Detikfoto (foto.detik.com) berisi berita foto
10) Detikoto (oto.detik.com) berisi informasi mengenai otomotif
11) Detiktravel (travel.detik.com) berisi informasi tentang liburan dan pariwisata
12) Detikevent (event.detik.com) berisi event-event yang diadakan dan kerjasama dengan detikcom
13) Detikforum (forum.detik.com) merupakan tempat diskusi online antar komunitas pengguna detikcom
14) Blogdetik (blog.detik.com) merupakan tempat pengakses mengisi info atau artikel, foto, video di halaman blog pribadi
15) Wolipop (wolipop.detik.com) berisi informasi tentang wanita dan gaya hidup
16) Iklan baris (iklanbaris.detik.com) berisi iklan yang langsung diisi konsumen
17) Pasangmata (pasangmata.detik.com) berisi informasi berita dari pengguna dan dimoderasi oleh admin.

3.3 *Data Minning*

Secara sederhana *data mining* adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar (Davies dkk, 2004). *Data mining* juga disebut sebagai serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data (Pramudiono, 2003). Data mining, sering juga disebut sebagai *knowledge discovery in database (KDD).* Adapun proses KDD ditunjukkan sebagai berikut.



Gambar 3.1. Tahapan *Knowledge Discovery from Data* (KDD)

Menurut Maclennan dkk (2009), fungsi *data mining* dapat dibagi menjadi sebagai berikut.

1. *Classificattion*

   Classificattion untuk mencari model atau fungsi yang menggambarkan dan membedakan kelas-kelas atau konsep data untuk mengklasifikasikan target class ke dalam kategori yang dipilih disebut *classification*. Dalam klasifikasi, terdapat target variabel kategori, misalnya pendapatan tinggi, sedang, dan rendah (Larose, 2005). Ada banyak metode untuk membangun klasifikasi seperti, *Support Vector Machine (*SVM*), Naïve-bayesian, Random forest, Maximum Entropy*, dan *Neighbor classification.*

2. *Clustering*

   Clustering berguna untuk mencari pengelompokkan atribut ke dalam segmentasi-segmentasi berdasarkan similaritas. Menurut Larose (2005), *cluster* berbeda dengan *classification* karena tidak adanya variabel target dalam *cluster*. Algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan *record* dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal.

3. *Association*

   Association berguna untuk mencari keterkaitan antara atribut atau item set, berdasarkan jumlah item yang muncul dari *association rule* yang ada. Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu (Larose, 2005).

4. *Regression*

   Regression berguna untuk mencari prediksi dari suatu pola yang ada, fungsinya hampir menyerupai dengan *classification*.

5. *Forecasting*

   Forecasting berguna untuk peramalan waktu yang akan datang berdasarkan trend yang telah terjadi di waktu sebelumnya.

6. *Sequence Analysis*

   Sequence analysis berguna untuk mencari pola urutan dari rangkaian kejadian.

7. *Deviation Analysis*

   Deviation analysis berguna untuk mencari kejadian abnormal yang sangat berbeda dari keadan umumnya.

3.4 *Text Minning*

Menurut Feldman & Sanger (2007), *text mining* merupakan proses pengalian informasi secara intensif yang bekerja menggunakan alat dan metode tertentu untuk menganalisis suatu kumpulan dokumen. *Text mining* digunakan untuk mendeskripsikan sebuah teknologi yang mampu menganalisis data teks semi-terstruktur maupun tidak terstruktur, sedangkan *data mining* digunakan dalam pengolahan data yang bersifat terstruktur (Jamil, 2017).

Adapun yang paling membedakan antara *text mining* dan *data mining* berada terhadap sumber data yang digunakan. Kesamaan antara keduanya yakni menggunakan data besar dan data berdimensi tinggi dengan struktur yang terus berubah. Pada *text mining*, pola–pola yang diekstrak dari data tekstual yang tidak terstruktur. Sedangkan pada *data mining*, data yang diolah umumnya sudah terstruktur dari proses *warehousing*. Sehingga *text mining* biasanya lebih sulit dari *data mining* karena berkaitan langsung dengan masyarakat dimana memiliki

struktur teks yang kompleks, struktur yang tidak lengkap, bahasa yang berbeda, dan arti yang tidak standar. Maka dari itu digunakan *Natural Languange Processing* untuk analisis teks yang tidak berstruktur tersebut. Secara umum tahap-tahap pada *text mining* dapat dibagi atas *text preprocessing* dan *feature selection* (Feldman & Sanger, 2007).

3.4.1 *Text preprocessing*

Dalam proses *text mining*, teks dokumen yang digunakan harus dipersiapkan terlebih dahulu sebelum dapat digunakan untuk proses utama. Proses mempersiapkan teks dokumen atau dataset mentah disebut juga dengan proses *text preprocessing*. *Text preprocessing* berfungsi untuk mengubah data teks yang tidak terstruktur atau sembarang menjadi data yang terstruktur. Secara umum proses yang dilakukan dalam tahapan preprocessing adalah sebagai berikut.

1. *Spelling Normalization*

   Spelling Normalization adalah proses subtitusi atau perbaikan kata-kata singkatan atau salah ejaan. Subtitusi kata dilakukan untuk menghindari jumlah perhitungan dimensi kata yang melebar. Perhitungan dimensi kata akan melebar jika kata yang salah eja atau disingkat tidak diubah karena kata tersebut sebenarnya mempunyai maksud dan arti yang sama tetapi akan dianggap sebagai entitas yang berbeda pada saat proses penyusunan matriks.

2. *Case Folding*

   Case folding adalah proses penyamaan *case* dalam sebuah dokumen. Hal ini dilakukan untuk mempermudah pencarian. Tidak semua dokumen teks konsisten dalam penggunaan huruf kapital. Oleh karena itu peran *case folding* dibutuhkan dalam mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (dalam hal ini huruf kecil atau *lowercase*).

5. *Tokenizing*

   Tokenizing adalah proses penguraian teks yang semula berupa kalimat-kalimat yang berisi kata-kata. Proses tokenisasi diawali dengan menghilangkan *delimiter-delimiter* yaitu simbol dan tanda baca yang ada pada teks tersebut seperti @, $, &, tanda titik (.), koma (,) tanda tanya (?), tanda seru (!). Proses

pemotongan *string* berdasarkan tiap kata yang menyusunnya, umumnya setiap kata akan terpisahkan dengan karakter spasi, proses tokenisasi mengandalkan karakter spasi pada dokumen teks untuk melakukan pemisahan. Hasil dari proses ini adalah kumpulan kata saja (Putri, 2016).

6. *Filtering*

   Filtering adalah proses mengambil kata-kata penting dari hasil token. Algoritma *stoplist/stopword* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata yang penting) dapat digunakan pada tahap ini. *Stopword* adalah kata-kata yang tidak deskriptif dan bukan merupakan kata penting dari suatu dokumen sehinggga dapat dibuang. Contoh *stopword* adalah "yang", "dan", "di", "dari" dan seterusnya (Putri, 2016). Dalam filtrasi ini menggunakan *stopword* agar kata-kata yang kurang penting dan sering muncul dalam suatu dokumen dibuang sehingga hanya menyisakan kata-kata yang penting dan mempunyai arti yang diproses ke tahap selanjutnya.

3.4.2 *Feature Selection*

Feature Selection merupakan tahap selanjutnya dari proses pengurangan dimensi. Meskipun pada tahap sebelumnya telah dilakukan penghapusan terhadap kata-kata yang tidak deskriptif (*stopword*), tidak seluruh kata-kata yang terdapat dalam dokumen memiliki makna penting. Pada tahap ini dilakukan pemilihan terhadap kata-kata yang relevan dan benar-benar merepresentasikan ini dari dokumen. Pemilihan dilakukan dengan melihat kata-kata yang memiliki intensitas kemunculan tinggi pada dokumen, serta kata-kata yang informatif secara keseluruhan.

3.5 Pembobotan Kata (*Term Weigthing*)

Pembobotan kata atau *term weighting* merupakan salah satu tahapan yang perlu diperhatikan dalam mencari informasi dari koleksi dokumen yang heterogen. Dalam dokumen umunya terdapat kata, frase, atau unit indeks lainnya yang menunjukan konteks dari dokumen tersebut, hal inilah yang disebut *term*. Term weighting digunakan untuk memberikan indicator dari setiap kata sesuai dengan tingkat kepentingan masing-masing kata dalam dokumen (Zafikri, 2008).

Salah satu metode pembobotan *term* terbaru yang paling banyak digunakan adalah metode *Term Frequency – Inverse Document Frequency* (TF-IDF). Dalam TF-IDF, perhitungan bobot *term* dari sebuah dokumen dilakukan dengan menghitung masing-masing nilai *Term Frequency* dan *Inverse Document Frequency.*

Menurut Zafikri (2008), perhitungan nilai TF-IDF dapat dilakukan dengan menggunakan rumus berikut.

1. *Term Frequency* (TF)

   Term Frequency merupakan faktor yang menentukan perhitungan bobot *term* berdasarkan jumlah dan bentuk kemunculan kata pada dokumen. Pada dasarnya dapat dikatakan bahwa semakin besar nilai jumlah kemunculan suatu term, maka semakin besar juga nilai bobot term tersebut dalam dokumen. Adapun perhitungan nilai *Term Frequency* dapat dilakukan dengan beberapa cara sebagai berikut.

   1) TF biner, pemberian bobot *term* dilihat berdasarkan ada tidaknya suatu kata dalam dokumen. Jika terdapat kata tersebut maka diberi nilai satu, jika tidak diberi nilai nol.
   2) TF murni atau *raw TF*, pemberian bobot *term* dilihat berdasarkan jumlah kemunculan suatu kata dalam dokumen. Missal jika kata tersebut muncul tiga kali maka diberi bobot tiga.
   3) TF logaritmit, pemberian bobot *term* pada dokumen yang memiliki sedikit kata dalam *query*, namun mempunyai frekuensi yang tinggi.

   $$tf = 1 + log(tf) \quad (3.1)$$

   4) TF normalisasi, pemberian bobot *term* diperoleh dengan membandingkan frekuensi sebuah kata dengan jumlah seluruh kata dalam dokumen

   $$tf = 0{,}5 + 0{,}05x \left(\frac{tf}{max\ tf}\right) \quad (3.2)$$

2. *Inverse Document Frequency* (IDF)

Inverse Document Frequency merupakan proses mengurangi dominasi *common term* yang sering muncul dalam dokumen. *common term* perlu dihilangkan umumnya kurang bernilai sehingga sering menyebabkananalisis kurang maksimal. Selain itu, IDF juga bertujuan untuk menjaga faktor kejarangmunculan kata (*term scarcity*). Pembobotan dalam IDF dilakukan dengan menghitung nilai faktor kebalikan dari frekuensi dokumen yang mempunyai suatu kata. Adapun perhitungan nilai *Inverse Document Frequency* dapat dilakukan dengan,

$$idf_j = log\left(\frac{D}{df_j}\right) \quad (3.3)$$

dimana

D : jumlah keseluruhan dokumen

df_j : jumlah dokumen yang mempunyai *term* t_j.

Adapun nilai TF-IDF diperoleh dari perkalian nilai *Term Frequency* dengan nilai *Inverse Document Frequency*. Maka pada perhitungan TF-IDF untuk *raw TF* digunakan rumus sebagai berikut.

$$w_{ij} = tf_{ij} \times idf_j \quad (3.4)$$

$$w_{ij} = tf_{ij} \times log\frac{D}{df_j} \quad (3.5)$$

dengan

w_{ij} : bobot *term* t_j terhadap dokumen d_i

tf_{ij} : jumlah kemunculan *term* t_j dalam dokumen d_i

3.6 Analisis Sentimen

Menurut Lee dan Pang (2008), analisis sentiment merupakan proses memperoleh informasi dengan cara memahami, mengekstrak, dan mengolah data tekstual secara otomatis. Analisis sentimen mulai terkenal pada tahun 2013 sebagai salah satu cabang *text mining* yang atau dikenal juga dengan *opinion mining.* Pada dasarnya, analisis sentimen digunakan untuk mengetahui tanggapan dan sikap dari suatu kelompok atau individu terhadap suatu topik bahasan kontekstual

keseluruhan dokumen. Tanggapan dan sikap tersebut dapat berupa pendapat atau penilaian atau evaluasi (teori appraisal), keadaan afektif (keadaan emosional penulis saat menulis), atau komunikasi emosional (efek emosional yang sampai pada pembaca) (Saraswati, 2011).

Secara umum domain yang sering membutuhkan analisis sentimen antara lain produk konsumen, layanan dan jasa, dan peristiwa-peristiwa sosial dan politik yang memerlukan opini publik. Analisis sentimen lebih memiliki kecenderungan terhadap penelitian mengenai pernyataan suatu pendapat yang mempunyai suatu sentimen baik positif atau negatif. Oleh karena pendapat memiliki pengaruh yang tinggi kepada perilaku seseorang, maka dapat dikatakan semua aktivitas seseorang terwakili dari pendapat tersebut. Hal ini terlihat dalam proses pengambilan keputusan dimana umumnya diambil dari pendapat orang-orang. Pada sektor bisnis dan organisasi, pendapat dan opini publik menjadi sangat penting terhadap penilaian suatu produk dan jasa (Liu 2012).

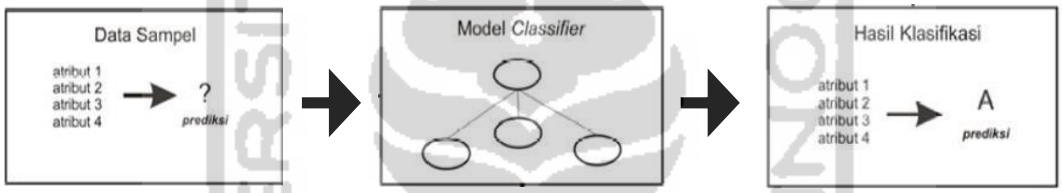
Bagi sektor bisnis, analisis sentimen dapat berguna dalam proses pelacakan produk, jasa, merek, dan target konsumen di pasar. Selain itu analisis sentimen juga dapat menilai keunggulan dan kelamahan suatu produk dan jasa. Secara umum analisis sentimen digunakan untuk mendeteksi keluhan atau rants buruk, persepsi produk atau layanan baru, dan presepsi dari suatu merek tertentu.

3.7 Klasifikasi

Menurut Prasetyo (2012), klasifikasi adalah proses pengelompokan teramati dari suatu objek data ke dalam suatu kelas tertentu berdasarkan kelas-kelas yang ada. Teknik klasifikasi lebih efektif digunakan dalam proses prediksi dan pengambaran suatu kumpulan data untuk jenis kategori biner atau nominal dibandingkan dengan kategori ordinal. Contohnya klasifikasi lebih cocok dalam mengklasifikasi seseorang dengan penghasilan dan tidak berpenghasilan dibandingkan mengklasifikasi seseorang berpenghasilan rendah, menengah, dan tinggi (Tan dkk, 2006).

Menurut Ham dan Kamber (2006), data klasifikasi terbagi menjadi dua proses tahapan. Tahap pertama merupakan *learned model* dimana dilakukan pembangunan model menggunakan hasil analisa *record database* dari serangkaian kelas data yang

ada. Masing-masing *record* diasumsikan mempunyai *predefined class* yang didasarkan pada atribut kelas label. Oleh karena masing-masing *record* memiliki kelas label maka klasifikasi termasuk ke dalam *supervied learning*. Hal inilah yang membedakan antara klasifikasi dan *clustering* yang tidak memerlukan kelas label (*unsupervied learning*). Tahap ini juga sering disebut dengan tahap pembelajaran atau pelatihan. Pelatihan dilakukan dengan menganalisis data latih hingga diperoleh informasi yang dibutuhkan untuk membangun suatu model algoritma klasifikasi. Proses pembangunan tersebut dapat dilihat sebagai proses pembentukan dan pemtaan fungsi $y = f(x)$ dengan y yaitu kelas label hasil prediksi dan x yaitu *record* yang akan diprediksikan. Adapun bagan proses klasifikasi dari data sampel hingga diperoleh hasil prediksi dapat dilihat pada Gambar 3.2 berikut.



Gambar 3.2. Bagan proses klasfikasi (Han & Kamber, 2006)

Menurut Ham dan Kamber (2006), untuk memperoleh hasil klasifikasi yang baik diperlukan beberapa persiapan sebagai berikut.

1. Pembersihan data

   Pembersihan data digunakan untuk mengurangi kecacatan data terutama dalam proses pembangunan model. Pembersihan yang biasa dilakukan adalah menghilangkan data noise, melengkapi data yang hilang, dan seterusnya.

2. Analisa relevansi

   Dalam proses klasifikasi terdapat atribut-atribut yang memiliki tingkat kemiripan tinggi dan sering kali saling berhubungan kuat satu dengan lainnya. Sehingga atribut-stribut tersebut perlu dihilangkan agar tidak mempengaruhi tingkat keoptimalan klasifikasi.

3.7.1 Ukuran Evaluasi Model Klasifikasi

Proses evaluasi dilakukan dengan menghitung suatu ukuran tertentu terhadap himpunan data uji, yakni data yang tidak digunakan dalam proses pembuatan model

klasifikasi tersebut. *Confusion matrix* merupakan matrix yang berisi informasi mengenai klasifikasi aktual yang akan diprediksi oleh sistem klasifikasi (Kohavi & Provost, 1998). Sistem klasifikasi dibentuk dari pemetaan suatu baris data dan *output* suatu hasil prediksi kelas dari data tersebut. Pada suatu klasifikasi baris data dapat menghasilkan empat kemungkinan yang digunakan untuk menilai dan mengevaluasi proses klasifikasi. Apabila data positif dan tepat diprediksi positif maka disebut *true positive*, namun jika salah dan terprediksi negatif maka disebut *false negative*. Apabila data negatif dan tepat diprediksi negatif maka disebut *true negative*, namun jika salah dan terprediksi positif maka disebut *false positive* (Fawcett, 2006).

Tabel 3.1. *Confusion matrik*

		Actual	
	Kelas	**Positif**	**Negatif**
Prediksi	**Positif**	*True Positive (TP)*	*False Negative (FN)*
	Negatif	*False Positive (FP)*	*True Negative (TN)*

Adapun ukuran yang umumnya digunakan dalam penilaian dan evaluasi model klasifikasi sebagai berikut.

