TUGAS AKHIR – KS184822

TEXT CLUSTERING UNTUK PENENTUAN TOPIK BERITA *ONLINE* MENGENAI KOTA SURABAYA DENGAN METODE K-*MEANS* DAN *SELF-ORGANIZING MAPS*

FONDA LEVIANY
NRP 062115 4000 0015

Dosen Pembimbing
Dr. Dra. Kartika Fithriasari, M.Si

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

TUGAS AKHIR – KS184822

TEXT CLUSTERING UNTUK PENENTUAN TOPIK BERITA *ONLINE* MENGENAI KOTA SURABAYA DENGAN METODE K-*MEANS* DAN *SELF-ORGANIZING MAPS*

FONDA LEVIANY
NRP 062115 4000 0015

Dosen Pembimbing
Dr. Dra. Kartika Fithriasari, M.Si

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

(*Halaman ini sengaja dikosongkan*)

FINAL PROJECT – KS184822

TEXT CLUSTERING TO DETERMINE THE ONLINE NEWS TOPICS ABOUT SURABAYA CITY USING K-MEANS AND SELF-ORGANIZING MAPS METHODS

FONDA LEVIANY
SN 062115 4000 0015

Supervisor
Dr. Dra. Kartika Fithriasari, M.Si

UNDERGRADUATE PROGRAMME
DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICS, COMPUTING, AND DATA SCIENCES
INSTITUTE TECHNOLOGY OF SEPULUH NOPEMBER
SURABAYA 2019

(*Halaman ini sengaja dikosongkan*)

LEMBAR PENGESAHAN

TEXT CLUSTERING UNTUK PENENTUAN TOPIK BERITA *ONLINE* MENGENAI KOTA SURABAYA DENGAN METODE K-*MEANS* DAN *SELF-ORGANIZING MAPS*

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

Oleh :
Fonda Leviany
NRP. 062115 4000 0015

Disetujui oleh Pembimbing:
Dr. Dra. Kartika Fithriasari, M.Si ( )
NIP. 19691212 199303 2 002

Mengetahui,
Kepala Departemen Statistika

Dr. Suhartono
NIP. 19710929 199512 1 001

SURABAYA, JULI 2019

(*Halaman ini sengaja dikosongkan*)

TEXT CLUSTERING UNTUK PENENTUAN TOPIK BERITA *ONLINE* MENGENAI KOTA SURABAYA DENGAN METODE K-*MEANS* DAN *SELF-ORGANIZING MAPS*

Nama Mahasiswa : Fonda Leviany
NRP : 062115 4000 0015
Departemen : Statistika
Dosen Pembimbing : Dr. Dra. Kartika Fithriasari, M.Si

Abstrak

Berita memberikan informasi mengenai peristiwa yang terjadi sehingga sampai di telinga masyarakat. Salah satu situs berita online yang memberikan informasi mengenai Kota Surabaya dan sekitarnya adalah SURYA.co.id yang beralamatkan http://surabaya.tribunnews.com/. Situs ini merupakan versi digital dari Koran Harian Surya yang pada Februari 2019 memperoleh penghargaan sebagai Surat Kabar Terbaik Regional Jawa versi IPMA. Berita yang dipublikasikan melalui situs ini diharapkan telah terkategorisasi dengan baik sehingga masyarakat Kota Surabaya dapat memperoleh informasi yang dicari dengan lebih cepat. Namun, fitur kategorisasi berita mengenai Kota Surabaya belum tersedia. Penelitian ini diharapkan mampu memberikan manfaat bagi masyarakat, pemerintah, dan pihak manajemen Tribunnews Surabaya. Korpus berita selama tahun 2018 yang diperoleh akan melewati tahap text preprocessing, tokenizing, feature selection, dan clustering menggunakan K-Means dan Self-Organizing Maps. Tahap tokenizing dilakukan dengan pendekatan unigram dan bigrams. Berdasarkan hasil evaluasi dengan average silhouette width diperoleh hasil bahwa metode K-Means memberikan hasil clustering lebih baik daripada Self-Organizing Maps dengan jumlah cluster optimum sebanyak 10 cluster. Topik berita yang sering dibahas selama tahun 2018 adalah adalah pelecehan seksual, kereta api, Universitas Airlangga, Kepolisian Kabupaten/Kota, Narkoba, RSUD dr. Soetomo, Kepolisian Daerah, Pelabuhan Tanjung Perak, pendidikan, serta hiburan, kriminalitas, peristiwa penting, dan lain-lain.