3. *Accuracy*

   Akurasi adalah jumlah proporsi prediksi yang benar. Akurasi digunakan sebagai tingkat ketepatan antara nilai actual dengan nilai prediksi. Adapun rumus penghitungan akurasi dapat dilihat pada persaman berikut.

$$Accuracy = \frac{TP+FN}{TP+FP+TN+FN} \quad (3.6)$$

4. *Precision*

   Precision adalah proporsi jumlah dokumen teks yang relevan terkenali diantara semua dokumen teks yang terpilih oleh sistem. *Precision* digunakan sebagai tingkat ketepatan antara informasi yang diminta dengan jawaban yang diberikan oleh sistem (Maning dkk, 2009). Rumus *precision* dapat dilihat pada persamaan berikut.

$$Precision = \frac{TP}{TP+FP} \quad (3.7)$$

5. *Recall*

   Recall adalah proporsi jumlah dokumen teks yang relevan terkenali diantara semua dokumen teks relevan yang ada pada koleksi. Nilai *recall* untuk kelas positif disebut juga dengan nilai *sensitivity*, sedangkan untuk kelas negatif disebut dengan nilai *specificity*. *Recall* digunakan sebagai ukuran keberhasialan sistem dalam menemukan kembali informasi (Maning dkk, 2009). Rumus *recall* dapat dilihat pada persamaan berikut.

$$Recall = \frac{TP}{TP+FN} \quad (3.8)$$

6. *F-Measure*

   F-measure merupakan rata-rata harmonis dari nilai *recall* dan nilai *precision* sehingga dapat memberikan peniaian kinerja yang lebih seimbang (Maning dkk, 2009). *F-measure* digunakan untuk mengukur kinerja sistem secara menyeluruh dalam pengklasifikasian. Rumus *F-measure* dapat dilihat pada persamaan berikut.

$$F - measure = \frac{2\,(recall \times precision)}{recall + precision} \quad (3.9)$$

3.7.2 *K-Fold Cross Validation*

Cross validation adalah teknik menilai validasi tingkat keakuratan sebuah model dari suatu dataset tertentu. Dataset terbagi menjadi data latih sebagai data yang digunakan untuk membangun model dan data uji sebagai data yang digunakan untuk memvalidasi model tersebut. Model klasifikasi digunakan dalam proses klasifikasi atau prediksi suatu data baru yang tidak termasuk dalam data pembangun model.

Menurut Refaeilzadeh dkk (2009), salah satu metode dari *cross validation* yang umumnya digunakan dalam perhitungan akurasi prediksi suatu sistem adalah *K-fold cross validation*. Proses dalam *K-fold cross validation* dilakukan dengan membagi dataset menjadi *K* segmen yang hampir sama ukuran proporsinya. Kemudian salah satu segmen *K* diambil sebagai data uji sedangkan *K-1* segmen

lainnya digunakan sebagai data latih dari pembentukan model baru. Proses pelatihan dan penilaian ini dilakukan sebanyak *K* kali iterasi. Nilai *K-fold cross validation* diperoleh dari rata-rata dari hasil iterasi yang dilakukan.

Jumlah *K* yang umumnya digunakan dalam *K-fold cross validation* yakni 5, 7, 10, dan 15. Adapun simulasi untuk *K-fold cross validation* untuk *K=5* sebagai berikut.

"Dataset = K1, K2, K3, K4, K5"

Tabel 3.2. *K-fold cross validation*

Iterasi	Data latih	Data uji
1	K2, K3, K4, K5	K1
2	K1, K3, K4, K5	K2
3	K1, K2, K4, K5	K3
4	K1, K2, K3, K5	K4
5	K1, K2, K3, K4	K5

Berdasarkan Tabel 3.2 dilakukan *K-fold cross validation* untuk *K=5* sehingga dilakukan iterasi sebanyak 5 kali. Iterasi dilakukan dengan mengambil data untuk segmen uji (1 segmen) dan segmen lainnya (4 segmen) pada tiap iterasinya (5 kali iterasi). Apabila terdapat 1000 data, untuk K=5 maka per segmen berjumlah masing-masing 200 data. Sehingga data latih berjumlah 800 data sedangkan untuk data uji berjumlah 200 data. Apabila K=10 dengan dataset sebanyak 1000 data maka data latih berjumlah 900 data sedangkan untuk data uji berjumlah 100 data.

3.8 *Maximum Entropy*

Menurut Nigam (1999), *Maximum Entropy* merupakan salah satu *machine learning* yang menggunakan proses pengestimasian probabilitas distribusi dalam pengklasifikasian data. Dalam metode *Maximum Entropy* dinyatakan bahwa untuk dataset yang tidak diketahui informasi mengenai distribusinya, maka data tersebut akan diasumsikan berditribusi seragam atau uniform. *Maximum Entropy* dapat digunakan untuk mengistimasi berbagai *natural language* taks seperti *languade modeling*, pelabelan *part of speech*, dan segmentasi pada teks lainnya.

Menurut Anggreini (2008), *Maximum Entropy* adalah metode klasifikasi berbasiskan probabilitas yang termasuk dalam kelas model eksponensial. Prinsip dari *Maximum Entropy* didasarkan pada distribusi *p(a|b)* yang akan memberikan nilai *entropy* maksimum. *Maximum Entropy* didefiniskan sebagai rata-rata nilai informasi yang maksimum untuk suatu himpunan kejadian X dengan distribusi nilai probabilitas yang seragam. Distribusi nilai probabilitas seragam yang dimaksud adalah distribusi yang menggunakan faktor ketidakpastian yang minimum atau dapat disebut sebagai distribusi yang memakai asumsi seminimal mungkin. Dengan menggunakan asumsi yang minimal, maksimal distribusi yang didapatkan merupakan distribusi yang paling mendekati kenyataan. Pencarian distribusi probabilitas yang paling memberikan nilai *entropy* yang maksimum dilakukan dengan tujuan mencari distribusi probabilitas terbaik.

3.8.1 Definisi *Entropy*

Entropy merupakan rata-rata dari himpunan informasi yang terdapat dalam suatu kumpulan kejadian $X = \{x_1, x_2, \ldots x_n\}$. Himpunan informasi yang terdapat pada suatu kejadian dapat dinyatakan dengan,

$$h(x) = log \frac{1}{p(x)} \tag{3.10}$$

dimana $h(x)$ adalah himpunan informasi dari suatu kejadian x yang dinyatakan dengan ukuran *bit*. Sedangkan $p(x)$ adalah probabilitas dari kemunculan kejadian x. Jumlah bit pada $h(x)$ adalah banyaknya bit yang perlukan untuk merepresentasikan himpunan informasi suatu kejadian. Semakin besar nilai $h(x)$, maka semakin besar pula informasi yang dimiliki oleh $h(x)$.

Pada klasifikasi teks nilai $h(x)$ dapat didefinisikan dengan *A* sebagai himpunan kelas klasifikasi dan *B* sebagai himpunan dokumen yang diklasifikasi. Nilai *entropy* dari himpunan dapat dinyatakan sebagai berikut.

$$H(p) = -\Sigma_{x \epsilon \varepsilon}\, p(x)\, log\, p(x) \tag{3.11}$$

dengan

$a \in A$ dan $b \in B$

$x = (a,b)$

$\varepsilon = \mathrm{A} \times B$

$p(x)$ yakni peluang kelas *a* terdapat pada dokumen *b*

Adapun hasil yang diharapkan dari metode algoritma *Maximum Entropy* yakni mendapatkan nilai (p) yang paling maksimal. Nilai *entropy* yang maksimal akan terpenuhi pada distribusi seragam sehingga mengakibatkan $(x) = \frac{1}{|\mathrm{X}|}$, dengan nilai $|X|$ merupakan kardinalitas dari X. Kardinalitas suatu himpunan merupakan nilai ukuran banyaknya elemen yang terdapat dalam suatu himpunan. Sehingga proses untuk mendapatkan nilai maksimal yang seragam dalam klasifikasi tidak semudah membagi nilai 1 dengan nilai kardinalitas dari X. Selain itu, pencarian distribusi probabilitas tentunya harus memenuhi batasan-batasan sesuai dengan jenis data yang diteliti.

3.8.2 Prinsip *Maximum Entropy*

Pada metode *Maximum Entropy* dinyatakan bahwa untuk memperoleh nilai nilai *entropy* maksimal maka seluruh distribusi akan diusahakan untuk *uniform*, apabila tidak terdapat informasi data yang lengkap. Pada klasifikasi teks dengan *Maksimum Entropy* dilakukan pengestimasian distribusi label dokumen. Dokumen sendiri akan direpresentasikan dengan fitur kemunculan kata. Perhitungan fitur didasari oleh penggunaan $f_{\mathrm{i}} \epsilon \{0,1\}$ dalam pencarian informasi kemunculan suatu fitur dalam suatu dokumen. Sehingga dasarnya metode algoritma *Maximum Entropy* digunakan untuk mencari distribusi probabilitas yang paling seragam

3.8.3 Algoritma Klasifikasi dengan *Maximum Entropy*

Adapun proses algoritma klasifikasi teks menggunakan metode *Maximum Entropy* dapat sebagai berikut

1) Mengidentifikasi kata-kata spesifik yang ada di dalam dokumen (kalimat).
2) Membentuk matriks yang berisi nilai kemunculan kata-kata spesifik tersebut dengan indeks berikut.

$$f_j(a,b) \begin{cases} 1; jika\ f\ j\ muncul\ di\ dokumen\ b\ pada\ kelas\ a \\ 0; jika\ f\ j\ tidak\ muncul\ di\ dokumen\ b\ pada\ kelas\ a \end{cases} \quad (3.12)$$

3) Membangun data latih untuk membuat model algoritma Maximum Entropy dengan menghitung nilai untuk setiap kelas αj.

$$a_j^{(0)} = 1 \tag{3.13}$$

$$a_j^{(n+1)} = a_j^{(n)} \left[\frac{E_{\tilde{p}} f_j}{E^{(n)} f_j}\right]^{\frac{1}{c}} \tag{3.14}$$

dimana,

$$E_{\tilde{p}} f_j = \sum_{x \epsilon \varepsilon} p(x) f_j(x) \tag{3.15}$$

$$E_{\tilde{p}}{}^{(n)} f_j = \sum_{x \epsilon \varepsilon} p^{(n)}(x) f_j(x) \tag{3.16}$$

$$p^{(n)}(x) = \pi \prod_{j=1}^{k} \left(a_j^{(n)}\right)^{f_j(x)} \tag{3.17}$$

$$\forall\, x \,\epsilon \sum_{j=1}^{k} f_j(x) = C$$

4) Menghitung *joint probability* $p(a,b)$ untuk perhitungan data uji.

$$a = \{positif, negatif\} \tag{3.19}$$

$$p^*(a,b) = \pi \prod_{j=1}^{k} a_j^{f_j(a,b)} \tag{3.20}$$

5) Menentukan kelas dari dokumen data uji dengan melihat nilai a^* tertinggi dari masing-masing kelas.

$$a^* = argmax\, p(a,b) \tag{3.21}$$

dengan $a \,\epsilon (positif, negatif)$

3.9 *Wordcloud*

Wordcloud merupakan salah satu metode untuk menampilkan kata-kata popular yang berkaitan dengan kata kunci internet dan data teks, khususnya pada analisis *test mining*. *Wordcloud* dapat digunakan dalam menyoroti trend ataupun istilah popular dikalangan pengguna. *Wordcloud* dibentuk berdasarkan frekuensi kemunculan kata, dimana kata yang paling sering muncul di dalam teks akan memiliki ukuran paling besar, begitu pula sebaliknya. Pendekatan menggunakan *wordcloud* dapat memberikan penjelasan terhadap pertanyaan penelitian dengan sangat cepat dan mudah serta dapat pula dilakukan analisis yang koprehensif (Graham dkk, 2015).

3.10 Asosiasi Teks

Asosiasi teks diperoleh dengan melakuan pendekatan pada perhitungan nilai korelasi. Pada umumnya, nilai korelasi digunakan dalam menyatakan hubungan dua atau lebih veriabel kuantitatif, namun pada asosiasi teks nilai korelasi dimaknai sebagai keeratan hubungan antar dua atau lebih variabel kualitatif (Ulwan, 2016). Korelasi bertujuan untuk menemukan tingkat hubungan antara variabel bebas (X) dan variabel bebas (Y), dalam ketentuan data memiliki syarat-syarat tertentu (Fadlisyah, 2014).

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{\{n\sum x_i{}^2 - (\sum x_i)^2\}\{n\sum y_i{}^2 - (\sum y_i)^2\}}} \quad (3.22)$$

dengan

r = nilai korelasi antara variabel x dan variabel y

n = banyaknya pasangan data x dan y

Σx_i = jumlah nilai pada variabel x i = 1, 2, 3, …, n

Σyi = jumlah nilai pada variabel y

Σx_i^2 = kuadrat dari total nilai variabel x

Σy_i^2 = kuadrat dari total nilai variabel y

Σx_i Σy_i = jumlah dari hasil perkalian antara nilai variabel x dan variabel y

Dalam perhitungan asosiasi teks, pertama-tama data teks ditransformasikan ke dalam *document term matrix* (dtm). Adapaun simulasi perhitungan dilakukan pada enam data berikut.

kata1
kata1 kata2
kata1 kata2 kata3
kata1 kata2 kata3 kata4
kata1 kata2 kata3 kata4 kata5
kata1 kata2 kata3 kata4 kata5 kata6

Kemudian ke 6 kata tersebut diubah dalam *document term matrix.*

Docs	kata1	kata2	kata3	kata4	kata5	kata6
1	1	0	0	0	0	0

2	1	1	0	0	0	0
3	1	1	1	0	0	0
4	1	1	1	1	0	0
5	1	1	1	1	1	0
6	1	1	1	1	1	1

Setelah diperoleh nilai *document term matrix*, selanjutnya dilakukan perhitungan nilai asosiasi. Nilai asosiasi diperoleh dengan menghitung rumus korelasi seperti pada simulasi kata 2 dan kata 4 berikut.

Docs	Kata2	Kata4	Kata2^2	Kata4^2	Kata2*4
1	0	0	0	0	0
2	1	0	1	0	0
3	1	0	1	1	1
4	1	1	1	1	1
5	1	1	1	1	1
6	1	1	1	1	1
Total	5	3	5	3	3

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{\{n\sum {x_i}^2 - (\sum x_i)^2\}\{n\sum {y_i}^2 - (\sum y_i)^2\}}}$$

$$r = \frac{(6 \times 3) - (5 \times 3)}{\sqrt{\{(6 \times 5) - 5^2\}\{(6 \times 3) - 3^2\}}}$$

$$r = \frac{3}{\sqrt{45}} = 0{,}447$$

Jadi, nilai korelasi kata 2 dan kata 4 sebesar 0,447. Hal ini menunjukkan bahwa besarnya asosiasi atau hubungan antara kata 2 dan kata 4 sebesar 0,447 atau 44,7%.

3.11 Diagram Fishbone

Diagram *fishbone* adalah sebuah alat visual yang berguna untuk proses identifikasi dan eksprorasi secara grafik dalam menggambarkan sebab akibat suatu permasalahan. Metode diagram *fishbone* atau dikenal juga dengan "Diagram Ishikawa" karena dikembangkan oleh Ishikawa pada sekitar tahun 1960-an. Nama fishbone diambil dari bentuk grafik yang menyerupai kerangka tulang ikan meliputi kepala, duri, dan sirip ikan, sedangkan sirip dan duri digambarkan sebagai penyebab permasalahannya. Konsep diagram *fishbone* secara umum adalah dengan

meletakkan permasalahan mendasar pada bagian kepala atau bagian paling kanan dari diagram kerangka tulang ikan.

Menurut Fritz (2016), pada dasarnya kegunaan diagram *fishbone* pengidentifikasian masalah dan penentuan penyebab dari masalah tersebut. Diagram *fishbone* dapat digunakan dalam berbagai level analisis permasalahan baik dalam individu, tim, ataupun organisasi. Adapun manfaat *fishbone* dalam analisis masalah, yaitu:

1. Memudahkan dalam mengilustrasikan gambaran singkat dalam tim/organisasi.
2. Lebih mempermudah untuk berfokus pada permasalahan utama. Diagram *fishbone* akan membantu tim/organisasi dalam menentukan masalah prioritas.
3. Diagram *fishbone* menghasilkan suatu solusi. Setelah dicari akar penyebab masalah, langkah dalam mengambil solusi akan lebih mudah dicapai.
4. Mempermudah untuk dilaksanakannya diskusi terutama pada pencarian sebab dan akibat suatu permasalahan.

Berdasarkan faktor-faktor penyebabnya diagram *fishbone* dapat dikategorisasikan berdasarkan konsep berikut. (Lovelock & Wright, 2005)

1. Konsep Fishbone Diagram 8P untuk Industri Marketing terbagi atas *Product, Price, Place, Promotion, People / personnel, Process, Physical Evidence,* dan *Performance.*
2. Konsep 5M untuk Industri Manufaktur terbagi atas Mesin (teknologi), Metode (proses), Material (termasuk *raw materials, consumables* dan informasi), *Man-power* (pekerja fisik) atau *Mind-power* (pekerja non-fisik), dan *Measurement* / pengukuran (inspeksi). Adapun terdapat penambahan (Bradley, 2016) yaitu *Milieu / Mother Nature* (faktor lingkungan), dan *Management / Money Power Maintenance.*
3. Konsep 4S untuk Industri Industri Jasa dan Pelayanan terdiri atas *Surroundings* (lingkungan), *Supplier*, *System* (sistem pelayanan konsumen), dan *Skill*.

BAB 4 METODOLOGI PENELITIAN

4.1 Populasi Penelitian

Populasi yang digunakan pada penelitian ini adalah semua data ulasan aplikasi berita *online* Detikcom yang terdapat dalam *website Google Play*. Sedangkan sampel yang digunakan dalam penelitian ini adalah semua data ulasan aplikasi berita *online* Detikcom yang dipublikasikan selama tahun 2019.

4.2 Jenis dan Sumber Data

Jenis data yang digunakan dalam penelitian ini adalah data kualitatif. Dimana data yang menjadi pokok penelitian ini berasal dari teks yang terdapat dalam ulasan yang diteliti. Sumber data yang digunakan dalam penelitian ini adalah data sekunder. Data tersebut diperoleh dengan menggunakan teknik *scraping* dengan bantuan aplikasi bawaan *Scraper* dari *Google Chrome* pada halaman situs aplikasi berita *online* Detikcom (https://play.google.com/store/apps/details?id=org. detikcom.rss).

4.3 Variabel Penelitian

Menurut Sugiyono (2011), variabel penelitian merupakan segala sesuatu yang ditentukan dan dipelajari oleh peneliti untuk memperoleh informasi yang digunakan dalam penarikan kesimpulan. Adapun variabel yang digunakan dalam penelitian ini adalah sebagai berikut.

1. *Date*, yakni tanggal dipublikasikan ulasan
2. *Rating*, yakni tingkat kepuasan pengguna terhadap aplikasi
3. Ulasan atau *review*, yakni pendapat pengguna terhadap aplikasi

4.4 Metode Analisis Data

Software yang digunakan dalam penelitian ini adalah *Microsoft Excel* dan *R 3.5.3.* Ada beberapa metode analisis data yang digunakan dalam penelitian ini, antara lain:

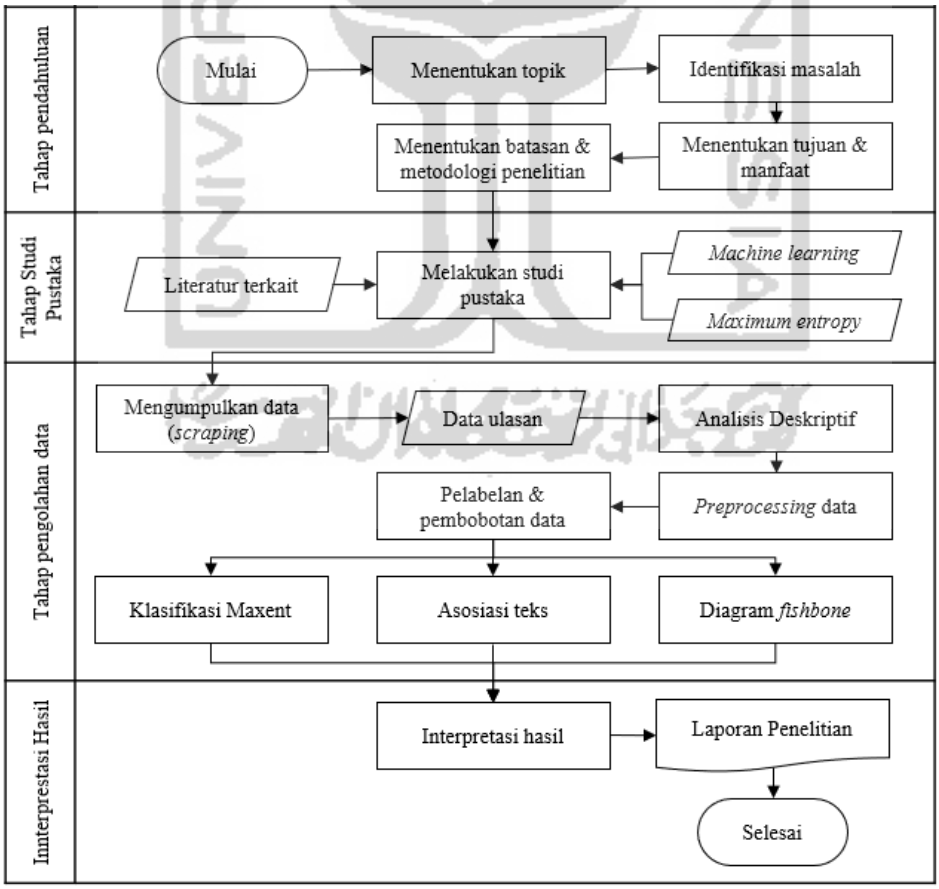
1. Analisis Deskriptif, digunakan untuk melihat gambaran umum mengenai aplikasi Detikcom selama tahun 2019.

2. Analisis sentimen dengan menggunakan metode *Maximum Entropy*, digunakan dalam proses pengklasifikasian *review* atau ulasan yang bersifat sentimen positif dan sentimen negatif.
3. Asosiasi teks, digunakan untuk mengidentifikasi dan membentuk pola kata yang dapat berasosiasi dengan kata lainnya untuk medapatkan informasi yang dianggap penting.
4. Diagram *Fishbone*, digunakan untuk mengidentifikasi faktor-faktor penyebab permasalahan yang didapatkan dari ulasan negatif sehingga dapat dilakukan rencan pemecahan masalah yang dihadapi.

4.5 Tahapan Penelitian

Tahapan yang dilakukan dalam penelitian ini dapat dipaparkan melalui diagram alir sebagai berikut.

Tabel 4.2. Diagram Alir Penelitian

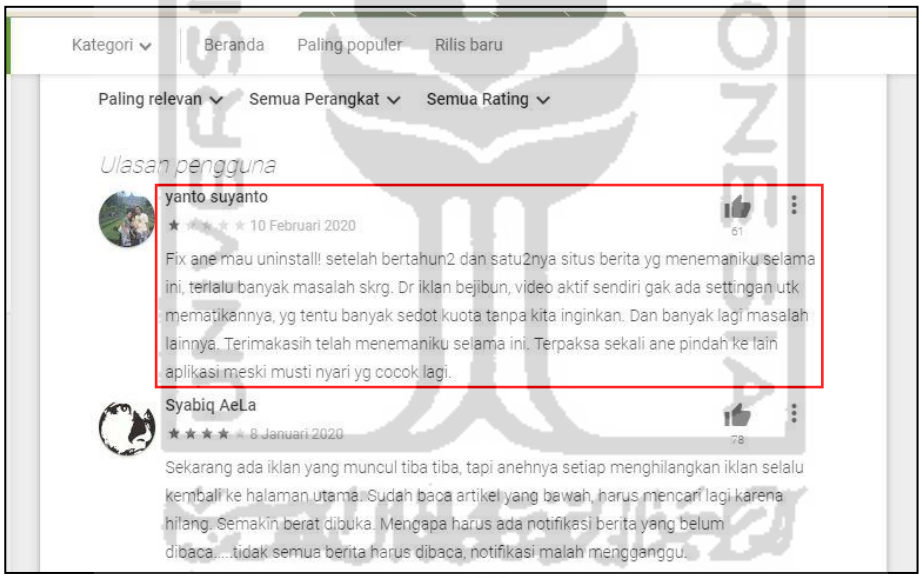


BAB 5 HASIL DAN PEMBAHASAN

5.1 Gambaran Umum

5.1.1 Pengumpulan Data

Data yang digunakan adalah ulasan pengguna aplikasi detikcom pada kolom komentar di *Google Play* pada halaman situs https://play.google.com/store/apps/details?id=org.detikcom.rss. Komponen-komponen yang diambil adalah nama pengguna, tanggal, *rating*, jumlah *like* dan ulasan pengguna, komponen tersebut dapat dilihat pada Gambar 5.1. Adapun data dikumpulkan menggunakan proses *scraping* dengan bantuan *tools Scrape Similar* pada *Google Chrome.*



Gambar 5.1. Laman kolom ulasan Detikcom

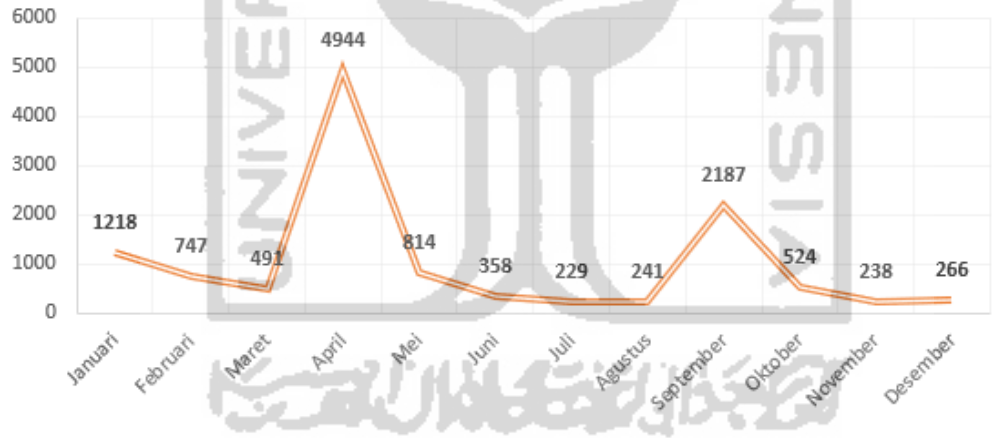
Hasil *scraping* yang diperoleh setelah proses pengolahan dapat dilihat pada Tabel 5.1 berikut. Adapun jumlah data berita diperoleh sebanyak 12.257 ulasan.

Tabel 5.1. Data ulasan Detikcom tahun 2019

	A	B	C	D	E	F	G	H
1	Rating	User	Tangga	Bulan	Tahun	Like	Review	
2	1	angga andi ardiansyah	1	1	2019	10	sekarang detikcom terlalu banyak iklan, dan saya sangat tidak suka sekali	
3	1	MinasaUpa alif	1	1	2019	19	Lihat histori editMakin hari makin ndag jelas, kemaren2 iklan makin banya	
4	1	Fitri Wulandari	2	1	2019	0	sering log out	
5	1	Tsamara Nayyara	2	1	2019	0	Berita yg dimuat sudah tidak berimbang. Saatnya uninstall. Bye detik	
6	1	jason lim	2	1	2019	1	iklan yg tiba2 muncul menjengkelkan,lgi asyik baca jadi gk asyik...terus yg	
7	1	shifuu ady	2	1	2019	1	popup iklan MENGGANGGU !!!!	
8	1	Heri Yanto	3	1	2019	1	makin lama detikcom makin kaya tai... lelet nya super lelet... gimana oran	
9	1	Sesairo San	3	1	2019	0	Beritanya tidak netral	
10	1	Bung Fazha	4	1	2019	1	terima kasih untuk detik sport, klo berita politik mah ampas, jilat bool kov	
11	1	emil sugiyanto	5	1	2019	12	Lihat histori editAPLIKASI DETIK CEBONG.. LU KLO MAU BUAT BERITA DAN	
12	1	Yogi Rafiqi	5	1	2019	0	sering keluar sendiri, padahal tdi ny kagak	
13	1	Adhy Stockphoto	5	1	2019	3	Uninstal aja lah, semenjak iklan" berbentuk video & full layar ga bermutu	
14	1	GR Channel	8	1	2019	0	detikInet isinya byk berita GRAB aj, kampret. nggak mutu.	
15	1	Pengguna Google	9	1	2019	2	11 tahun jd pembaca detikcom, makin kesini makin jd jongos penguasa, s	

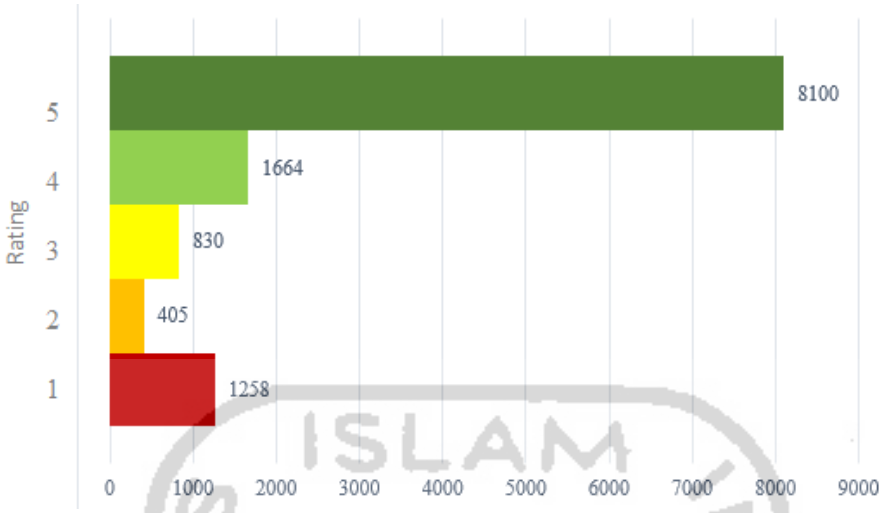
5.1.2 Analisis Deskiptif

Analisis statistik deskriptif bertujuan untuk melihat gambaran umum mengenai informasi dari aplikasi Detikcom berdasarkan jumlah ulasan dan *rating* yang diberikan pengguna. Selama tahun 2019 diketahui jumlah ulasan pengguna sebanyak 12.257 ulasan.



Gambar 5.2. Grafik jumlah ulasan tahun 2019

Berdasarkan grafik pada Gambar 5.2 dilihat jumlah ulasan dari para pengguna pada setiap bulannya tahun 2019. Jumlah ulasan terbanyak terdapat pada bulan April dengan 4944 ulasan. Jumlah ulasan terkecil terdapat pada bulan Juli dengan 229 ulasan. Pada bulan April dan September terjadi kenaikan jumlah ulasan yang tinggi dibandingkan bulan-bulan lainnya. Hal ini disebabkan tingginya kebutuhan masyarakat terhadap berita terutama informasi seputar pemilihan umum (pemilu) Indonesia. Bulan April merupakan pelaksanaan pemilu dan September merupakan pengumuman hasil pemilu.

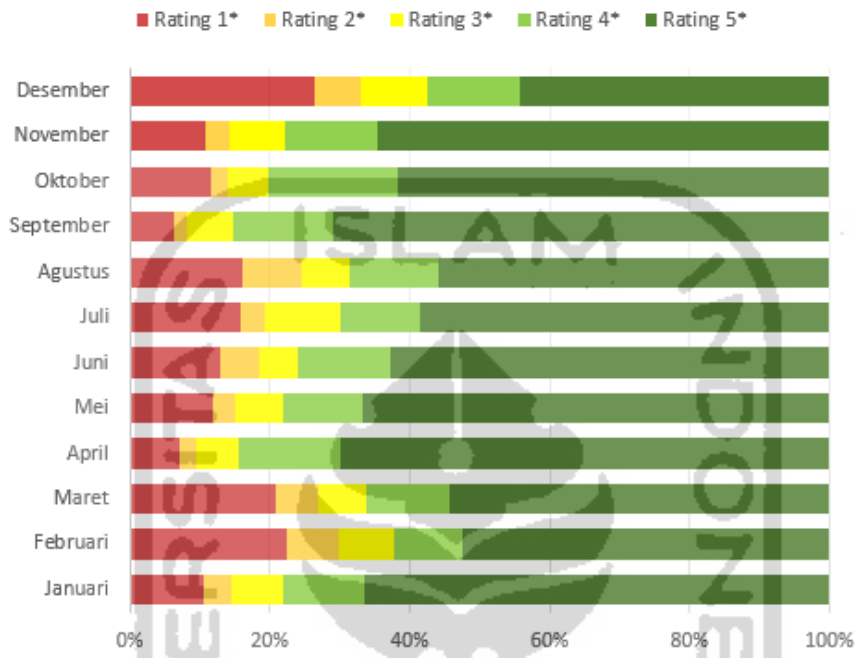


Gambar 5.3. Grafik *rating* pengguna tahun 2019

Berdasarkan Gambar 5.3 dapat dilihat jumlah ulasan dari *rating* pengguna pada tahun 2019. Rating pada situs *Google Play* mempunyai sekala 1 sampai 5 dengan skor 1 untuk kategori "Sangat Tidak Suka", skor 2 untuk kategori "Sangat Suka", skor 3 untuk kategori "Cukup", skor 4 untuk kategori "Suka", dan skor 5 untuk kategori "Sangat Suka". Dari penilaian pengguna tersebut diketahui pengguna Detikcom mempunyai penilain yang baik terhadap aplikasi. Hal ini terlihat dari mayoritas pengguna memberikan *rating* 5 atau sangat suka terhadap aplikasi dengan jumlah 8100 ulasan dan diikuti oleh *rating* 4 atau suka terhadap aplikasi dengan jumlah 1664 ulasan. Adapun pengguna memberikan *rating* 3 atau cukup terhadap aplikasi sebanyak 830 ulasan, untuk *rating* 2 atau tidak suka terhadap aplikasi sebanyak 405 ulasan, dan untuk *rating* 1 atau sangat tidak suka terhadap aplikasi sebanyak 1258 ulasan.

Berdasarkan Gambar 5.4 dapat dilihat proporsi jumlah *rating* pada setiap bulannya selama tahun 2019. Jumlah proporsi *rating* 1 tertinggi terdapat pada bulan Desember sebesar 26,3% sedangkan yang terendah pada bulan September sebesar 6,2%. Jumlah proporsi *rating* 2 tertinggi terdapat pada bulan Agustus sebesar 8,3% sedangkan yang terendah pada bulan September sebesar 1,9%. Jumlah proporsi *rating* 3 tertinggi terdapat pada bulan Juli sebesar 10,9% sedangkan yang terendah pada bulan Juni sebesar 5,6%. Jumlah proporsi *rating* 4 tertinggi terdapat pada bulan Oktober sebesar 18,5% sedangkan yang terendah

pada bulan Februari sebesar 9,8%. Jumlah proporsi *rating* 5 tertinggi terdapat pada bulan September sebesar 70,8% sedangkan yang terendah pada bulan Desember sebesar 44,4%.



Gambar 5.4. Grafik proporsi jumlah *rating* pengguna tahun 2019

Secara umum pada bulan September, April, dan Januari memiliki proporsi tingkat kepuasan tertinggi dimana proporsi *rating* 5 dan *rating* 4 lebih tinggi dibandingkan bulan-bulan lainnya. Pada bulan Januari, peningkatan disebabkan keoptimisan pengguna terhadap perkembangan Detikcom kedepannya di awal tahun. Sedangkan pada bulan April dan September, peningkatan lebih disebabkan kebutuhan pengguna akan informasi seputar pemilu Indonesia terutama oleh pengguna-pengguna baru. Pada bulan Desember dan Februari memiliki proporsi tingkat kepuasan terendah dimana proporsi *rating* 1 dan *rating* 2 lebih tinggi dibandingkan bulan-bulan lainnya. Pada bulan Desember, penurunan disebabkan tidak terpenuhinya permintaan pengguna terhadap suatu fitur atau perubahan Detikcom selama satu tahun terakhir. Sedangkan bulan Februari, penurunan lebih disebabkan Detikcom gagal memenuhi keoptimisan pengguna pada bulan sebelumnya.

5.2 *Text Mining*

5.2.1 *Preprocesing Data*

Sebelum masuk ke dalam proses klasifikasi, data ulasan yang masih belum terstruktur dengan baik dan belum seragam, perlu dilakukan tahap *preprocessing* atau prapemprosesan. Tahap *preprocessing* bertujuan untuk menyeragamkan dan menstrukturkan data yakni dengan mengurangi volume kosakata terutama pada karakter-karakter selain huruf. Karakter pada data ulasan yang akan dihilangkan berupa angka, tanda baca, emoticon, serta kata-kata lainnya yang dianggap kurang bermanfaat. Adapun ulasan yang masih berbahasa Inggris akan diseragamkan ke dalam Bahasa Indonesia. Pada tahapan *prepocessing*, proses pembersihan data dilakukan dengan menggunakan *text mining*. Adapun tahap-tahap yang akan dilakukan yakni *translating, spelling normalization, case folding, tokenizing,* dan *filtering* yang akan dipaparkan sebagai berikut.

1. *Translating*

   Translating adalah proses penyeragaman bahasa, karena analisis dilakukan dalam Bahasa Indonesia maka semua ulasan berbahasa asing (Inggris) akan terjemahkan ke dalam Bahasa Indonesia. Adapun proses *translating* dapat dilihat pada Tabel 5.2 berikut.

Tabel 5.2. *Translating*

Input	*Output*
I need an easy & free app for updated news, and detik provide it.	Saya membutuhkan aplikasi yang mudah & gratis untuk berita terbaru, dan detik menyediakannya.