Kata Kunci**: **Berita Online, K-Means, N-Gram, Self-Organizing Maps, Text Clustering

(*Halaman ini sengaja dikosongkan*)

TEXT CLUSTERING TO DETERMINE THE ONLINE NEWS TOPICS ABOUT SURABAYA CITY USING K-MEANS AND SELF-ORGANIZING MAPS METHODS

Name : **Fonda Leviany**
Student Number : **062115 4000 0015**
Department : **Statistics**
Supervisor : **Dr. Dra. Kartika Fithriasari, M.Si**

Abstract

News provides information to the people about events that occured. One of the online news sites that provide information about the Surabaya City and around is SURYA.co.id which has the complete URL http://surabaya. tribunnews.com/. This site is the digital version of Surya Daily Newspaper which got the achievement on February 2019 as The Best Newspaper in Java Region by IPMA. The news published through this site is expected to have been well categorized so that the citizen of Surabaya City can obtain information faster. However, the news categorization feature regarding the Surabaya City is not yet available. This research is expected to be able to provide benefits to the citizens, government, and the management of the Surabaya Tribunnews. The news corpus during 2018 year, will be processed which start from text pre-processing, feature selection, and text clustering using K-Means and Self-Organizing Maps. The tokenizing phase is carried out with unigram and bigrams approach. Based on the evaluation results by average silhouette width, the K-Means method gives better clustering results than Self-Organizing Maps with the optimum number of clusters is 10 clusters. The news topics that are often discussed during 2018 are sexsual harassment,, trains, Airlangga University, city police, drugs, RSUD dr. Soetomo, regional police, Tanjung Perak Harbour, education, and also about entertainment, crimes, breaking news, and etc.

Keywords *:* ***K-Means, Online News, N-Gram, Self-Organizing Maps, Text Clustering***

(*Halaman ini sengaja dikosongkan*)

KATA PENGANTAR

Puji syukur atas kehadirat Allah SWT yang telah memberikan rahmat, taufik, dan hidayah-Nya sehingga atas izin-Nya penulis dapat menyelesaikan laporan Tugas Akhir yang berjudul "***Text Clustering* untuk Penentuan Topik Berita *Online* Mengenai Kota Surabaya dengan Metode K-*Means* dan *Self-Organizing Maps***".

Penyusunan laporan Tugas Akhir ini dapat terselesaikan dengan baik dan lancar karena tidak lepas dari bantuan, bimbingan, serta dukungan dari berbagai pihak. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada :

1. Ayah, Ibu, Adik, dan semua keluarga besar atas doa, motivasi, dukungan, dan fasilitas yang telah diberikan kepada penulis sehingga diberi kelancaran selama masa perkuliahan hingga menyelesaikan Tugas Akhir.
2. Ibu Dr. Dra. Kartika Fithriasari, M.Si selaku dosen pembimbing yang telah sabar membimbing, mengarahkan, memberikan masukan, dan dukungan bagi penulis untuk dapat menyelesaikan Tugas Akhir ini.
3. Ibu Dr. Irhamah, M.Si dan Ibu Pratnya Paramitha Oktaviana, M.Si selaku dosen penguji yang telah memberikan masukan dan saran demi kesempurnaan Tugas Akhir ini.
4. Bapak Dr. Suhartono, M.Sc. selaku Kepala Departemen Statistika ITS dan Ibu Dr. Santi Wulan Purnami, S.Si., M.Si. selaku Ketua Program Studi Sarjana Statistika ITS yang telah memberikan nasihat dan bimbingan kepada penulis untuk menyelesaikan Tugas Akhir ini.
5. Ibu Dr. Dra. Ismaini Zain, M.Si. selaku dosen wali yang telah sabar dalam memberikan bimbingan, nasihat, arahan, dan motivasi bagi penulis selama menempuh pendidikan sarjana di Departemen Statistika ITS.
6. Muhammad Adhitya Muslim selaku partner organisasi sekaligus partner 24/7 yang selalu memberikan bantuan, semangat, dan dukungan bagi penulis untuk menyelesaikan laporan Tugas Akhir ini. Dan juga Qathrunnada, Bina Astri

Sitoresmi, Deliar Mahardika Candra, Zeffri Irawan, Daud Muhajir, beserta seluruh teman-teman dan karyawan Kopma dr. Angka ITS yang tidak bisa penulis sebutkan namanya satu per satu yang telah memberikan pengalaman dan dukungan bagi penulis selama menjalani perkuliahan di ITS.

7. Lianna Dwi Rahmawati, Risda Ikfina Putri, Taufiqotul Masrukha Tesha Nisva, Shindy Sari Utami, Icha Tirhiss Febriana, Dian Vitiana Ningrum, Waode Melvy Agrina Jalil Silea, Hikmatul Munawaroh, Dewi Muslimatul Azizah, Rahayu Prihatini Saputri, Farizah Rizka Rahmaniar, Cahya Buana Putri, I Gusti Putu Surya Darma, Narendra Saguna, dan teman-teman dari Departemen Statistika ITS angkatan 2015 yang telah memberikan perhatian, kasih sayang, dan dukungan baik urusan akademik dan non akademik selama menjalani perkuliahan di ITS.
8. Seluruh dosen dan karyawan Departemen Statistika ITS atas ilmu dan pengalaman yang dibagikan kepada penulis.

Penulis menyadari bahwa laporan Tugas Akhir ini masih jauh dari kata sempurna, oleh karena itu penulis sangat mengharapkan kritik dan saran yang membangun agar berguna untuk perbaikan berikutnya. Semoga laporan Tugas Akhir ini dapat memberikan manfaat.

Surabaya, Juli 2019

Penulis

DAFTAR ISI

Halaman

Halaman

DAFTAR TABEL

Halaman

Halaman

DAFTAR GAMBAR

Halaman

(*Halaman ini sengaja dikosongkan*)

DAFTAR LAMPIRAN

Halaman

(*Halaman ini sengaja dikosongkan*)

BAB I
PENDAHULUAN

1.1 Latar Belakang

Berita adalah laporan tercepat mengenai fakta atau ide terbaru yang benar, menarik, dan penting bagi sebagian besar khalayak, melalui media berkala seperti surat kabar, radio, televisi, atau media *online* internet (Sumadiria, 2011). Melalui berita yang disebarkan, informasi mengenai peristiwa yang terjadi dapat sampai di telinga masyarakat. Berdasarkan hasil survei Nielsen Consumer & Media View (CMV) kuartal III 2017, diungkapkan bahwa media cetak (termasuk koran, majalah, dan tabloid) memiliki penetrasi sebesar 7% dan dibaca oleh 4,5 juta orang. Dari jumlah tersebut, 83% di antaranya membaca koran. Alasan utama para pembaca masih memilih koran adalah karena nilai beritanya dapat dipercaya. Jika dilihat dari profil pembaca, media cetak di Indonesia cenderung dikonsumsi oleh konsumen dari rentang usia 20-49 tahun (74%), memilki pekerjaan sebagai karyawan (32%) dan mayoritas pembacanya berasal dari kelas atas (54%). Ini menunjukkan bahwa pembaca media cetak masih produktif dan dari kalangan yang mapan (Lubis, 2017).

Seiring dengan kemajuan teknologi informasi dan komunikasi, media yang digunakan oleh masyarakat dalam mengakses berita mengalami pergesaran. Media konvensional yang dulu digunakan untuk mengakses berita dalam rangka memperoleh informasi mengenai peristiwa yang telah terjadi seperti media cetak, berangsur-angsur ditinggalkan dan beralih menggunakan media elektronik dan media *online*. Hal ini dikarenakan karakteristik media elektronik dan media *online* yang dapat memberikan informasi lebih *up to date* daripada media cetak. Survei Nielsen Consumer & Media View hingga triwulan ketiga 2017 menyatakan, kebiasaan membaca orang Indonesia telah mengalami pergeseran. Pada 2017, penetrasi pasar pengguna internet sebesar 44% sedangkan penetrasi pasar pengguna koran sebesar 7%. Hal itu menunjukkan bahwa kepembacaan melalui internet atau digital lebih tinggi jika dibandingkan media cetak yakni mencapai 6 juta orang. Selain itu,

hasil survei tersebut juga memberikan informasi bahwa tingkat pembelian koran secara personal hanya sebesar 20%, menurun dibandingkan 2013 yang mencapai 28%. Menariknya lagi, berita dalam versi digital mampu menjangkau pembaca dari Generasi Z dengan rentang usia 10-19 tahun (17%). Para Generasi Z diprediksi menjadi konsumen media masa depan. Hal ini juga didukung dari data APJII menunjukkan bahwa pada tahun 2017, pengguna internet di Indonesia sebanyak 143,26 juta jiwa atau setara dengan 54,7 persen dari total populasi masyarakat Indonesia (Asosiasi Penyelenggara Jasa Internet Indonesia, 2018). Hal ini menempatkan Indonesia sebagai negara dengan peringkat ke-5 di dunia dalam hal penggunaan Internet. Hal ini mendukung pula perilaku masyarakat Indonesia yang menggunakan internet untuk *browsing* berita (Internet World Statistics, 2017).

Secara umum, media *online* adalah media yang tersaji secara *online* di internet. Dengan media *online*, berita yang memuat informasi penting dan terkini dapat dilaporkan dan disebarkan secara cepat sehingga masyarakat tidak tertinggal informasi mengenai kejadian atau peristwa yang telah terjadi. Berita yang disajikan pun tidak hanya sebatas tulisan, namun dapat berupa audio, video, dan gambar yang memiliki kualitas di atas media cetak. Selain itu, dengan adanya sistem teknologi yang semakin maju, artikel berita *online* yang telah diterbitkan secara otomatis akan terarsip secara digital. Sehingga, pembaca dapat membuka kembali artikel berita di masa lampau sewaktu-waktu. Kelebihan media *online* lainnya sebagai media publikasi berita adalah tidak ada batasan halaman atau waktu seperti di media cetak dan media elektronik sehingga masyarakat dapat memperoleh informasi selengkap-lengkapnya. Beberapa situs berita *online* yang sering dikunjungi oleh masyarakat di antaranya adalah *detik.com*, *kompas.com*, *tribunnews.com*, dan lain-lain.