2. *Spelling Normalization*

   Spelling Normalization adalah proses penyeragaman ejaan kata. Perbaikan dilakukan pada kata-kata singkatan, kata-kata yang masih salah ejaan, atau kata-kata sinonim dengan makna serupa. Misalnya pada kata tidak, tdk, tak, ga, nggak tdak, dan seterusnya yang memiliki makna sama akan diseragamkan menjadi satu kata.

Tabel 5.3. *Spelling normalization*

Input	*Output*
Kok sekarang **gak** ada berita olahraga ya. **Cm yg** berita **yg** ada -_-	Kok sekarang **tidak** ada berita olahraga ya. **Cuma yang** berita **yang** ada -_-

3. *Case folding*

Case folding adalah proses penyeragaman bentuk huruf. Karakter lain selain huruf seperti angka, tanda baca, dan emoticon akan dianggap sebagai *delimiter* sehingga karakter tersebut akan dihapus. Selain itu huruf juga diseragaman menjadi huruf non kapital supaya kata yang dituliskan menggunakan huruf awal kapital dan huruf non-kapital tidak terjadi perbedaan arti.

Tabel 5.4. *Case folding*

Input	*Output*
Kok sekarang tidak ada berita olahraga ya. **C**uma yang berita yang ada -_-	kok sekarang tidak ada berita olahraga ya cuma yang berita yang ada

4. *Tokenizing*

Tokenizing adalah proses pemisahan teks dokumen menjadi kata per kata yang tidak saling berpengaruh (*independent*). Potongan kata tersebut dinamakan token yakni sebuah entitas yang mempunyai nilai dalam penyusunan matriks dokumen. *Tokenizing* berguna untuk mempermudah perhitungan frekuensi kemunculan kata dalam dokumen.

Tabel 5.5. *Tokenizing*

Input	*Output*		
kok sekarang tidak ada berita olahraga ya cuma yang berita yang ada	kok sekarang tidak ada	berita olahraga ya cuma	yang berita yang ada

5. *Filtering*

   Filtering adalah proses penyaringan atau pemilihan kata dalam dokumen. Tahap ini bertujuan untuk mengurangi dimensi kata di dalam *corpus* dengan menggunakan *stopwords*. *Stopwords* terdiri dari kata-kata yang kurang informatif atau kurang berpengaruh secara keseluruhan yang sering kali muncul dalam dokumen. Kata-kata yang masuk ke dalam *stopwords* umumnya terdiri dari kata ganti orang, kata penghubung, kata seruan, kata pertanyaan, dan kata lainnya yang tidak mempunyai makna penting dalam penentuan bahasan dari dokumen. Adapun contoh kata-kata yang masuk dalam penyaringan sebagai berikut.

   a) Kata penghubung, contoh: atau. dan, lalu, kemudian, serta dsb.
   b) Kata preposisi, contoh: di, ke, pada, dsb.
   c) Kata-kata lainnya yang tidak diinginkan (perintah *remove words*).

Tabel 5.6. *Filtering*

Input			*Output*		
kok sekarang tidak ada	berita olahraga ya cuma	yang berita yang ada	sekarang tidak	berita olahraga	berita

5.2.2 Pelabelan Kelas Sentimen

Proses klasifikasi menggunakan *machine learning* masih termasuk ke dalam klasifikasi terawasi (*supervied classifications*), sehingga dibutuhkan pelabelan kelas sentimen terhadap data ulasan. Pada umumnya, pelabelan klasifikasi yang dilakukan pada analisis sentimen dapat dibagi menjadi tiga kelas sentimen yakni sentiment positif, sentimen netral, dan sentimen negatif.

Berdasarkan *rating* yang diberikan pengguna dapat diketahui gambaran umum dari penilaian pengguna terhadap aplikasi Detikcom. Maka dari itu, ulasan dengan *rating* 1 untuk kategori “Sangat Tidak Suka” dan *rating* 2 untuk kategori “Tidak Suka” dapat dilabelkan ke dalam kelas negatif, sedangkan ulasan dengan

rating 3 untuk kategori "Cukup" dapat dilabelkan ke dalam kelas netral, dan ulasan dengan *rating* 4 untuk kategori "Suka" dan *rating* 5 untuk "Sangat Suka" dapat dilabelkan ke dalam positif. Adapun hasil pelabelan kelas sentimen diperoleh jumlah data seperti berikut.

Tabel 5.7. Perbandingan jumlah data pada kelas sentimen

Kelas	Jumlah
Positif	9764
Netral	830
Negatif	1663

Berdasarkan Tabel 5.7, hasil pelabelan kelas sentimen bahwa ulasan tertinggi terdapat pada kelas sentimen positif dengan jumlah 9764 ulasan, diikuti dengan kelas sentimen negatif sebanyak 1663 ulasan dan kelas sentimen netral sebanyak 830 ulasan. Kelas sentimen positif terdiri dari ulasan pengguna yang merasa puas dengan Detikcom. Sedangkan, kelas sentimen negatif terdiri dari ulasan pengguna yang merasa tidak puas dengan Detikcom. Pada kelas sentimen netral, umumnya pengguna yang merasa cukup dengan Detikcom, tidak banyak berkomentar, ataupun memiliki opini yang bercampur antara ulasan positif dan negatif. Hal ini mengakibatkan kelas sentiment netral dianggap kurang dapat memberikan informasi yang lebih spesifik baik dalam masukaan maupun keluhan.

Pada penelitian ini dilakukan pereduksian kelas dengan membagi ulasan sentimen netral ke dalam sentimen positif dan sentimen negatif. Proses pelabelan dapat dilakukan dengan menghitung skor sentiment menggunakan bantuan kamus lexicon. Pembobotan kata dilakukan dengan menghitung frekuensi kemunculan kata pada sebuah dokumen teks. Menurut Yates & Neto (1999), semakin tinggi frekuensi sebuah kata muncul pada dokumen teks, maka semakin besar pula bobot kata tersebut dan dapat dianggap kata tersebut merupakan representasi dari dokumen teks tersebut. Adapun tahapan melakukan pelabelan menggunakan *software R* sebagai berikut.

Tabel 5.8. Sintaks *R* proses pelabelan

Sintaks *R*	**Fungsi**

`library(tm)`
`setwd("E://SK/TA/DATA")`
`kalimat2<-read.csv ("DETIK_netral.csv",`
`  header=TRUE)`
`positif <- scan("kata-pos.txt",`
`  what="character",comment.char=";")`
`negatif <- scan("kata-neg.txt",`
`  what="character",comment.char=";")`
`kata.positif = c(positif, "membaik")`
`kata.negatif = c(negatif, "kemunduran")`
	1. *Running packages* tm
	2. Membuka file csv data ulasan
	3. Membuka file txt untuk daftar kata positif dan kata negatif

`score.sentiment = function(kalimat2,`
`  kata.positif, kata.negatif,`
`  .progress='none')`
`{`
`require(plyr)`
`require(stringr)`
`scores = laply(kalimat2, function(kalimat,`
`  kata.positif, kata.negatif) {`
`kalimat = gsub('[[:punct:]]', '', kalimat)`
`kalimat = gsub('[[:cntrl:]]', '', kalimat)`
`kalimat = gsub('\\d+', '', kalimat)`
`kalimat = tolower(kalimat)`
 	
`list.kata = str_split(kalimat, '\\s+')`
`kata2 = unlist(list.kata)`
 	
`positif.matches = match(kata2,`
`  kata.positif)`
`negatif.matches = match(kata2,`
`  kata.negatif)`
`positif.matches = !is.na(positif.matches)`
`negatif.matches = !is.na(negatif.matches)`
 	
`score = sum(positif.matches) -`
`  (sum(negatif.matches))`
`return(score)`
`}, kata.positif, kata.negatif,`
`  .progress=.progress )`
`scores.df = data.frame(score=scores,`
`  text=kalimat2)`
`return(scores.df)`
`}`
	4. Membuat function untuk proses skoring dengan tahapan:
	a) *Running packages* plyr dan stringr
	b) Membuat array dari daftar inisial
	c) Melakukan *case folding* dan menghapus *noise*
	d) Melakukan *tokoneizing* dan menyederhanakan daftar kata
	e) Mengidentifikasi *term* dari setiap kata positif dan kata negatif
	f) Membuat logika dari kata positif dan kata negatif
	g) Menghitung jumlah skor sentimen
	h) Menyimpan tabel dari data skor dan data ulasan

`hasil = score.sentiment(kalimat2$text,`
`  kata.positif, kata.negatif)`
`hasil$klasifikasi<- ifelse(hasil$score==0,`
`  "Netral", ifelse(hasil$score<0,`
`  "Negatif", "Positif"))`
	5. Memanggil *function* hasil skoring
	6. Mengkonversi nilai skor ke dalam kelas positif dan kelas negatif

Sintaks *R*	Fungsi

`hasil$klasifikasi`
`data <- hasil[c(3,1,2)]`
`View(data)`
`write.csv(data, file = "DETIKLabel.csv")`
	7. Menyimpan file hasil pelabelan data dalam format csv

Berdasarkan Tabel 5.8 diperoleh nilai skor dan kelas sentimen baru menggunakan *software R.* Adapun proses perhitungan nilai skor sentimen secara manual dilakukan dengan rumus perhitungan skor sentimen sebagai berikut.

$$Skor = (Jumlah\ kata\ positif) - (Jumlah\ kata\ negatif) \qquad (5.1)$$

Simulasi perhitungan skor pada ulasan "cepat memberitakan akurat jenis berita lengkap buka susah" yang dapat dilihat pada Tabel 5.9. Pada uasan tersebtu terdapat tiga kata positif dan 1 kata negatif yang terdeteksi pada kamus lexicon, yakni "cepat", "akurat", "lengkap" sebagai kata positif, dan "susah" sebagai kata negatif.

Tabel 5.9. Pembobotan kata positif dan negatif

Ulasan	Positif	Negatif
cepat memberitakan **akurat** jenis berita **lengkap** buka **susah**	3	1

Sehingga diperoleh:

Skor = (Jumlah kata positif) – (Jumlah kata negatif)

Skor = 3 – 1 = 2

Skor akhir yang diperoleh dari simulasi perhitungan bernilai > 0, maka ulasan tersebut masuk dalam kelas sentimen positif. Contoh hasil pelabelan kelas baru untuk data sentimen netral seperti pada Tabel 5.10 berikut.

Tabel 5.10. Contoh hasil pelabelan baru data ulasan netral

Kelas Baru	Skor	Ulasan
Positif	2	cepat memberitakan akurat jenis berita lengkap buka susah
Negatif	-1	iklan mengganggu
Positif	1	berita cepat
Negatif	-1	terkadang respon lambat

Selain menggunakan kamus *lexicon*, pengkategorian secara manual juga dilakukan untuk sentimen netral pada kasus tertentu. Apabila ulasan kelas sentiment netral memperoleh skor 0 sulit untuk dianalisis maka akan dimasukan ke dalam pelabelan baru secara manual sesuai dengan prespektif peneliti. Umumnya ulasan yang sulit diidentifikasi akan dimasukan ke dalam kelas sentimen negatif dengan pertimbangan informasi negatif akan lebih dieksplorasi pada analisis selanjutnya. Adapun hasil pelabelan baru yang diperoleh sebagai berikut.

Tabel 5.11. Hasil reduksi pelabelan kelas sentimen

Kelas	Jumlah
Positif	10199
Negatif	1997

Berdasarkan Tabel 5.11 diketahui jumlah hasil reduksi untuk kelas pelabelan sentimen baru menunjukan bahwa ulasan tertinggi terdapat pada kelas sentimen positif dengan jumlah 10199 dengan penambahan 435 ulasan. Sedangkan kelas sentimen positif berjumlah 1997 diperoleh penambahan 334 ulasan. Sebanyak 61 ulasan pada kelas sentimen netral tidak memenuhi syarat (*blank text*) sehingga digunakan dalam analisis.

5.3 *Machine Learning*

5.3.1 Pembuatan Data Latih dan Data Uji

Data latih digunakan oleh algoritma klasifikasi untuk membentuk sebuah model *classifer*. Model ini merupakan representasi pengetahuann yang akan digunakan untuk prediksi kelas data baru yang belum pernah ada, semakin besar data latih yang digunakan, maka akan semakin baik *machine* dalam memahami pola data. Data uji digunakan untuk mengukur sejauh mana classifer berhasil melakukan klasifikasi dengan benar. Berdasarkan Paretto Principle, rasio yang umum digunakan dalam pembagian data latih dan data uji adalah 80%:20%. Adapun pembagian jumlah data latih dan data uji dapat dilihat sebagai berikut.

Tabel 5.12. Pembagian jumlah data latih dan data uji

Kelas	**Jumlah**	**Data Latih (80%)**	**Data Uji (20%)**
Positif	10199	8159,2 ≈ 8159	2039,8 ≈ 2040
Negatif	1997	1597,6 ≈ 1598	399,2 ≈ 399
Total		9757	2439

Berdasarkan Tabel 5.12 diketahui perbandingan jumalah data latih dan data uji. Dengan proporsi perbandingan sebesar 80%:20% dari keseluruhan 12.196 ulasan diperoleh data uji sebanyak 9.757 ulasan dan data uji sebanyak 2.439 ulasan.

5.3.2 Klasifikasi dengan Metode *Machine Learning*

Proses klasifikasi dilakukan dengan menggunakan data latih dan data uji dalam proses pembuatan *machine learning*. Model *machine learning* sendiri dibuat dengan melakukan pelatihan pada data latih yang selanjutnya dapat diujikan menggunakan data uji. Pengujian model dilakukan untuk mengetahui tingkat keakurasian model dan sejauh mana model dapat melakukan klasifikasi dengan benar. Dalam menentukan *meachine learning* terdapat beberapa metode algoritma dapat digunakan. Percobaan dilakukan untuk menentukan metode yang memiliki nilai akurasi tertinggi. Metode algoritma yang digunakan dalam

penelitian ini adalah *Maximum Entropy* (Maxent). Adapun tahapan klasifikasi menggunakan program *R* sebagai berikut.

Tabel 5.13. Sintaks R klasifikasi *machine learning*

Sintaks R	**Fungsi**

`setwd("E://SK/TA/DATA")`
`positif = readLines("PL.csv")`
`negatif = readLines("NL.csv")`
`positiftes = readLines("PT.csv")`
`negatiftes = readLines("NT.csv")`
`ulasan = c(positif, negatif)`
`ulasan_test= c(positiftes, negatiftes)`
`ulasan_all = c(ulasan, ulasan_test)`
`sentiment = c(rep("positif",`
`  length(positif)), rep("negatif",`
`  length(negatif)))`
`length(sentiment)`
`sentiment_test = c(rep("positif",`
`  length(positiftes)), rep("negatif",`
`  length(negatiftes)))`
`sentiment_all = as.factor(c(sentiment,`
`  sentiment_test))`
`all= data.frame(ulasan_all,sentiment_all)`
	8. Membuka file csv untuk data latih & data uji
	9. Mendefinisikan masing-masing data latih & data uji
	10. Menggabungkan data latih & data uji yang telah didefinisikan
	11. Mendefinisikan label kelas untuk data latih & data uji
	12. Menggabungkan data latih & data uji dengan label kelas

`library(RTextTools)`
`library(e1071)`
`mat = create_matrix(ulasan_all, language =`
`  "indonesian", removeStopwords = FALSE,`
`  removeNumbers = TRUE, stemWords =`
`  FALSE, tm::weightTfIdf)`
`mat = as.matrix(mat)`
 	
`ln = length(sentiment)`
`ln1 = ln+1`
`la = length(sentiment_all)`
`container = create_container(mat,`
`  sentiment_all, trainSize=1:ln,`
`  testSize=ln1:la, virgin=FALSE)`
	13. *Running packages* RTextTools dan e1071
	14. Membuat *term weighting - document term matrix* (tf-dtm)
	15. Mengubah data ulasan ke dalam bentuk matriks
	16. Menghitung jumlah data latih & data uji
	17. Membuat container untuk masing-masing data latih & data uji

`# Maximum Entropy`
`model = train_model(container, 'MAXENT')`
`results = classify_model(container, model)`
`summary(model)`
 	
`table(as.character(sentiment_all[ln1:la]),`
`  as.character(results[,"MAXENTROPY_LABEL`
`  "]))`
`recall_accuracy(sentiment_all[ln1:la],`
`  results[,"MAXENTROPY_LABEL"])`
`create_precisionRecallSummary(container,`
`  results)`
	5. Membuat model *Maximum Entropy*
	6. Menentukan klasifikasi data uji
	7. Menghitung tabel *confusion matrix*
	8. Menghitung nilai akurasi
	9. Menghitung nilai presisi dan *recall*

Berdasarkan Tabel 5.13 diperoleh hasil klasifikasi dengan menggunakan metode algoritma *Maximum Entropy*. Adapun nilai akurasi yang diperoleh dinilai cukup tinggi yakni sebesar 0,916018 atau 91,6%.

5.3.3 Evaluasi *Machine Learning*

Setelah diperoleh metode klasifikasi, diperlukan evaluasi dari sistem keakuratan dari model yang dibangun. Pembangunan model terdiri dari pembentukan data latih data uji. *Cross validation* merupakan salah satu teknik untuk menilai/memvalidasi keakuratan sebuah model yang dibangun berdasarkan dataset tertentu. Model yang dibuat bertujuan untuk melakukan prediksi atau klasifikasi terhadap suatu data baru yang belum ada di dalam dataset.

Percobaan dilakukan sebanyak 5 kali iterasi pada dataset acak sebagai *cross validation* untuk mendapatkan dataset dengan nilai akurasi prediksi terbaik. Adapun hasil masing-masing percobaan *machine learning* menggunakan metode *Maximum Entropy* sebagai berikut.

Tabel 5.14. Perbandingan nilai akurasi *Maximum Entropy*

Machine learning	**Akurasi**
Iterasi 1	91,60%
Iterasi 2	95,66%
Iterasi 3	91,97%
Iterasi 4	90,54%
Iterasi 5	95,69%

Berdasarkan Tabel 5.14 diketahui nilai akurasi tertinggi diperoleh dari dataset pada *machine learning* iterasi 5 sebesar 95,69%. Hasil perhitungan tingkat akurasi diperoleh dari jumlah data uji yang terklasifikasi dengan benar dibandingkan dengan total semua data yang di uji. Dari kelima iterasi tersebut diperoleh hasil rata-rata akurasi sebesar 93,09% dengan menggunakan *Maximum Entropy*.

Salah satu metode dari *cross validation* lainnya yang digunakan adalah *K-fold cross validation.* Dalam *K-fold cross validation* data dibagi menjadi k bagian

yang memiliki ratio yang sama. Salah satu dari bagian dipilih sebagai data uji, sedangkan sisanya sebagai data latih. Selama proses ini diulang sebanyak k kali. Adapun hasil perbandingan nilai akurasi K-*fold cross validation* untuk K=5 sebagai berikut.