Berita yang dipublikasikan melalui media *online* diharapkan telah terkategorisasi dengan baik. Sehingga, masyarakat dapat mengakses berita kapan pun melalui berdasarkan topik informasi yang dicari. Akan tetapi, belum semua situs berita *online* telah

memiliki fitur pelabelan topik secara spesifik. Contoh situs berita *online* yang memberikan informasi mengenai Kota Surabaya dan sekitarnya adalah SURYA.co.id yang beralamatkan *http://surabaya.tribunnews.com/*. Situs ini merupakan versi digital dari Koran Harian Surya yang pada Februari 2019 memperoleh penghargaan sebagai Surat Kabar Terbaik Regional Jawa versi IPMA (Koloway, 2019). Pada situs berita *online* ini filter berita yang digunakan masih bersifat umum untuk memberitakan kejadian berdasarkan wilayahnya seperti "Berita Surabaya", "Berita Sidoarjo", "Berita Gresik", dan lain-lain. Filter "Berita Surabaya" merupakan fitur pada situs berita *online* tersebut untuk menampilkan berita mengenai Kota Surabaya. Pada filter ini, belum ada pengelompokkan topik yang lebih spesifik lagi untuk mengetahui apakah artikel tersebut membahas topik politik, kriminalitas, pendidikan, sosial, dan lain-lain. Hal ini dapat menghambat pembaca untuk mencari topik berita yang dicarinya khususnya artikel berita mengenai Kota Surabaya. Selain itu, dengan adanya pengelompokkan berita secara lebih spesifik diharapkan mampu memberikan informasi bagi pemerintah mengenai kejadian atau peristiwa yang terjadi di wilayahnya. Sehinga, berdasarkan himpunan data artikel berita *online* yang telah dikelompokkan menurut topik tertentu dapat membantu pemerintah dalam mengambil kebijakan dan keputusan terkait isu topik yang sering diberitakan. Oleh karena itu, diperlukan suatu metode statistika untuk menjawab permasalahan ini yaitu *text clustering*.

Text clustering adalah proses *unsupervised learning* (proses pembelajaran sendiri) yang mengelompokkan kumpulan dokumen berdasarkan hubungan kemiripannya dan memisahkan ke dalam beberapa kelompok. Beberapa metode *text clustering* diantaranya adalah K-*Means*, K-*Medoids*, *Single Linkage Method*, DBSCAN, OPTIC, *Self-Organizinf Maps*, dan lain-lain. Tahap pertama dalam melakukan *text clustering* adalah melakukan *pre-processing* data yang dimulai dari tahap *case folding*, *stemming*, *stopwords*, *tokenizing*, dan pembobotan dengan menggunaka TF-IDF. Dengan adanya pengelompokkan ini, diharapkan dapat membantu pembaca

untuk menemukan artikel berita yang ingin dicari. Selain itu, dengan visualisasi *word cloud* dapat membantu pemerintah untuk mengetahui isu permasalahan Kota Surabaya berdasarkan apa yang dijadikan bahan pemberitaan bagi masyarakat Surabaya.

Penelitian terdahulu mengenai pengelompokkan data atau *clustering* berdasarkan deret waktu pernah dilakukan oleh Kartika F dkk yang berjudul "*Clustering Stationary and Non-Stationary Time Series Based on Autocorrelation Distance of Hierarchical and* K-*Means Algorithms*" memberikan hasil bahwa metode K-*Means* memberikan performansi yang paling baik dalam mengelompokkan data deret waktu stasioner dan non-stasioner. Selain itu, penelitian sebelumnya yang berjudul "*Clustering* Berita Berbahasa Indonesia" pada tahun 2008 yang dilakukan oleh Yudi Wibisono dan Masayu Leyla memberikan hasil bahwa penggunaan log TF-IDF tanpa *stemming* menghasilkan kualitas *cluster* terbaik. Namun, dapat dikatakan bahwa kualitas hasil *cluster* yang dihasilkan masih rendah. Hasil penelitian tersebut sejalan dengan hasil penelitian oleh Fadillah Z Tala di tahun 2003 yang berjudul "*A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*" yang juga memberikan hasil bahwa proses *stemming* kurang optimal apabila digunakan untuk tujuan penemuan kembali informasi (*information retrieval*) dari sekumpulan dokumen berita. Selain itu, pada tahun 2010 Diah Pudi Langgeni dkk melakukan penelitian yang berjudul "*Clustering* Artikel Berita Berbahasa Indonesia menggunakan *Unsupervised Feature Selection*" memberikan hasil bahwa *unsupervised feature selection* dapat memperbaiki performansi hasil *clustering*. Penelitian mengenai analisis *clustering* pada dokumen berita juga pernah dilakukan pada tahun 2014 oleh Ambarwati dan Edi Winarko dengan judul "Pengelompokan Berita Indonesia berdasarkan Histogram Kata Menggunakan *Self-Organizing Map*" memberikan hasil bahwa sistem yang dirancang mampu menampilkan visualisasi dengan *smoothed data histograms* berupa *island map* dari artikel berita majalah Tempo yang diproses. Namun, hasil dari penelitian tersebut belum dilakukan suatu evaluasi untuk mengetahui performansi

dari hasil *cluster* yang terbentuk serta pada tahap *tokenizing* tidak memperhatikan bentuk idiom atau frasa. Algoritma K-*Means* yang sederhana terbukti memiliki performansi lebih baik daripada algoritma *Self-Organizing Maps* yang diperoleh dari hasil penelitian tugas akhir dengan judul "Pengelompokan Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Kesehatan Masyarakat Menggunakan Metode Kohonen *SOM* dan K-*Means*" oleh Marina Marsudi Putri dan Kartika Fithriasari di tahun 2015 memberikan kesimpulan bahwa nilai *icdrate* metode Kohonen SOM yaitu sebesar 0,962. Nilai ini lebih kecil dibandingkan dengan nilai *icdrate* hasil pengelompokan metode K-*Means* yaitu sebesar 0,988. Oleh karena itu, pada Tugas Akhir ini akan dilakukan penelitian mengenai *clustering* pada data tekstual dengan metode K-*Means* dan *Self-Organizing Map* untuk mengetahui perbandingan hasil performansi kedua metode jika diterapkan pada data tekstual.

K-*Means* banyak digunakan karena metode ini sederhana dan dapat digunakan untuk berbagai tipe data serta cukup efisien. Sementara itu, *Self-Organizing Maps* digunakan karena algoritma ini merupakan salah satu jenis model *neural network* yang melakukan *unsupervised learning* dan mampu memetakan data dari berdimensi tinggi ke berdimensi rendah. Keunggulan lain dari metode *Self-Organizing Maps* adalah dapat mengelompokkan data yang mengandung *overlapping* dan mengatasi sifat non linier pada klaster yang terbentuk. Oleh karena itu, pada Tugas Akhir ini akan dilakukan penelitian mengenai pembentukan klaster untuk penentuan topik berita *online* pada *http://surabaya.tribunnews.com/*. Dari kumpulan artikel berita *online* mengenai Kota Surabaya tahun 2018, selanjutnya akan dilakukan *pre-processing* dimana pada tahap *tokenizing* akan menggunakan pendekatan *unigram* dan *bigrams*. Untuk memperoleh kata kunci pada masing-masing topik akan dilakukan *feature selection* berdasarkan *term* dan *document frecuency*. Selanjunya, akan dilakukan pembentukan *cluster* dengan metode K-*Means* dan *Self-Organizing Maps*, dimana parameter yang digunakan untuk membandingkan kinerja kedua metode tersebut menggunakan nilai *silhouette coefficient*. Pada

masing-masing *cluster* yang terbentuk akan digali informasi lebih lanjut (*information retrieval*) dengan bantuan visualisasi *wordcloud*. Dari hasil penelitian ini, diharapkan dapat memberikan manfaat bagi masyarakat dalam mencari informasi yang dibutuhkan secara cepat dan tepat; membantu pemerintah Kota Surabaya dalam merumuskan kebijakannya; serta membantu pihak manajemen *http://surabaya.tribunnews.com* dalam mengelompokkan dokumen artikel berita yang diterbitkan.

1.2 Rumusan Masalah

Belum adanya pengelompokkan topik berita *online* secara spesifik pada situs berita *http://surabaya.tribunnews.com/* sehingga diperlukan suatu metode statistika untuk menyelesaikan permasalahan tersebut. Metode statistika yang digunakan dalam penelitian ini adalah *text clustering* dengan K-*Means* dan *Self-Organizing Maps*. Selanjutnya, performansi hasil *clustering* dengan kedua metode tersebut akan dibandingkan berdasarkan nilai *average silhouette width* tertinggi untuk menentukan metode *text clustering* yang optimal dalam menentukan topik berita *online*. Selain itu, diperlukan suatu teknik untuk menggali informasi lebih dalam lagi mengenai berita *online* yang diterbitkan seperti berita apa yang sedang hangat dibicarakan oleh masyarakat sehingga dapat menjadi informasi tambahan bagi Pemerintah Kota Surabaya dalam mengambil kebijakan publik melalui visualisasi *word cloud*.

1.3 Tujuan

Berdasarkan rumusan masalah yang telah dipaparkan sebelumnya, tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut.

1. Mengetahui karakteristik artikel berita *online* dan *text preprocessing* pada situs *http://surabaya.tribunnews.com/* mengenai Kota Surabaya.
2. Mengetahui hasil *text clustering* dengan metode K-*Means* dan *Self-Organizing Maps* dengan pendekatan *tokenizing* menggunakan *unigram* dan *bigrams*.