Tabel 5.15. Perbandingan nilai akurasi *5-fold cross validation*

Machine learning	**Akurasi**
Iterasi K1	90,25%
Iterasi K2	90,33%
Iterasi K3	89,92%
Iterasi K4	90,41%
Iterasi K5	89,35%

Berdasarkan Tabel 5.15 diketahui nilai akurasi tertinggi diperoleh dari dataset pada *machine learning* iterasi K2 sebesar 90,41%. Hasil pengujian performa sistem dalam melakukan klasifikasi dengan menggunakan *5-fold cross validation* diperoleh hasil rata-rata akurasi sebesar 90,05% untuk metode *Maximum Entropy*.

Selain sistem keakuratan dari model diperlukan juga evaluasi dari data uji tersebut. Penilaian dalam proses evaluasi menggunakan metode *confusion matrix* untuk mencari nilali akurasi, nilai presisi, dan nilai *recall*. *Confusion matrix* merupakan salah satu alat dalam metode evaluasi yang digunakan pada *machine learning* yang biasanya membuat dua kategori atau lebih (Manning dkk, 2009). Setiap unsur matriks menunjukkan jumlah contoh data uji untuk kelas sebenarnya dalam bentuk baris sedangkan kelas yang diprediksi dalam bentuk kolom.

Tabel 5.16. Hasil klasifikasi data uji

Aktual	**Prediksi**		**Jumlah**
	Positif	Negatif	
Positif	2007	33	2040
Negatif	72	327	399
Jumlah	2079	360	2439

Banyaknya data observasi berkategori positif yang mampu diprediksi positif (diprediksi dengan tepat) oleh *machine learning* disebut dengan *true positif* (TP) sebesar 2007 ulasan. Banyaknya data observasi berkategori negatif yang mampu diprediksi negatif (diprediksi dengan tepat) oleh *machine learning* disebut dengan *true negative* (TN) sebesar 327 ulasan. Banyaknya data observasi yang berkategori positif akan tetapi terdapat kesalahan prediksi disebut dengan *false positif* (FP) sebesar 33 ulasan. Banyaknya data observasi yang berkategori negatif akan tetapi terdapat kesalahan prediksi disebut dengan *false negatif* (FN) sebesar 72 ulasan.

Adapun proses perhitungan mengevaluasi dilakukan dengan menggunakan rumus berikut.

1. Akurasi adalah jumlah proporsi prediksi yang benar. Akurasi digunakan sebagai tingkat ketepatan antara nilai actual dengan nilai prediksi.

$$Akurasi = \frac{TP + FN}{TP + FP + TN + FN} = \frac{2007 + 327}{2007 + 33 + 327 + 72}$$

$$= 0{,}9569496 \text{ atau } 95{,}69\%$$

2. *Precision* adalah proporsi jumlah dokumen teks yang relevan terkendali diantara semua dokumen teks yang terpilih oleh sistem. *Precision* digunakan sebagai tingkat ketepatan antara informasi yang diminta dengan jawaban yang diberikan oleh sistem.

$$Precision_{Positif} = \frac{TP}{TP + FP} = \frac{2007}{2007 + 72} = 0{,}96537 \text{ atau } 96{,}54\%$$

$$Precision_{Negatif} = \frac{TN}{TN + FN} = \frac{327}{327 + 33} = 0{,}90833 \text{ atau } 90{,}83\%$$

3. *Recall* adalah proporsi jumlah dokumen teks yang relevan terkendali diantara semua dokumen teks relevan yang ada pada koleksi. *Recall* digunakan sebagai ukuran keberhasilan sistem dalam menemukan kembali informasi.

$$Recall_{Positif} = \frac{TP}{TP + FN} = \frac{2007}{2007 + 33} = 0{,}98382 \text{ atau } 98{,}39\%$$

$$Recall_{Negatif} = \frac{TN}{TN + FP} = \frac{327}{327 + 72} = 0{,}81955 \text{ atau } 81{,}96\%$$

4. *F-measure* merupakan rata-rata harmonis dari nilai *recall* dan nilai *precision* untuk memperoleh peniaian kinerja yang lebih seimbang. *F-measure* digunakan sebagai pengukur dari kinerja sistem dalam pengklasifikasian.

$$F_{measure\ Positif} = \frac{2\ (recall \times precision)}{recall + precision} = \frac{2\ (0{,}98382 \times 0{,}96537)}{0{,}98382 + 0{,}96537}$$
$$= 0{,}97451 \text{ atau } 97{,}45\%$$
$$F_{measure\ Negatif} = \frac{2\ (recall \times precision)}{recall + precision} = \frac{2\ (0{,}81955 \times 0{,}90833)}{0{,}81955 + 0{,}90833}$$
$$= 0{,}86166 \text{ atau } 86{,}17\%$$

Tabel 5.17. Hasil *confusion matrix*

Aktual	**Prediksi**		***Recall***
	Positif	Negarif	
Positif	2007	33	98,38%
Negatif	72	327	81,95%
Precisition	96,54%	90,83%	-
F-measure	97,45%	86,17%	-
Akurasi	95,69%		

Berdasarkan Tabel 5.17 dapat diketahui hasil evaluasi dari *machine learning* terhadap data uji. *Precision* adalah proporsi masing-masing kelas sentimen prediksi yang tepat dari semua kelas sentimen yang diprediksi. Pada kelas positif kemampuan sistem dalam memprediksi kelas positif dengan tepat dari semua kelas positif yang diprediksi sebesar 96,54%, sedangkan kelas negatif sebesar 90,83%.

Recall adalah proporsi masing-masing kelas sentimen aktual yang tepat terprediksi dari semua kelas sentimen yang diprediksi. Pada kelas positif kemampuan sistem untuk menemukan kelas positif aktual yang terprediksi dengan tepat sebesar 98,38%, sedangkan kelas negatif sebesar 81,95%.

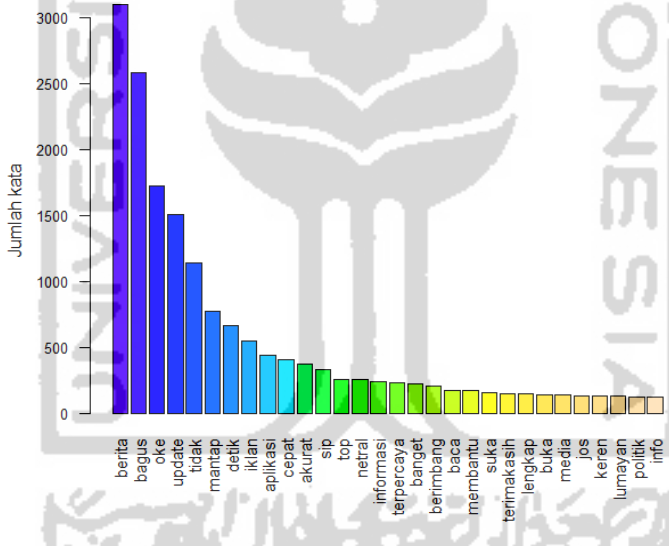
F-measure adalah proporsi pengukur dari kinerja sistem dari masing-masing kelas sentimen. Kinerja sistem dalam pengklasifikasian kelas positif sebesar 97,45% sedangkan kelas negatif sebesar 86,17%.

Dari tabel tersebut diketahui nilai dari *precision*, *recall,* dan *f-measure* untuk kedua kelas sentimen cukup tinggi. Adapun nilai untuk kelas sentimen positif lebih tinggi dibanding kelas sentimen negatif, dapat disebabkan karena jumlah data untuk kelas positif lebih banyak daripada kelas negatif. Dengan nilai akurasi sebesar 95,69% dapat dikatakan tingkat ketepatan hasil prediksi menggunakan metode *Maximum Entropy* terhadap data aktual secara keseluruhan tinggi.

5.4 Analisis Performa Detikcom

5.4.1 Visualisasi Data Sentimen

Visualisasi data memberikan gambaran secara umum informasi, topik dan bahasan yang sering diulas oleh pengguna aplikasi Detikcom dengan lebih mudah dan menarik.



Gambar 5.5. Kata yang paling banyak muncul

Berdasarkan Gambar 5.5 dapat dilihat kata yang paling banyak muncul secara keseluruhan. Kata berita (3094 kali), bagus (2576 kali), oke (1725 kali), *update* (1504 kali), tidak (1137 kali), mantap (773 kali), detik (666 kali), iklan (546 kali), aplikasi (443 kali), cepat (404 kali), dan seterusnya. Kumpulan kata-kata yang sering muncul tersebut dapat dibuat dalam bentuk *wordcloud* seperti pada Gambar 5.6 berikut.

a) b)

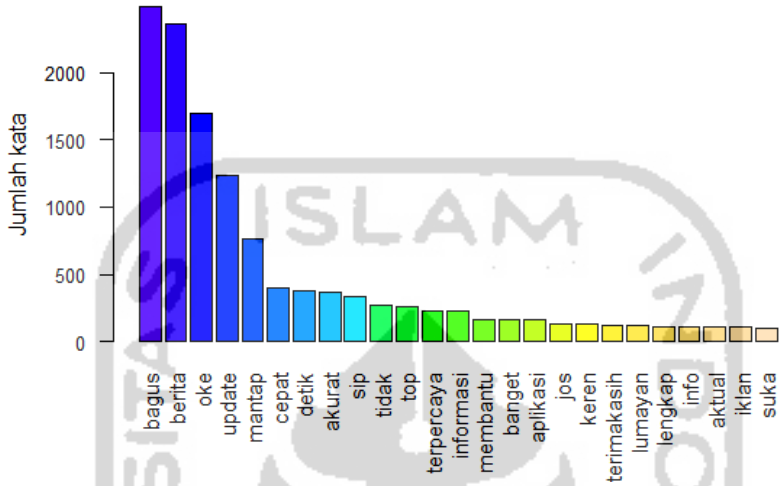
Gambar 5.6. *Wordcloud* a) semua kata; b) tanpa 5 kata terbanyak

Berdasarkan Gambar 5.6 dapat dilihat *wordcloud* untuk kata yang paling banyak muncul secara keseluruhan. Dengan menggunakan *wordcloud* dapat dilihat dengan lebih jelas kata-kata yang sering muncul, semakin besar ukuran teks maka semakin besar juga frekuensi yang dimiliki oleh kata tersebut. Berdasarkan *wordcloud* tersebut dapat dilihat bahwa kata berita, tidak, iklan, aplikasi, dan bagus merupakan kata yang mendominasi artinya semakin besar ukuran teks maka semakin besar juga frekuensi yang dimiliki oleh kata tersebut. Pada dasarnya karena aplikasi Detikcom adalah portal berita, maka sudah dipastikan bahwa kata berita dan aplikasi adalah topik yang paling sering diulas.

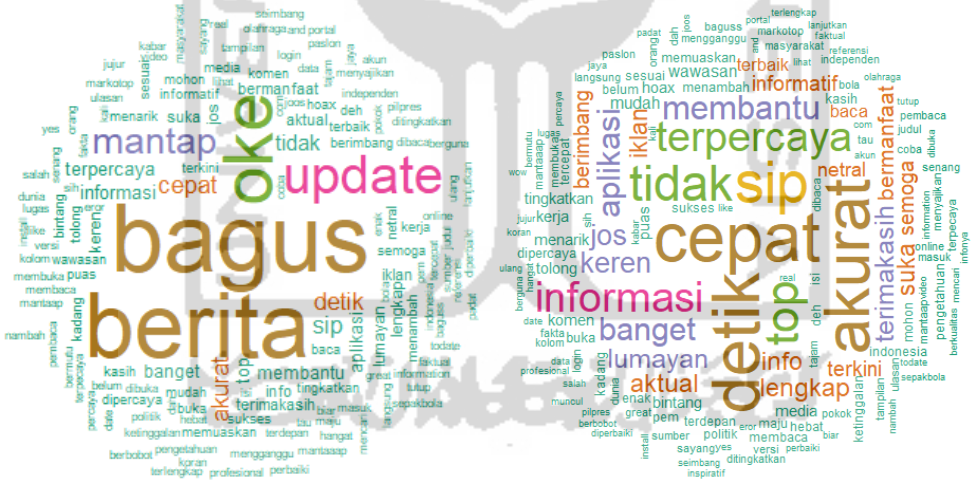
Secara umum dapat dilihat bahwa frekuensi terbanyak umumnya terdapat pada kata positif terlihat pada kata bagus, oke, mantap, dan cepat. Hal ini dapat diartikan bahwa mayoritas pengguna merasa puas terhadap aplikasi Detikcom. Namun ini juga dapat dipengaruhi oleh jumlah perbandingan ulasan kelas sentimen, dimana jumlah ulasan sentimen positif yang jauh lebih banyak dibandingkan sentimen negatif. Adapun visualisasi dari masing-masing kelas sentimen sebagai berikut.

1. Sentimen Positif

Kelas sentimen positif berisi ulasan terhadap Detikcom yang dinilai baik oleh penguna. Jumlah ulasan untuk tanggapan positif adalah 10199. Berikut informasi dan visualisasi untuk kelas sentimen positif.



Gambar 5.7. Kata yang paling banyak muncul kelas positif



a) b)

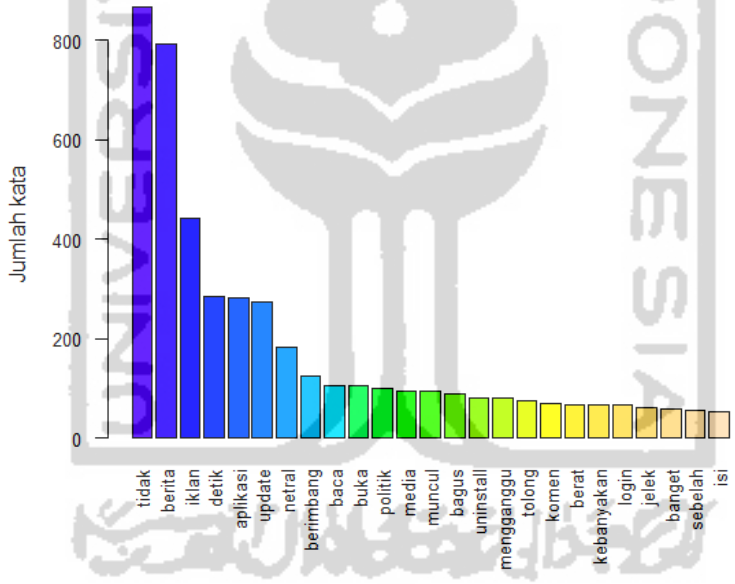
Gambar 5.8. *Wordcloud* kelas positif a) semua kata; b) tanpa 5 kata terbanyak

Berdasarkan grafik tersebut dapat dilihat kata yang paling banyak muncul secara keseluruhan. Kata bagus (2486 kali), berita (2361 kali), oke (1694 kali), *update* (1230 kali), mantap (767 kali), cepat (394 kali), detik (382 kali), akurat (365 kali), sip (331 kali), dan seterusnya. Kata-kata yang muncul seperti pada Gambar 5.7 merupakan kata yang memiliki sentimen positif

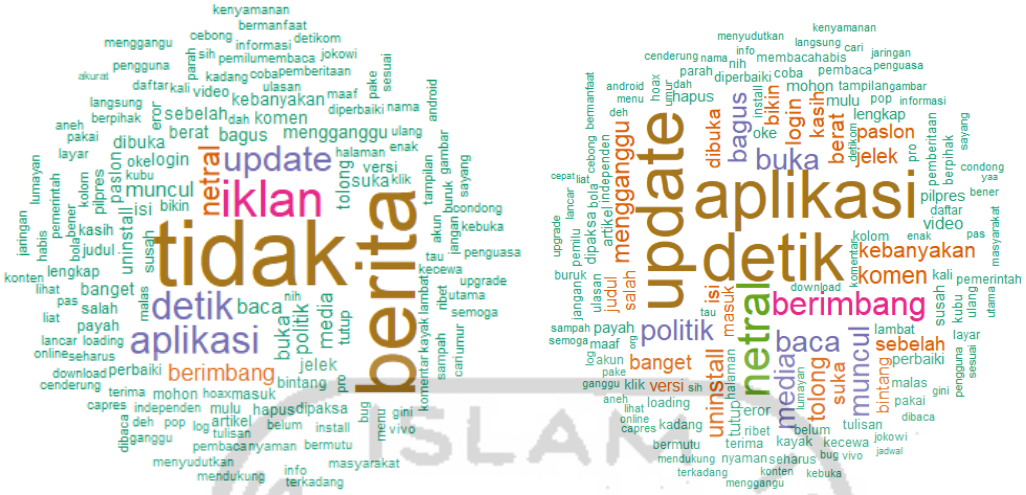
yakni kata-kata yang menjadi topik pembicaraan atau bahasan positif yang paling banyak diulas oleh pengguna detikcom. Kumpulan kata-kata yang sering muncul tersebut juga dapat ditampilkan dalam bentuk *wordcloud* seperti terlihat pada Gambar 5.8. Pada *wordcloud* dapat dilihat bahwa kata bagus, berita, oke, *update*, mantap, dan cepat merupakan kata yang mendominasi artinya semakin besar ukuran teks maka semakin besar juga frekuensi yang dimiliki oleh kata tersebut.

2. Sentimen Negatif

Kelas sentiment negatif berisi ulasan terhadap Detikcom yang dinilai kurang baik oleh penguna. Jumlah ulasan untuk tanggapan negarif adalah 1269. Berikut informasi dan visualisai untuk kelas sentimen negarif.



Gambar 5.9. Kata yang paling sering muncul kelas negatif



Gambar 5.10. *Wordcloud* kelas negatif a) semua kata; b) tanpa 3 kata terbanyak

Berdasarkan Gambar 5.9 dapat dilihat kata yang paling sering muncul secara keseluruhan. Kata tidak (886 kali), berita (793 kali), iklan (441 kali), detik (284 kali), aplikasi (283 kali), *update* (274 kali), netral (182 kali), berimbang (125 kali), dan selanjutnya. Kata-kata yang muncul seperti pada Gambar 5.10 merupakan kata yang memiliki sentimen negatif yakni kata-kata yang menjadi topik pembicaraan atau bahasan negatif yang paling banyak diulas oleh pengguna Detikcom. Kumpulan kata-kata yang sering muncul tersebut juga dapat ditampilkan dalam bentuk *wordcloud* seperti terlihat pada Gambar 5.10. Pada *wordcloud* dapat dilihat bahwa kata tidak, berita, iklan, detik, aplikasi, dan *update* merupakan kata yang mendominasi artinya semakin besar ukuran teks maka semakin besar juga frekuensi yang dimiliki oleh kata tersebut.