3. Menarik informasi kembali berdasarkan karakteristik berita pada masing-masing *cluster* dengan menggunakan visualisasi *word cloud*.

1.4 Manfaat

Manfaat yang diharapkan dari penelitian ini adalah.

1. Membantu masyarakat dalam menemukan informasi yang cepat sesuai dengan topik label berita yang ingin dicari.
2. Membantu pihak manajemen Tribunnews Surabaya dalam membuat topik artikel berita *online*.
3. Memberikan informasi tambahan kepada Pemerintah Kota Surabaya mengenai isu yang beredar di masyarakat berdasarkan hasil *clustering* data berita *online* yang dipublikasikan melalui *http://surabaya.tribunnews.com/*.

1.5 Batasan Masalah

Penelitian ini menggunakan batasan masalah yaitu data bersumber dari artikel berita *online* mengenai Kota Surabaya yang yang dipublikasikan melalui situs http://surabaya.tribunnews.com/ pada 1 Januari 2018 - 31 Desember 2018. Selain itu, penggunaan metode N-*Gram* dalam penelitian ini menggunakan basis teks dengan membentuk *unigram* dan *bigrams*. Untuk jumlah klaster yang dibentuk dalam penelitian ini berukuran 2 hingga 10 klaster serta jenis topologi yang digunakan pada metode SOM adalah "*hexagonal*".

(*Halaman ini sengaja dikosongkan*)

BAB II
TINJAUAN PUSTAKA

2.1 *Data Mining*

Data mining didefinisikan sebagai proses komputasi untuk menganalisis data dalam jumlah besar dengan mengekstrak pola dan informasi yang berguna (Gullo, 2015). Dalam beberapa dekade terakhir, *data mining* telah banyak mendapat sebutan lain seperti *knowledge discovery*, *business intelligence*, *predictive modelling*, *predictive analytics*, dan beberapa lainnya (Linoff & Berry, 2011). Tetapi, tidak sedikit orang yang mendefinisikan *data mining* sebagai sinonim dari istilah populer lainnya yaitu *knowledge discovery from data* (KDD) dan yang lain melihat *data mining* hanya sebagai salah satu tahapan dari *knowledge discovery.*

Pentingnya *data mining* saat ini terutama didorong oleh banyaknya data yang dikumpulkan dan disimpan dengan berbagai aplikasi terkemuka terkini, seperti data web, data *e-commerce*, data pembelian, transaksi bank, dan sebagainya. Data yang dihasilkan oleh aplikasi-aplikasi tersebut umumnya merupakan jenis *big data* dimana data tersebut sulit diolah atau dimengerti secara sederhana. *Big Data* merupakan data yang mempunyai tiga karakteristik yaitu jumlah (*volume*) dan variasi (*variety*) besar, serta bergerak cepat (*velocity*), sehingga melampaui kapasitas pengolahan database konvensional (Dumbill, 2014). Hingga saat ini, *data mining* telah banyak diakui sebagai suatu alat analisis data serbaguna yang bisa diaplikasikan untuk menganalisis *big data* dalam berbagai bidang, tidak hanya dalam bidang teknologi informasi tetapi juga dalam dunia pengobatan klinis, sosiologi, fisika, dan banyak lainnya.

Penggunaan *data mining* dibedakan menjadi dua jenis fungsi yaitu prediktif dan deskriptif (Gullo, 2015). Penggalian prediktif mengacu pada pembangunan model yang berguna untuk memprediksi perilaku atau nilai-nilai di masa depan. Tugas deskriptif meliputi klasifikasi dan prediksi, tugas yang dilakukan seperti membangun beberapa model (atau fungsi) yang menggambarkan kelas atau konsep data oleh satu set objek data yang label kelasnya

diketahui (*training set*), sehingga dapat memprediksi kelas yang labelnya tidak diketahui; deteksi penyimpangan, yaitu berurusan dengan penyimpangan data, yang didefinisikan sebagai perbedaan antara nilai yang terukur dan nilai referensi; analisis evolusi, yaitu, mendeteksi dan menggambarkan pola yang teratur dalam data yang perilakunya berubah dari waktu ke waktu. Sedangkan tujuan penggalian deskriptif yaitu membangun model untuk mendeskripsikan data menjadi bentuk yang mudah dimengerti, efektif, dan efisien. Contoh dari tugas deskriptif di antaranya adalah karakterisasi data, yang tujuan utamanya adalah untuk meringkas karakteristik umum atau fitur dari kelas target data; *association rule*, yaitu, menemukan aturan yang menunjukkan kondisi atribut-nilai yang sering muncul bersama-sama dalam himpunan data; dan *clustering*, yang bertujuan untuk membentuk kelompok yang memiliki kohesif tinggi dan terpisahkan dengan baik dari satu set objek data.