Adapun pada penelitian ini algoritma yang digunakan masih belum dapat memperoses kata ‘tidak’ secara maksimal. Sehingga kata ‘tidak’ yang seharusnya masuk ke dalam tahapan *prepocessing* masih belum dapat dihilangkan. Pada Gambar 5.10 (b) dapat dilihat lebih jelas bentuk *wordcloud* tanpa kata tersebut.

5.4.2 Asosiasi Kata

Asosiasi kata diperoleh dengan menggunakan pendekatan nilai korelasi terhadap masing-masing kata terhadap kemungkinan suatu kata di ulas bersamaan dengan kata lainya. Berdasarkan kata-kata yang paling banyak muncul dapat diperoleh asosiasi antarkata pada masing-masing kelas sentimen secara bersamaan guna memperoleh informasi. Adapun asosiasi kata dari masing-masing kelas sentimen sebagai berikut.

1. Sentimen Positif

   Asosasi jata pada klasifikasi kelas sentimen positif dapat dilihat pada Tabel 5.18 berikut.

Tabel 5.18. Asosiasi kata positif

Bagus		Detik		Terpercaya	
Kerja	0,11	Handal	0,22	Tajam	0,21
Performa	0,05	Paslon	0,20	Sumber	0,13
Penayangan	0,04	Nyaman	0,18	Aktual	0,10
Informasi	0,03	Paham	0,18	Media	0,06
Berita		Kelompok	0,18	Isi	0,06
Condong	0,16	Condong	0,18	Tanggungjawab	0,06
Kampanye	0,16	Kampanye	0,18	Membantu	
Paslon	0,14	Cepat		Lokal	0,15
Lengkap	0,13	Respon	0,07	Sibuk	0,13
Berimbang	0,13	Materi	0,07	Pencarian	0,11
Handal	0,12	Install	0,06	Suka	0,11
Update		Info	0,05	Aplikasi	
Berkali	0,11	Informasi		Handal	0,29
Ergonomis	0,11	Menambah	0,13	Menyuguhkan	0,29
Pertimbangan	0,11	Reformasi	0,13	Nyaman	0,28
Akurat		Perselisihan	0,13	Cocok	0,22
Terpecaya	0,09	Bangsa	0,11	Performa	0,22
Tajam	0,07	Pengetahuan	0,11	Tujuan	0,22

Berdasarkan tabel tersebut kata “bagus” berasosiasi dengan beberapa kata seperti “kerja”, “performa”, “penayangan”, dan “informasi” dengan nilai asosiasi ≥ 0.03. Berdasarkan asosiasi tersebut, dapat diketahui pengguna telah menganggap kinerja dan performa aplikasi Detikcom secara keseluruhan bagus. Khususnya pada cara penyanganan dan informasi yang diberikan.

Kata “berita” berasosiasi dengan beberapa kata seperti “condong”, “kampanye”, “paslon”, “lengkap”, “berimbang”, dan “handal” dengan nilai asosiasi ≥ 0.12. Berdasarkan asosiasi tersebut, dapat diketahui pengguna telah menganggap berita yang disampaikan lengkap, berimbang, dan handal. Namun pada pengguna juga masih menganggap berita yang disampaikan masih terdapat kampanye dan condong pada pasangan calon (paslon) tertentu selama pemilu.

Kata “*update*” berasosiasi dengan beberapa kata seperti “berkali”, “ergonomis”, dan “pertimbangan” dengan nilai asosiasi ≥ 0.11. Berdasarkan asosiasi tersebut, dapat diketahui pengguna merasa *update* yang dilakukan masih butuh pertimbangan dan dilakukan berkali-kali agar diperoleh aplikasi yang ergonomis.

Kata “akurat” berasosiasi dengan beberapa kata seperti “terpercaya”, “tajam”, dan “dijaga” dengan nilai asosiasi ≥ 0.06. Berdasarkan asosiasi tersebut, dapat diketahui pengguna telah menganggap Detikcom adalah aplikasi yang akurat, tajam dan terpecaya dan perlu dijaga.

Kata “detik” berasosiasi dengan beberapa kata seperti “handal”, “paslon”, “nyaman”, “paham”, “kelompok”, “condong”, dan “kampanye” dengan nilai asosiasi ≥ 0.18. Berdasarkan asosiasi tersebut, dapat diketahui pengguna telah menganggap Detikcom sebagai aplikasi yang handal, nyaman, dan memberikan informasi yang mudah dipahami. Namun pada pengguna juga masih menganggap Detikcom masih terdapat kampanye dan condong pada kelompok ataupun paslon tertentu.

Kata “cepat” berasosiasi dengan beberapa kata seperti “respon”, “materi”, “install”, dan “info” dengan nilai asosiasi ≥ 0.05. Berdasarkan asosiasi tersebut, dapat diketahui pengguna telah menganggap Detikcom cepat dalam memberikan respon dan penyaluran materi infomasi. Selain itu Detikcom juga cepat dalam proses penginstallan.

Kata “informasi” berasosiasi dengan beberapa kata seperti “menambah”, “reformasi”, “perselisihan”, “bangsa”, dan “pengetahuan” dengan nilai asosiasi ≥ 0.11. Berdasarkan asosiasi tersebut, dapat diketahui

pengguna telah menganggap informasi yang disampaikan menambahkan pengetahuan. Salah satu ulasan yang sedang banyak diinformasikan terutama saat pemilu adalah tentang reformasi dan perselisihan bangsa.

Kata "terpercaya" berasosiasi dengan beberapa kata seperti "tajam", "sumber", "aktual", "media", "isi", dan "tanggungjawab" dengan nilai asosiasi ≥ 0.06. Berdasarkan asosiasi tersebut, dapat diketahui pengguna telah menganggap Detikcom adalah aplikasi yang terpercaya, tajam, dan akurat dengan isi media yang terpercaya dan dapat dipertanggungjawabkan.

Kata "membantu" berasosiasi dengan beberapa kata seperti "lokal", "sibuk", "pencarian", dan "suka" dengan nilai asosiasi ≥ 0.11. Berdasarkan asosiasi tersebut, dapat diketahui pengguna menganggap Detikcom adalah aplikasi yang membantu dalam pencarian berita-berita lokal terutama bagi pengguna yang sibuk, sehingga banyak disukai pengguna.

Kata "aplikasi" berasosiasi dengan beberapa kata seperti "handal", "reformasi", "menyuguhkan", "nyaman", "cocok", "perfoma", dan "tujuan" dengan nilai asosiasi ≥ 0.22. Berdasarkan asosiasi tersebut, dapat diketahui pengguna merasa perfoma aplikasi dinilai handal, nyaman, dan cocok digunakan. Salah satu ulasan yang sedang banyak disunguhkan adalah tentang reformasi.

2. Sentimen Negatif

Asosiasi kata pada klasifikasi kelas sentimen negatif dapat dilihat pada Tabel 5.19. Berdasarkan table 5.19 diketahui bahwa kata "tidak" berasosiasi dengan beberapa kata seperti "*login*", "harapan", "jelas", dan "sejalan" dengan nilai asosiasi ≥ 0.23. Berdasarkan asosiasi tersebut, dapat diketahui pengguna merasa kesulitan dalam login kedalam aplikasi. Selain itu, pengguna juga merasa berita yang dipaparkan tidak jelas, tidak sejalan, ataupun tidak sesuai harapan mereka.

Tabel 5.19. Asosiasi kata negatif

Tidak		Berita		Iklan	
Login	0,31	Dukungan	0,30	Merusak	0,48
Harapan	0,26	Vulgar	0,27	Senonoh	0,44
Jelas	0,26	Paslon	0,26	Hapus	0,34
Sejalan	0,23	Bingung	0,24	Ganti	0,32
Detik		Pemilu	0,23	Dominasi	0,24
Kolega	0,37	Komplain	0,22	Risih	0,24
Kritis	0,30	Aplikasi		*Update*	
Skip	0,28	Propaganda	0,39	Kalinya	0,33
Video	0,25	Menghapus	0,37	Kinerja	0,33
Uninstall	0,24	Unduhan	0,35	Waktu	0,33
Netral		Ganti	0,30	Berubah	0,29
Paslon	0,29	Berimbang		Tampilan	0,26
Mendukung	0,25	Pilpres	0,27	Politik	
Negatif	0,22	*Uninstall*	0,25	Situasi	0,23
Nilai	0,21	Miris	0,25	Cenderung	0,22
Informasi	0,21	Komplain	0,17	Bisnis	0,19
		Sosmed	0,17	Membenahi	0,19
				Panas	0,19

Kata “berita” berasosiasi dengan beberapa kata seperti “dukungan”, “vulgar”, “paslon”, “bingung”, “pemilu”, dan “komplain” dengan nilai asosiasi ≥ 0.22. Berdasarkan asosiasi tersebut, dapat diketahui pengguna menganggap berita yang disampaikan berisi dukungan kepada pasangan calon (paslon) tertentu selama pemilu. Selain itu, pengguna mengkomplain tentang berita vulgar dan membingungkan.

Kata “iklan” berasosiasi dengan beberapa kata seperti “merusak”, “senonoh”, “hapus”, “ganti”, “dominasi”, dan “risih” dengan nilai asosiasi ≥ 0.24. Berdasarkan asosiasi tersebut, dapat diketahui pengguna menganggap iklan yang terdapat diaplikasi terlalu mendominasi sehingga merusak dan menganggu/risih pengguna. Selain itu, isi iklan dianggap kurang senonoh dan perlu dihapus ataupun diganti.

Kata “detik” berasosiasi dengan beberapa kata seperti “kolega”, “kritis”, “skip”, “video”, dan “uninstall” dengan nilai asosiasi ≥ 0.24. Berdasarkan asosiasi tersebut, dapat diketahui pengguna menganggap

Detikcom sebagai aplikasi yang masih memihak pada kolega dan dianggap kurang kritis. Selain itu, masih terdapat video-video yang tidak ingin ditonton dan sering kali di-skip pengguna di Detikcom. Pengguna yang tidak merasa puas sering meng-*uninstall* aplikasi Detikcom.

Kata "aplikasi" berasosiasi dengan beberapa kata seperti "propaganda", "menghapus", "unduhan", dan "ganti" dengan nilai asosiasi ≥ 0.30. Berdasarkan asosiasi tersebut, dapat diketahui pengguna menganggap aplikasi berisi propaganda. Beberapa pengguna memutuskan untuk menghapus ataupun menganti unduhan aplikasi Detikcom.

Kata "*update*" berasosiasi dengan beberapa kata seperti "kalinya", "kinerja", "waktu", "berubah" dan "tampilan" dengan nilai asosiasi ≥ 0.26. Berdasarkan asosiasi tersebut, dapat diketahui pengguna merasa *update* yang dilakukan terlau sering (berkali-kali) dalam kurun waktu yang singkat. Namun, perubahan tidak dimbangi kinerja maupun tampilan yang dianggap tidak lebih baik.

Kata "netral" berasosiasi dengan beberapa kata seperti "paslon", "mendukung", "negatif", "nilai", dan "informasi" dengan nilai asosiasi ≥ 0.21. Berdasarkan asosiasi tersebut, dapat diketahui pengguna menganggap aplikasi Detikcom kurang netral dalam penyampaian nilai dan informasi terutama pada pemilu. Pengguna merasa Detikcom mendukung pasangan calon (paslon) tertentu dan memberi informasi negatif pada lainnya.

Kata "berimbang" berasosiasi dengan beberapa kata seperti "pilpres", "uninstall", "miris", "komplain", dan "sosmed" dengan nilai asosiasi ≥ 0.17. Berdasarkan asosiasi tersebut, dapat diketahui pengguna menganggap aplikasi Detikcom kurang berimbang terutama pada saat pemilihan presiden (pilpres). Ketidak berimbangan tersebut membuat pengguna merasa miris ataupun komplain di sosial media, hingga memutuskan untuk meng-*uninstall* aplikasi Detikcom.

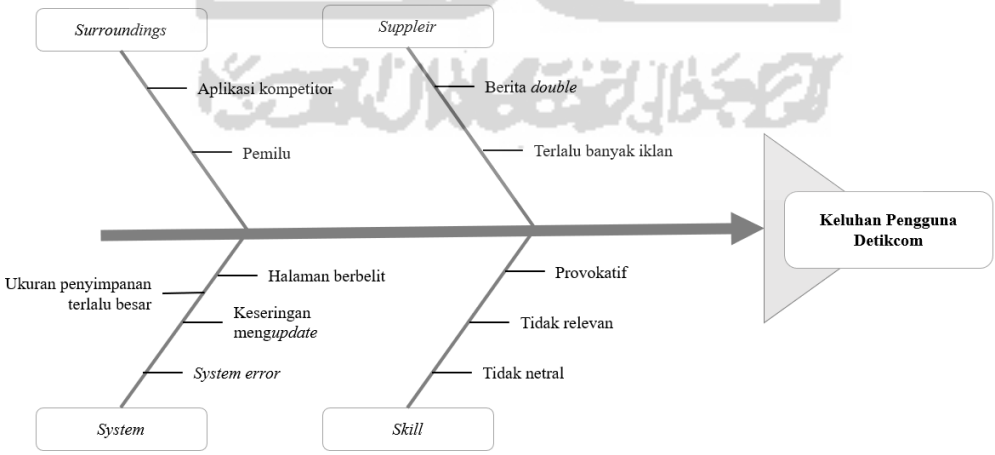
Kata "politik" berasosiasi dengan beberapa kata seperti "situasi", "cenderung", "bisnis", "membenahi", dan "panas" dengan nilai asosiasi ≥ 0.19. Berdasarkan asosiasi tersebut, dapat diketahui pengguna menganggap

Detikcom membuat situasi politik semakin memanas. Pengguna menganngap Detikcom menciptakan politik yang cenderung memihak dan sering kali dianggap bisnis sehingga pengguna meminta pihaknya untuk segera membenahi hal tersebut.

Secara umum dapat dilihat bahwa nilai asosiasi untuk kelas sentimen positif lebih rendah dibandingkan sentiemen negatif. Salah satu penyebabnya adalah jumlah perbandingan ulasan kelas sentimen, dimana ulasan sentimen positif yang jauh lebih banyak dibadingkan sentimen negatif. Pada sentimen positif, ulasan pengguna umumnya lebih sering memberikan pujian singkat dan jarang membicarakan kelebihan aplikasi Detikcom secara spesifik. Sedangkan pada kelas sentimen negatif, ulasan pengguna umumnya lebih kritis dan lebih terperinci terhadap keluhan yang dirasakan.

5.4.3 Diagram *Fishbone*

Diagram fishbone merupakan salah satu alat untuk mengidentifikasi secara garfik dari sebab dan akibat suatu permasalahan. Berdasarkan pemaparan hasil ulasan negatif yang diperoleh dari data ulasan, maka didapatkan informasi mengenai permasalahan yang terjadi terkait dengan keluhan pengguna Detikcom seperti berikut.



Gambar 5.11. Diagram *fishbone* komplain penguna Detikcom

Pada Gambar 5.11 dapat diketahui informasi mengenai faktor-faktor penyebab aplikasi Detikcom mendapatkan ulasan negatif dari pengguna. Karena media berita berada pada industry layanan dan jasa maka konsep penjelasan *fishbone* dilihat dari segi *surroundings*, *supplier*, *system*, dan *skill*. Setelah diketahui penyebab keluhan pengguna dapat ditentukan pemecahan masalah.

Adapun rencana pemecahan masalah seperti pada Tabel 5.20 berikut.

Tabel 5.20. Rencana pemecahan masalah

No	Faktor	Permasalahan	Pemecahan Masalah
1	*Surroundings*	Aplikasi kompetitor	Meningkatkan performa Detikcom agar tidak kalah dengan aplikasi berita lainnya baik dari segi aplikasi itu sendiri ataupun dengan peningkatan mutu isi berita yang disajikan.
		Pemilu	Menyampaikan berita dengan lebih netral lagi dan tidak terpengaruh ataupun mengambil sisi dari kubu politik tertentu terutama di masa-masa kritis seperti pada pemilihan umum (pemilu).
2	*Supplier*	Berita *double* (diulang-ulang)	Mengurangi berita yang tidak terlalu relevan ataupun berita serupa yang sudah dipublikasi. Selain itu, perlu koordinir yang lebih baik lagi dari tugas wartawan agar tidak memberikan berita *double*.
		Terlalu banyak iklan	Mengurangi jumlah iklan terutama pada berita-berita singkat seperti kategori 20detik.
3	*System*	Keseringan *update*	Mengurangi jumah *update* aplikasi dalam kurun waktu yang dekat apabila tidak terlalu urgen.
		Halaman berbelit	Mengurangi pemotongan berita menjadi beberapa halaman terutama untuk isi berita yang tidak terlalu panjang, apabila diharuskan dimaksimal hanya sampai 2 halaman.

No	Faktor	Permasalahan	Pemecahan Masalah
		Ukuran penyimpanan aplikasi (*size*) terlalu besar	Mengurangi fitur-fitur aplikasi yang tidak terlalu penting dan melakukan press (pengecilan ukuran aplikasi yang dianggap terlalu membutuhkan ruang penyimpanan yang besar).
		Sistem *error*	Membenahi sistem aplikasi Detikcom yang *error* terutama pada sistem *error* yang biasanya terjadi yaitu susah login, aplikasi menutup sendiri, lama *loading* dan sebagainya.
4	*Skill*	Isi berita tidak relevan	Memberikan berita yang lebih relevan lagi sesuai dengan kejadian dan masalah yang terjadi saat itu. Memfokuskan pada penyampian informasi penting bagi pengguna dengan tidak mengulang-ulang berita yang mungkin tidak ada sangkut pautnya berkali-kali.
		Isi berita tidak netral	Memberikan berita yang lebih beragam lagi, dan berusaha lebih netral lagi dengan memfokuskan penyampaian fakta dan tidak mengiring opini yang memihak ataupun menjatuhkan kepada pihak tertentu terutama pada politik.
		Berita provokartif	Menyapaikan berita dengan lebih mengedepankan fakta dari pada opini, tidak mengadudomba pihak ataupun mengiring suatu opini. Lebih berhati-hati dalam menyampaikan berita yang kurang jelas kebenarannya ataupun berita sensitif yang dapat menimbulkan kegaduhan di masyarakat.

BAB 6 PENUTUP

6.1 Kesimpulan

Berdasarkan hasil analisis dan pembahasan dapat diperoleh beberapa kesimpulan sebagai berikut.