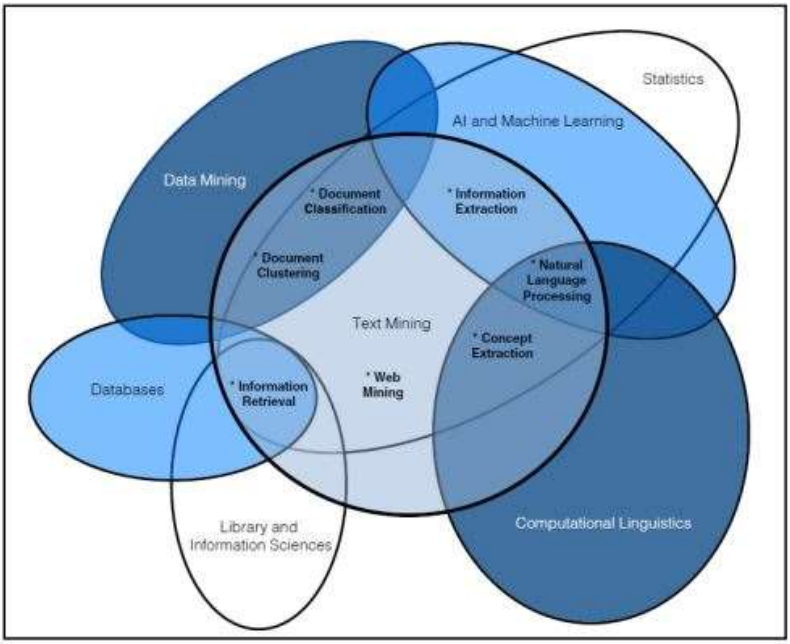
2.2 *Text Mining*

Text mining atau *text analytics* adalah istilah yang mendeskripsikan sebuah teknologi yang mampu menganalisis data teks semi-terstruktur maupun tidak terstruktur. Hal inilah yang membedakannya dengan *data mining* dimana *data mining* mengolah data yang sifatnya terstruktur. Pada dasarnya, *text mining* merupakan bidang interdisiplin yang mengacu pada perolehan informasi *(information retrieval*), *data mining*, pembelajaran mesin (*machine learning*), statistik, dan komputasi linguistik (Jiawei, Kamber, & Pei, 2012). Saat ini teknik *text mining* secara berkelanjutan diaplikasikan dalam dunia industry, akademik, aplikasi web, internet, dan berbagai bidang lainnya.

Text mining mengekstrak indeks numerik yang bermakna dari teks dan kemudian informasi yang terkandung dalam teks akan diakses dengan menggunakan berbagai algoritma *data mining* (statistik dan *machine learning*) (Miner dkk, 2012). *Text mining* dapat menganalisis dokumen, mengelompokkan dokumen berdasarkan kata-kata yang terkandung di dalamnya, serta menentukan kesamaan di antara dokumen untuk mengetahui bagaimana

mereka berhubungan dengan variabel lainnya (Statsoft, 2015). Aplikasi yang paling umum dilakukan *text mining* saat ini misalnya penyaringan spam, analisis sentimen, mengukur preferensi pelanggan, meringkas dokumen, pengelompokan topik penelitian, dan banyak lainnya. Menurut Miner dkk (2012), pekerjaan *text mining* dikelompokkan menjadi 7 daerah praktek yang diilustrasikan seperti Gambar 2.1.

- Pencarian dan perolehan informasi (*search and information retrieval*), yaitu penyimpanan dan penggalian dokumen teks misalnya dalam mesin pencarian (*search engine*) dan pencarian kata kunci (*keywords*)
- Pengelompokan dokumen, yaitu pengelompokan dan pengkategorian kata, istilah, paragraf, atau dokumen dengan menggunakan metode klaster (*clustering*) *data mining*.
- Klasifikasi dokumen, yaitu pengelompokan dan pengkategorian kata, istilah, paragraf, atau dokumen dengan menggunkan metode klasifikasi (*classification*) *data mining* berdasarkan model terlatih yang sudah memiliki label.
- *Web mining*, yaitu penggalian informasi dari internet dengan skala fokus yang spesifik.
- Ekstraksi informasi (*information extraction*), yaitu mengidentifikasi dan mengekstraksi informasi dari data yang sifatnya semi-terstruktur atau tidak terstruktur dan mengubahnya menjadi data yang terstruktur.
- *Natural Language Processing* (NLP), yaitu pembuatan program yang memiliki kemampuan untuk memahami bahasa manusia.
- Ekstraksi konsep, yaitu pengelompokan kata atau frase ke dalam kelompok yang mirip secara semantik.