1. Selama tahun 2019 terdapat 12.257 ulasan aplikasi berita *online* Detikcom *website Google Play*. Jumlah ulasan tertinggi pada bulan April dan September dipengaruhi oleh pelaksanaan pemilihan umum (pemilu). Secara umum pengguna merasa puas dengan kinerja Detikcom terlihat dari mayoritas pengguna memberikan *rating* 5 (8100 ulasan) diikuti *rating* 4 (1664 ulasan), *rating* 1 (1258 ulasan), *rating* 3 (830 ulasan), dan terendah *rating* 2 (830 ulasan). Adapun pada bulan September pengguna merasa paling puas dengan proporsi *rating* 5 tertinggi (70,8%) dan bulan Desember paling tidak puas dengan proporsi *rating* 1 tertinggi (26,3%).
2. Pengklasifikasian data menjadi kelas sentimen positif sebanyak 10.199 ulasan dan sentimen negatif sebanyak 1997 ulasan. Hasil dari penerapan metode algoritma *Maximum Entropy* dalam mengklasifikasikan data ulasan pengguna Detikcom menjadi kelas positif dan negatif dengan perbandingan data latih dan data uji sebesar 80%:20% diperoleh hasil klasifikasi sentimen dengan tingkat akurasi sebesar 91,6%.
3. Pembangunan model dalam pembentukan data latih data uji dapat meningkatkan ketepatan *machine learning* dari klasifikasi. Dari 5 percobaan diperoleh dataset dengan akurasi tertinggi pada iterasi ke 5 sebesar 95,69% dengan rata-rata akurasi 5-*fold cross validation* sebesar 90,05%. Berdasarkan *cross validation,* kinerja sistem dalam pengklasifikasian kelas positif sebesar 97,45% sedangkan kelas negatif sebesar 86,17%. Dari nilai-nilai tersebut dapat dikatakan tingkat ketepatan hasil prediksi menggunakan metode *Maximum Entropy* terhadap data aktual secara keseluruhan cukup tinggi.

4. Berdasarkan informasi yang diperoleh dari hasil klasifikasi dan asosiasi teks yang dilakukan, diketahui bahwa kata yang paling sering dibicarakan pengguna aplikasi Detikcom mengenai berita, bagus, oke, *update*, tidak, mantap, detik, iklan, aplikasi, dan cepat. Adapun kata yang banyak dibicarakan pada sentimen positif mengenai bagus, berita, oke, *update*, dan mantap, sedangkan pada sentimen negatif mengenai tidak, berita, iklan, detik, dan aplikasi. Kemudian dari diagram *fishbone* diketahui faktor-faktor yang menyebabkan aplikasi Detikcom memiliki ulasan negatif yaitu dari segi *surroundings* (perbandingan dengan aplikasi lain dan pemilu), *supplier* (berita *double* dan terlalu banyak iklan), *system* (keseringan *update*, halaman berbelit, *size* aplikasi terlalu besar, dan sistem *error*), dan *skill* (isi berita tidak relevan, tidak netral, dan provokartif).

6.2 Saran

Berdasarkan hasil analisis dan kesimpulan, dapat diberikan beberapa saran sebagai berikut.

1. Bagi pihak Detikcom, hasil dari ekstraksi informasi yang diperoleh dari ulasan-ulasan pengguna khususnya ulasan yang berbentuk negatif dapat dijadikan bahan evaluasi dalam peningkatan kepuasan pengguna dan memberikan pelayanan semaksimal mungkin, serta untuk pengembangan pembaharuan aplikasi selanjutnya.
2. Proses pelabelan kelas sentimen yang dilakukan dalam penelitian ini baru sebatas pada pendeteksian sentimen antar kata menggunakan kamus *lexicon*, sehingga kata-kata negasi belum dapat teridentifikasi dengan baik terutama pada penggunaan kata ‘tidak’. Pada penelitian selanjutnya sebaiknya dapat menggunakan sistem pelabelan yang memiliki tingkatan lebih tinggi, yakni mampu mendeteksi sentimen pada frasa dan kalimat.
3. Bagi peneliti selanjutnya, dapat menggunakan pendekatan *machine learning* lain sebagai pembanding performa algoritma *Maximum Entropy* untuk mengklasifikasi ulasan aplikasi berita *online* Detikcom ataupun pada situs lainnya selain *Google Play.*

DAFTAR PUSTAKA

Alexa. 2020. *Top site in* Indonesia. https://www.alexa.com/topsites/countries/ID. Diakses pada 26 Maret 2020.

Anggreini, Dyta. 2008. Klasifikasi Topik Menggunakan Metode *Naive Bayes* dan *Maximum Entropy* pada Artikel Media Massa dan Abstrak Tulisan. Skripsi. Universitas Indonesia.

Berry, M.W., & Kogan, J. 2010. *Text Mining Aplication and Theory*. United Kingdom: WILEY.

Davies & Beynon, P. 2004. *Database Systems Third Edition.* New York: Palgrave Macmillan.

Detik. 2018. Ini 74 Media yang Terverifikasi Dewan Pers. https://news.detik.com/berita/d-3413992/ini-74-media-yang-terverifikasi-dewan-pers. Diakses pada 30 Maret 2020.

Dewan Pers. 2020. Data Perusahaan Pers. https://dewanpers.or.id/data/perusahaanpers. Diakses pada 30 Maret 2020.

Fadlisyah, B.D.A. 2014. Statistika: Terapannya di Informatika, Edisi Pertama. Yogyakarta: Graha Ilmu.

Fawcett, T. 2006. *An Introduction to ROC Analysis. Pattern Recognition Letters* 27 (8): 861–874.

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

Fritz, G. 2016. Analisa *Bad Hike* pada Kran Lavatory Tipe S11234R Menggunakan Metode *Nominal Group Technique* dan Metode *Fishbone* di PT Surya Toto Indonesia Tbk. Skripsi. Program Diploma Teknik Mesin Sekolah Vokasi UGM Yogyakarta.

Google Play. 2020. Detikcom – Berita Terbaru & Terlengkap. https://play.google.com/store/apps/details?id=org.detikcom.rss Diakses pada 25 Maret 2020.

Gumilang, Z.A.N. 2018. Implementasi *Naïve Bayes Classifier* dan Asosiasi untuk Analisis Sentimen Data Ulasan Aplikasi *E-Commerce Shopee* pada Situs *Google Play*. Skripsi. Program Studi Statistika FMIPA UII Yogyakarta.

Han, K.J. 2001. *Data Mining: Concepts and Technique*. San Fransisco: John Wiley & Sons Inc.

Han, J., & Kamber, M. 2006. *Data Mining: Concepts and Techniques Second Edition*. San Francisco: Morgan Kauffman.

Han, J. & Kamber, M. 2012. *Data Mining: Concepts and Techniques Third Edition*. Waltham, MA: Morgan Kaufmann.

Jamil, H.N. 2017. Analisis Sentimen pada *Online Review* Menggunakan Kombinasi Metode *Lexicon Based* dan *Naïve Bayes Classifier*. Skripsi. Program Studi Statistika FMIPA UII Yogyakarta.

Kohavi, R., & Provost, F. 1998. *On applied research in machine learning*. *Machine Learning*, 30 (2): 127-132.

Larose, T. 2005, *Discovering Knowledge in Data: an Introduction To Data Mining*. San Fransisco: John Wiley & Sons Inc.

Lee, L., & Pang, B. 2008. *Opinion Mining and Sentiment Analysis. Foundation and Trends in Information Retrieval*, 2 (1-2): 1-135.

Liu, B. 2012. *Sentiment Analysis and Subjectivity. Synthesis Lectures on Human Language Technologies*. USA: Morgan & Claypool Publishers.

Lovelock, C., & Wirtz, J. 2005. Manajemen Pemasaran Jasa. Indonesia: Kelompok Gramedia Indeks.

MacLennan, J., Tang, Z., & Crivat, B. 2009. *Data Mining with Microsoft*. SQL Server 2008. USA: Wiley Publishing Inc.

Manning, C. D., Raghavan, P., & Schutze, H. 2009. *An Introduction to Information Retrieval – Online Edition*. Cambridge: Cambridge University Press.

Marres, R., Joan, V.F., & Wilson, P.G. 2013. *Document-level Sentiment Classification: An Empirical Comparison between SVM and ANN. Expert Systems with Applications*, 40 (2): 621–633.

Nigam, K., Lafferty, J., & Mccallum, A. 1999. *Using Maximum Entropy for Text Classification*. IJCAI-99 *Workshop on Machine Learning for Information Filtering*, 61-67.

Okezone. 2019. Geser Detikcom, Okezone.com Jadi Portal Berita Nomor 2 di Indonesia. https://techno.okezone.com/read/2019/05/20/207/2057940/. Diakses pada 30 Maret 2020.

Prakoso, R.W., Novianti, A., & Setianingsih, C. 2017. Analisis Sentimen Menggunakan *Support Vector Machine* dan *Maximum Entropy*. *E-Proceeding of Engineering* Vol. 4 (2): 2389-2395.

Pramudiono, I. 2003. Pengantar *Data Mining*: Menambang Permata Pengetahuan di Gunung Data. Materi Kuliah Umum IlmuKomputer.com

Praptiwi, D.Y. 2018. Analisis Sentimen *Online Review* Pengguna *E-Commerce* Menggunakan Metode *Support Vector Machine* dan *Maximum Entropy*. Skripsi. Program Studi Statistika FMIPA UII Yogyakarta.

Prasetyo, E., 2012. *Data Mining* Konsep dan Aplikasi menggunakan Matlab. Yogyakakara: Andi.

Putranti, N.D., & Winarko, E. 2014. Analisis Sentimen *Twitter* untuk Teks Berbahasa Indonesia dengan *Maximum Entropy* dan *Support Vector Machine*. *Indonesian Journal of Computing and Cybernetics Systems* Vol 8 (1): 91-100.

Putri, D.U.K. 2016. Implementasi Inferensi *Fuzzy* Mamdani untuk Keperluan Sistem Rekomendasi Berita Berbasis Konten. Skripsi. Program Studi Ilmu Komputer FMIPA UGM Yogyakarta.

Refaeilzadeh, P., Tang, L., & Liu, H. 2009. *Cross Validation. Editors: M. Tamer dan Ling Liu. Encyclopedia of Database Systems, Springer*. New York.

Sabily, A.F., Andikara, P.P., & Fauzi, M.A. 2019. Analisis Sentimen Pemilihan Presiden 2019 pada *Twitter* menggunakan Metode *Maximum Entropy*. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer VoI. 3 (5): 4204-4209.

Saraswati, N.S. 2011. *Text Mining* dengan *Metode Naive Bayes Classifier* dan *Support Vector Machines untuk Sentiment Analysis*. Skripsi. Program Studi Teknologi Informasi Fakultas Teknik UGM Yogyakarta.

Stephens, M. 2007. *A Histori of News, Third edition*. New York: Oxford University Press.

Sugiyono. 2011. Metode Penelitian Kualitataif dan R&D. Bandung: Alfabeta.

Susanti, A.R. 2016. Analisis Klasifikasi Sentimen *Twitter* Terhadap Kinerja Layanan *Provider* Telekomunikasi Menggunakan Varian *Naive Bayes*. Tesis. Institut Pertanian Bogor.

Syah, A.P., Adiwijaya, & Al Faraby, S. 2017. *Sentiment Analysis on Online Store Product Reviews with Maximum Entropy Method. E-Proceeding of Engineering* Vol. 4: 4632-4640.

Tan, P., Steinbach, M., & Karpatne, A. 2006. *Introduction to Data Mining*. USA: Addison-Wesley.

Ulwan, M.N. 2016. *Pattern Recognition* pada *Unstructured* Data Teks Menggunakan *Support Vector Machine* dan *Association*. Skripsi. Program Studi Statistika FMIPA UII Yogyakarta.

Wijayanti, W.N. 2014. Analisis Sentimen pada *Review* Pengguna Sistem Operasi *Windows Phone* dengan Menggunakan Metode *Support Vector Machine* (SVM). Skripsi. Program Studi Teknologi Informasi Fakultas Teknik UGM Yogyakarta.

Yates, R.B. & Neto, B.R. 1999. *Modern Information retrieval*. New York: Addison-Wesley

Yunus, Syarifudin. 2010. Jurnalistik Terapan. Bogor: Ghalia Indonesia.

Zafikri, A. 2008. Implementasi *Metode Term Frequency Inverse Document Frequency* (TF-IDF) pada Sistem Temu Kembali Informasi. Skripsi. Program Studi S-1 Ilmu Komputer FMIPA USU.

LAMPIRAN

Lampiran 1 Data Ulasan

No	*Rate*	*User*	Tgl	Bln	*Like*	*Review*
1	1	angga andi ardiansyah	1	1	10	sekarang detikcom terlalu banyak iklan, dan saya sangat tidak suka sekali mengganggu saat kita sedang fokus membaca berita. apalagi ada iklan tentang uplive atau sejenisnya yang menampilkan adegan wanita tidak sepatutnya. tolong perbaiki, jika masih banyak iklan seperti itu. saya akan uninstall deti...Ulasan Lengkapsekarang detikcom terlalu banyak iklan, dan saya sangat tidak suka sekali mengganggu saat kita sedang fokus membaca berita. apalagi ada iklan tentang uplive atau sejenisnya yang menampilkan adegan wanita tidak sepatutnya. tolong perbaiki, jika masih banyak iklan seperti itu. saya akan uninstall detikcom karena tidak ada manfaatnya. terima kasih
2	1	MinasaUpa alif	1	1	19	Makin hari makin ndag jelas, kemaren2 iklan makin banyak dan full screen, sekarang d tambah lagi iklan pake berita yang d dominasi grab, klo iklan ndag usah d jadikan berita dong. risih baca nya. Makin hari bukan nya makin berkembang dan manjakan pengguna tapi makin k urusan nyari duit. beriklan bol...Ulasan LengkapMakin hari makin ndag jelas, kemaren2 iklan makin banyak dan full screen, sekarang d tambah lagi iklan pake berita yang d dominasi grab, klo iklan ndag usah d jadikan berita dong. risih baca nya. Makin hari bukan nya makin berkembang dan manjakan pengguna tapi makin k urusan nyari duit. beriklan boleh tapi sewajar nya dong, terimakasih dengan ini saya uninstall aplikasi ini karena terlalu banyak iklan grab
3	3	umi iffa	1	1	0	Mantap
4	5	Pengguna Google	1	1	0	beritanya updet banget. mantap.
5	5	Iqbal Rizki	1	1	0	Aplikasi berita terbaik. update berita sangat cepat, jadi tidak takut lagi ketinggalan berita. pokoknya baguslah. saya kasi bintang 5 ☆☆☆☆☆
6	1	Fitri Wulandari	2	1	0	sering log out
7	1	Tsamara Nayyara	2	1	0	Berita yg dimuat sudah tidak berimbang. Saatnya uninstall. Bye detik
8	1	jason lim	2	1	1	iklan yg tiba2 muncul menjengkelkan,lgi asyik baca jadi gk asyik...terus yg 20 detik juga apaan itu iklan juga mesti...
9	1	shifuu ady	2	1	1	popup iklan MENGGANGGU !!!!
10	3	Pengguna Google	2	1	1	mau kasih komen nanti di kolom komen nanti susah amat. tolong jika tidak bener2 paham dengan berita jangan diposting dulu. mana ada selang cuci darah 7000 s/d 17000an, bedakan "blood tranfution set" dengan "arteri venous blood line" Terima kasih
11	3	Anthoni Anthoni	2	1	1	iklan sangat menggangu
12	1	Heri Yanto	3	1	1	makin lama detikcom makin kaya tai... lelet nya super lelet... gimana orang mau baca klo sering error... dikit2 detikcom tidak menanggapi...
13	1	Sesairo San	3	1	0	Beritanya tidak netral
14	2	Danang Wijayanto	3	1	1	iklanya gak bermutu..tau2 muncul
15	3	Pengguna Google	3	1	0	ini parasit tuh iklannya😒

No	*Rate*	*User*	Tgl	Bln	*Like*	*Review*
…	…	…	…	…	…	…
12243	5	abu nawas	29	12	0	Detikcom, oce !
12244	5	ahmad donda	29	12	0	Nur
12245	5	Yakob Pangihutan	29	12	0	Mantap
12246	5	Diding Suhardi	29	12	2	Informatif dan aktual.. keren..
12247	1	Adhy Blokn	30	12	0	Aplikasih anjing
12248	1	Gabriel Zena	30	12	0	Beritanya gak bermutu..
12249	2	har yanto	30	12	4	Semakin tidak nyaman menggunakan nya dengan tampilan yang sekarang , bila ingin memilih kategori olahraga harus masuk dulu baru pilih, model yang lama lebih baik .
12250	4	Abdul Cholik	30	12	0	oke
12251	5	Mukti Wibowo	30	12	0	Jooz
12252	5	Bil Busri	30	12	1	Trimakasih sudah di update, nonton tv sudah bisa full screen.
12253	5	Brian Angelus	30	12	0	Good
12254	5	Soedjali Putra	30	12	0	Mantab
12255	5	Nurul Hidayat	30	12	0	Yes..... Good....
12256	1	Zaeski Abdi	31	12	0	Lg asik baca berita tiba, kembali ke halaman utama terus
12257	5	bunga pamitha	31	12	0	Mantulll

Untuk 15 ulasan pertama dan 15 ulasan terakhir selama tahun 2019

Data dapat diakses pada laman https://play.google.com/store/apps/details?id=org.detikcom.rss&hl=id

Lampiran 2 Jumlah *rating* pengguna per bulan selama tahun 2019

Bulan	***Rating***					**Total**
	1*	**2***	**3***	**4***	**5***	
Januari	127	51	90	140	810	**1218**
Februari	168	54	60	73	392	**747**
Maret	103	29	34	58	267	**491**
April	350	117	307	720	3450	**4944**
Mei	96	27	56	92	543	**814**
Juni	46	20	20	47	225	**358**
Juli	36	8	25	26	134	**229**
Agustus	39	20	17	30	135	**241**
September	136	41	146	315	1549	**2187**
Oktober	61	12	31	97	323	**524**
November	26	8	19	31	154	**238**
Desember	70	18	25	35	118	**266**
Total	**1258**	**405**	**830**	**1664**	**8100**	**12257**

Lampiran 3 Proporsi *rating* pengguna per bulan selama tahun 2019

Bulan	***Rating***				
	1*	**2***	**3***	**4***	**5***
Januari	10,4%	4,2%	7,4%	11,5%	66,5%
Februari	22,5%	7,2%	8,0%	9,8%	52,5%
Maret	21,0%	5,9%	6,9%	11,8%	54,4%
April	7,1%	2,4%	6,2%	14,6%	69,8%
Mei	11,8%	3,3%	6,9%	11,3%	66,7%
Juni	12,8%	5,6%	5,6%	13,1%	62,8%
Juli	15,7%	3,5%	10,9%	11,4%	58,5%
Agustus	16,2%	8,3%	7,1%	12,4%	56,0%
September	6,2%	1,9%	6,7%	14,4%	70,8%
Oktober	11,6%	2,3%	5,9%	18,5%	61,6%
November	10,9%	3,4%	8,0%	13,0%	64,7%
Desember	26,3%	6,8%	9,4%	13,2%	44,4%

Lampiran 4 Sintaks *R Preprocessing* Data

```
# Load packages
library("NLP")
library("tm")
library("SnowballC")
library("RColorBrewer")
library("wordcloud")
library(stringr)

setwd("E://SK/TA/DATA")
docs<-readLines("detik-teks.csv")
summary(docs)

# Load the data as a corpus
docs <- Corpus(VectorSource(docs))

#Inspect the content of the document
inspect(docs)

#Replacing "/", "@" and "|" with space:
toSpace <- content_transformer(function (x , pattern )
gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")

#Cleaning the text
# Convert the text to lower case
docs <- tm_map(docs, content_transformer(tolower))

#Remove punctuation
docs <- tm_map(docs, toSpace, "[[:punct:]]")

#Remove numbers
docs <- tm_map(docs, toSpace, "[[:digit:]]")

# add two extra stop words: "available" and "via"
myStopwords = readLines("stopword_id.csv")

# remove stopwords from corpus
docs <- tm_map(docs, removeWords, myStopwords)

# Remove your own stop word
# specify your stopwords as a character vector
docs <- tm_map(docs, removeWords, c("you","nya"))

# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)

# Remove URL
removeURL <- function(x) gsub("http[[:alnum:]]*", " ", x)
docs <- tm_map(docs, removeURL)

#Replace words
docs <- tm_map(docs, gsub, pattern="kgk", replacement="tidak")
```

```
docs <- tm_map(docs, gsub, pattern="skrg", replacement="sekarang")
docs <- tm_map(docs, gsub, pattern="tgl", replacement="tolong")
docs <- tm_map(docs, gsub, pattern="bgt", replacement="banget")
docs <- tm_map(docs, gsub, pattern="thx", replacement="terimakasih")
docs <- tm_map(docs, gsub, pattern="jd", replacement="jadi")
docs <- tm_map(docs, gsub, pattern="ga", replacement="tidak")
docs <- tm_map(docs, gsub, pattern="nggak", replacement="tidak")
docs <- tm_map(docs, gsub, pattern="baik", replacement="bagus")
docs <- tm_map(docs, gsub, pattern="update",
replacement="terbaru")
docs <- tm_map(docs, gsub, pattern="detikcom",
replacement="detik")
docs <- tm_map(docs, gsub, pattern="login", replacement="buka")

# specify your stopwords as a character vector
#docs <- tm_map(docs, removeWords, c("berita","tidak"))

#Build a term-document matrix
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 25)

#barplot
k<-barplot(d[1:20,]$freq, las = 2, names.arg =
d[1:20,]$word,cex.axis=1,cex.names=1,
        main ="Most frequent words",ylab = "Frekuensi kata",col
=topo.colors(20))

#Generate the Word cloud
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=100, random.order=TRUE, rot.per=0.35,
          colors=brewer.pal(7, "Dark2"))

dataframe<-data.frame(text=unlist(sapply(docs, `[`)),
stringsAsFactors=F)

write.csv(dataframe, "E://SK/TA/DATA/DETIKCleaning.csv")
save.image()
```

Lampiran 5 Sintaks *R* Pelabelan

```
 library(tm)
setwd("E://SK/TA/DATA")
kalimat2 <- read.csv("DETIKCleaning.csv", header=TRUE)

#skoring
positif <- scan("kata-pos.txt", what="character",comment.char=";")
negatif <- scan("kata-neg.txt", what="character",comment.char=";")
kata.positif <- c(positif, "is near to")
kata.negatif <- c(negatif, "cant")
score.sentiment <- function(kalimat2, kata.positif, kata.negatif,
.progress='none')
{
 require(plyr)
 require(stringr)
 scores = laply(kalimat2, function(kalimat, kata.positif,
kata.negatif) {
 kalimat = gsub('[[:punct:]]', '', kalimat)
 kalimat = gsub('[[:cntrl:]]', '', kalimat)
 kalimat = gsub('\\d+', '', kalimat)
 kalimat = tolower(kalimat)

list.kata = str_split(kalimat, '\\s+')
 kata2 = unlist(list.kata)
 positif.matches = match(kata2, kata.positif)
 negatif.matches = match(kata2, kata.negatif)
 positif.matches = !is.na(positif.matches)
 negatif.matches = !is.na(negatif.matches)
 score = sum(positif.matches) - (sum(negatif.matches))
 return(score)
 }, kata.positif, kata.negatif, .progress=.progress )
 scores.df = data.frame(score=scores, text=kalimat2)
 return(scores.df)
}
hasil <- score.sentiment(kalimat2$text, kata.positif,
kata.negatif)

#CONVERT SCORE TO SENTIMENT
hasil$klasifikasi<-
ifelse(hasil$score==0,"Netral",ifelse(hasil$score<0,
"Negatif","Positif"))
hasil$klasifikasi
View(hasil)

#EXCHANGE ROW SEQUENCE
data <- hasil[c(3,1,2)]
View(data)
write.csv(data, file = "DETIKlabel.csv")
```

Lampiran 6 Sintaks *R Wordcloud* dan Asosiasi kata

```
# Load packages
library("tm")
library("wordcloud")
library("RColorBrewer")
library(stringr)

setwd("E://SK/TA/DATA/DETIK")
docs<-readLines("ULASAN POSITIF.csv")

# Load the data as a corpus
docs <- Corpus(VectorSource(docs))

# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)

#Replace words
docs <- tm_map(docs, gsub, pattern="towels", replacement="towel")
docs <- tm_map(docs, gsub, pattern="varieties",
replacement="variety")

#Build a term-document matrix
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 25)

#Generate the Word cloud
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=50, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))

#Explore frequent terms and their associations
findFreqTerms(dtm, lowfreq = 4)

#barplot
K <- barplot(d[1:20,]$freq, las = 2, names.arg = d[1:20,]$word,
       cex.axis=1.2,cex.names=1.2, main ="Most frequent words",
       ylab = "Word frequencies",col =topo.colors(20))
K
#asosiasi kata
V <- as.list(findAssocs(dtm, terms
=c("tidak","iklan","berita","aplikasi",
"detik","buka","terbaru","muncul","tolong","baca"),
corlimit = c(0.20,0.20,0.20,0.20,0.20,0.20,0.20,0.20,0.20,0.20)))
V
termFrequency <- rowSums(as.matrix(dtm))
termFrequency <- subset(termFrequency, termFrequency>=5)

text(k,sort(termFrequency, decreasing = T)- 1,
labels=sort(termFrequency, decreasing = T),pch = 6, cex = 1)
```

Lampiran 7 Sintaks R *Machine Learning*

```
 setwd( "E://SK/TA/DATA")
positif = readLines("PL.csv")
negatif = readLines("NL.csv")
positiftes = readLines("PT.csv")
negatiftes = readLines("NT.csv")
length(ulasan_test)

ulasan = c(positif, negatif)
ulasan_test= c(positiftes,negatiftes)
ulasan_all = c(ulasan, ulasan_test)
sentiment = c(rep("positif", length(positif) ),
              rep("negatif", length(negatif)))
length(sentiment)
sentiment_test = c(rep("positif", length(positiftes) ),
                   rep("negatif", length(negatiftes)))
sentiment_all = as.factor(c(sentiment, sentiment_test))
all = data.frame(ulasan_all,sentiment_all)
write.csv(all, file = "sentiment.csv")
length(sentiment_all)
length(all)
summary(all)
head(all)

library(RTextTools)
library(e1071)

mat = create_matrix(ulasan_all, language = "indonesian",
removeStopwords = FALSE,
    removeNumbers = TRUE, stemWords = FALSE, tm::weightTfIdf)

mat = as.matrix(mat)

ln=length(sentiment)
ln1=ln+1
la=length(sentiment_all)

#Maximum Entropy
container <- create_container(mat, sentiment_all,
trainSize=1:ln,testSize=ln1:la, virgin=FALSE)
model <- train_model(container, 'MAXENT',kernel='linear')
results <- classify_model(container, model)
summary(model)
table(as.character(sentiment_all[ln1:la]),
as.character(results[,"MAXENTROPY_LABEL"]))
recall_accuracy(sentiment_all[ln1:la],
results[,"MAXENTROPY_LABEL"])
create_precisionRecallSummary(container, results)
```

Lampiran 8 Hasil klasifikasi pada iterasi dataset

```
> ME_a01 = summary(model)
> ME_b01 = table(as.character(sentiment_all[1n1:1a]), as.character(results[,"MAXENTROPY_LABEL"]))
> ME_c01 = recall_accuracy(sentiment_all[1n1:1a], results[,"MAXENTROPY_LABEL"])
> ME_d01 = create_precisionRecallSummary(container, results)
```

```
> ME_a01 # Summary model
Length  Class   Mode
     1 maxent     S4
> ME_b01# Confusion matrix

          negatif positif
  negatif     346      54
  positif     151    1890
> ME_c01# Akurasi
[1] 0.916018
> ME_d01# Evaluasi: Precision, Recall, F-measure
        MAXENTROPY_PRECISION MAXENTROPY_RECALL MAXENTROPY_FSCORE
negatif                 0.70              0.86               0.77
positif                 0.97              0.93               0.95
```

```
> ME_a02 # Summary model
Length  Class   Mode
     1 maxent     S4
> ME_b02# Confusion matrix

          negatif positif
  negatif     336      64
  positif      42    1999
> ME_c02# Akurasi
[1] 0.9565752
> ME_d02# Evaluasi: Precision, Recall, F-measure
        MAXENTROPY_PRECISION MAXENTROPY_RECALL MAXENTROPY_FSCORE
negatif                 0.89              0.84               0.86
positif                 0.97              0.98               0.97
```

```
> ME_a03 # Summary model
Length  Class   Mode
     1 maxent     S4
> ME_b03# Confusion matrix

          negatif positif
  negatif     339      61
  positif     135    1906
> ME_c03# Akurasi
[1] 0.919705
> ME_d03# Evaluasi: Precision, Recall, F-measure
        MAXENTROPY_PRECISION MAXENTROPY_RECALL MAXENTROPY_FSCORE
negatif                 0.72              0.85               0.78
positif                 0.97              0.93               0.95
```

```
> ME_a04 # Summary model
Length  Class   Mode
     1 maxent     S4
> ME_b04# Confusion matrix

          negatif positif
  negatif     347      53
  positif     178    1863
> ME_c04# Akurasi
[1] 0.9053667
> ME_d04# Evaluasi: Precision, Recall, F-measure
        MAXENTROPY_PRECISION MAXENTROPY_RECALL MAXENTROPY_FSCORE
negatif                 0.66              0.87               0.75
positif                 0.97              0.91               0.94
```

```
> ME_a05 # Summary model
Length  Class   Mode
     1 maxent     S4
> ME_b05# Confusion matrix

          negatif positif
  negatif     328      72
  positif      33    2008
> ME_c05# Akurasi
[1] 0.9569848
> ME_d05# Evaluasi: Precision, Recall, F-measure
        MAXENTROPY_PRECISION MAXENTROPY_RECALL MAXENTROPY_FSCORE
negatif                 0.91              0.82               0.86
positif                 0.97              0.98               0.97
```

Lampiran 9 Hasil klasifikasi pada 5-*cross falidation*

```
> ME_a1 = summary(model)
> ME_b1 = table(as.character(sentiment_all[1n1:1a]), as.character(results[,"MAXENTROPY_LABEL"]))
> ME_c1 = recall_accuracy(sentiment_all[1n1:1a], results[,"MAXENTROPY_LABEL"])
> ME_d1 = create_precisionRecallSummary(container, results)
```

```
> ME_a1 # Summary model
Length  Class   Mode
     1 maxent     S4
> ME_b1 # Confusion matrix

          negatif positif
  negatif     288     112
  positif     126    1915
> ME_c1 # Akurasi
[1] 0.902499
> ME_d1 # Evaluasi: Precision, Recall, F-measure
        MAXENTROPY_PRECISION MAXENTROPY_RECALL MAXENTROPY_FSCORE
negatif                 0.70              0.72               0.71
positif                 0.94              0.94               0.94
```

```
> ME_a2 # Summary model
Length  Class   Mode
     1 maxent     S4
> ME_b2 # Confusion matrix

          negatif positif
  negatif     251     149
  positif      87    1954
> ME_c2 # Akurasi
[1] 0.9033183
> ME_d2# Evaluasi: Precision, Recall, F-measure
        MAXENTROPY_PRECISION MAXENTROPY_RECALL MAXENTROPY_FSCORE
negatif                 0.74              0.63               0.68
positif                 0.93              0.96               0.94
```

```
> ME_a3 # Summary model
Length  Class   Mode
     1 maxent     S4
> ME_b3 # Confusion matrix

          negatif positif
  negatif     251     149
  positif      97    1944
> ME_c3 # Akurasi
[1] 0.8992216
> ME_d3 # Evaluasi: Precision, Recall, F-measure
        MAXENTROPY_PRECISION MAXENTROPY_RECALL MAXENTROPY_FSCORE
negatif                 0.72              0.63               0.67
positif                 0.93              0.95               0.94
```

```
> ME_a4  # Summary model
Length  Class   Mode
     1 maxent     S4
> ME_b4 # Confusion matrix

          negatif positif
  negatif     246     154
  positif      80    1961
> ME_c4 # Akurasi
[1] 0.9041376
> ME_d4 # Evaluasi: Precision, Recall, F-measure
        MAXENTROPY_PRECISION MAXENTROPY_RECALL MAXENTROPY_FSCORE
negatif                 0.75              0.62               0.68
positif                 0.93              0.96               0.94
```

```
> ME_a5  # Summary model
Length  Class   Mode
     1 maxent     S4
> ME_b5 # Confusion matrix

          negatif positif
  negatif     247     153
  positif     107    1934
> ME_c5 # Akurasi
[1] 0.8934863
> ME_d5 # Evaluasi: Precision, Recall, F-measure
        MAXENTROPY_PRECISION MAXENTROPY_RECALL MAXENTROPY_FSCORE
negatif                 0.70              0.62               0.66
positif                 0.93              0.95               0.94
```

Lampiran 10 Rincian hasil iterasi dataset 5 untuk perhitungan model (*models*)

No	Weight	Label	Feature
1	11,3	positif	2445
2	7,32	positif	3445
3	2,58	positif	1878
4	-2,58	negatif	1878
5	11,3	positif	1607
6	0,396	positif	2878
7	-0,396	negatif	2878
8	2,72	positif	2607
9	-2,72	negatif	2607
10	0,0811	positif	3878
11	7,76	positif	3607
12	5,37	positif	1151
13	4,73	negatif	2151
14	2,24	positif	3151
15	0,102	positif	1584
16	7,1	negatif	1313
17	6,23	negatif	2584
18	4,29	negatif	2313
19	8,21	positif	3584
20	5,45	positif	3313
21	4,58	negatif	1746
22	0,351	negatif	2746
23	18,7	negatif	3746
24	0,762	positif	1908
25	7,5	positif	148
26	0,463	negatif	2908
27	0,146	negatif	248
28	7,1	negatif	3908
29	4,65	positif	348
30	-4,65	negatif	348
31	0,529	positif	1
32	4,12	positif	448
33	10,5	positif	2
34	4,94	negatif	548
35	1,53	positif	3
…	…	…	…

No	Weight	Label	Feature
4579	1,64	negatif	3438
4580	0,0559	negatif	1961
4581	21,8	positif	3961
4582	15,6	negatif	1144
4583	0,0446	negatif	2144
4584	10,3	positif	3144
4585	0,0445	positif	1577
4586	2,72	positif	1306
4587	1,9	negatif	2577
4588	-0,269	positif	2306
4589	0,269	negatif	2306
4590	14,1	negatif	3577
4591	4,43	positif	3306
4592	9,77	negatif	1739
4593	4,7	negatif	2739
4594	4,49	positif	104
4595	12,5	positif	3739
4596	36,4	positif	204
4597	0,152	negatif	304
4598	2,16	negatif	404
4599	0,975	positif	504
4600	0,353	negatif	604
4601	0,966	positif	704
4602	-0,966	negatif	704
4603	0,0934	positif	804
4604	0,136	positif	904
4605	11,3	negatif	1283
4606	15,2	negatif	1012
4607	-0,936	positif	2283
4608	0,936	negatif	2283
4609	4,09	positif	2012
4610	-5,82	positif	3283
4611	5,82	negatif	3283
4612	4,87	negatif	3012
4613	9,85	positif	1445

Untuk 35 ulasan pertama dan 35 ulasan terakhir selama tahun 2019

Lampiran 11 Rincian hasil iterasi dataset 5 untuk penentuan label (*results*)

No	Label	Probabilitas
1	positif	0,5000000
2	positif	0,9879513
3	positif	1,0000000
4	positif	0,9999968
5	positif	0,9749938
6	positif	0,5000000
7	positif	0,9818210
8	positif	1,0000000
9	positif	1,0000000
10	positif	0,9987823
11	positif	0,9999459
12	positif	0,8682297
13	positif	1,0000000
14	positif	0,9999931
15	positif	0,9998224
16	positif	0,9879513
17	positif	0,9911267
18	positif	0,9818210
19	positif	0,9525253
20	positif	0,9999267
21	positif	0,9879513
22	positif	0,9879513
23	positif	0,8015761
24	positif	1,0000000
25	positif	0,9973008
26	positif	0,9922233
27	positif	0,9989127
28	positif	0,9999999
29	positif	0,9879513
30	positif	0,9999870
31	positif	0,7551909
32	positif	0,9536798
33	positif	0,8467968
34	positif	0,9997346
35	positif	0,9999657
…	…	…

No	Label	Probabilitas
2407	negatif	0,9999998
2408	negatif	0,9999870
2409	negatif	0,9999998
2410	negatif	0,9999948
2411	negatif	1,0000000
2412	negatif	0,9986007
2413	negatif	1,0000000
2414	negatif	0,9999870
2415	positif	0,6143058
2416	negatif	0,9965419
2417	positif	0,5000000
2418	negatif	0,6879627
2419	negatif	1,0000000
2420	negatif	0,9781052
2421	negatif	1,0000000
2422	negatif	0,9999870
2423	negatif	0,9365051
2424	negatif	1,0000000
2425	negatif	0,9999994
2426	negatif	1,0000000
2427	positif	0,5000000
2428	positif	0,5000000
2429	negatif	0,9844066
2430	negatif	1,0000000
2431	positif	0,8682297
2432	negatif	0,9219111
2433	negatif	0,9999998
2434	negatif	0,5295171
2435	negatif	0,9999434
2436	negatif	1,0000000
2437	negatif	0,9978142
2438	negatif	1,0000000
2439	negatif	0,9999999
2440	negatif	1,0000000
2441	positif	0,9982752

Untuk 35 ulasan pertama dan 35 ulasan terakhir selama tahun 2019