Gambar 2.1 Diagram Venn 6 Bidang Terkait dan 7 Area Praktik *Text Mining* (Sumber: Miner dkk, 2012)

2.3 Tahapan *Text Mining*

Beberapa tahapan proses yang harus dilakukan untuk mencapai tujuan dari *text mining* ditunjukkan pada sub bab berikut. Data terpilih yang akan dianalisis pertama kali melewati tahap praproses dan representasi teks, hingga akhirnya dapat dilakukan *knowledge discovery* menggunakan *text clustering* yang dilanjutkan dengan tahap akhir yaitu melakukan evaluasi terhdap pembentukan *cluster*. Pada bagian ini akan dijelaskan proses *text mining* dengan data yang bersumber dari situs berita *online*.

2.3.1 *Web Scraping*

Web scraping adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa laman web yang dibangun dengan bahasa markup seperti HTML atau XHTML yang bertujuan untuk mengambil informasi dari halaman tersebut baik secara keseluruhan atau sebagian untuk digunakan bagi kepentingan lain (Johnson & Gupta, 2012). Secara umum, ada empat tahapan dalam penggunaan *web scraping* untuk mengambil data secara otomatis dari sebuah laman *web* sebagai berikut:

1. Mempelajari dokumen HTML dari *website* yang akan diambil informasinya untuk tag HTML yang mengapit informasi yang akan diambil.
2. Menelusuri mekanisme navigasi pada *website* yang akan diambil informasinya untuk ditirukan pada program *web scraper* yang akan dibuat.
3. Berdasarkan informasi yang didapat pada langkah 1 dan 2 di atas, program *web scraper* dibuat untuk mengotomatisasi pengambilan informasi dari *website* yang ditentukan.
4. Informasi yang didapat dari langkah 3 disimpan dalam format data tertentu.

Dalam penelitian ini *web scraping* digunakan untuk mengambil data dari sebuah laman situs berita *online* kemudian melakukan transformasi dari bentuk yang tidak terstruktur, umumnya dalam format HTML menjadi suatu format data terstruktur yang dapat disimpan ke dalam database untuk keperluan repositori maupun analisis lebih lanjut.

2.3.2 *Text Pre-processing*

Setiap langkah dalam tahap praproses teks memegang peranan yang sangat penting dalam aplikasi dan teknik pada *text mining*. Pra proses teks terdiri dari tahapan-tahapan yang dilakukan sebelum mengolah data teks. Pra proses ini perlu dilakukan karena data teks mentah yang diperoleh biasanya merupakan data yang belum terstruktur dan belum dapat dilakukan proses *text mining*. Selain itu, dengan melakukan pra proses teks memberikan keun-

tungan karena dapat mempercepat proses dalam melakukan pengolahan data. Adapun tahapan-tahapan praproses teks yang diterapkan dalam penelitian ini adalah sebagai berikut yang dirangkum dari situs resmi *software* R (CRAN, 2018).

1. *Case Folding*

Case folding adalah tahap pra proses yang mengubah semua huruf dalam dokumen teks menjadi huruf non kapital. Tahap ini bertujuan untuk menyamakan *case* dalam artikel.

2. *Remove Punctuation*

Proses menghapus karakter berupa tanda baca pada dokumen teks yang tidak diperlukan seperti tanda baca koma (,), titik (.), apostrof (‘), tanda hubung (-), *hashtag* (#), dan lain-lain.

3. *Remove Number*

Tahap ini menghapus karakter berupa angka yang ditemukan pada keseluruhan dokumen teks.

4. *Remove Stopwords*

Stopwords adalah kata – kata yang sering kali muncul dalam dokumen namun artinya tidak deskriptif dan tidak memiliki keterkaitan dengan tema tertentu. Di dalam bahasa Indonesia, *stopwords* dapat disebut sebagai kata hubung yang tidak penting, seperti “di”, “oleh”, “pada”, “sebuah”, “karena” dan lain sebagainya.

5. *Strip White Space*

Proses penghapusan tanda baca, angka, dan *stopwords*, meninggalkan spasi berlebih pada dokumen teks. Selain itu masih ditemukan karakter HTML pada dokumen teks. Spasi berlebih dan karakter HTML yang ditemukan dapat menimbulkan *noise* sehingga perlu dihapus.

6. *Tokenizing*

Tokenizing adalah proses pemotongan string input berdasarkan tiap kata yang menyusunnya. Pemecahan kalimat menjadi kata-kata tunggal dilakukan dengan men-*scan* kalimat dengan pemisah (*delimiter*) dan *white space* (spasi, tab, dan *newline*). Teknik *tokenizing* yang dilakukan dalam penelitian ini menggunakan pendekatan statistika yaitu dengan N-*Gram* yang akan dibahas lebih lanjut pada sub bab 2.6.

2.4 *Feature Selection*

Dalam *machine learning* dan statistika, *feature selection* adalah suatu proses untuk memilih sekelompok *feature* yang dianggap relevan untuk membangun model. Salah satu teknik dalam *feature selection* di antaranya adalah dengan menggunakan pendekatan *filter*. Pendekatan *filter* menggabungkan ukuran independen untuk mengevaluasi subset *feature* tanpa melibatkan algoritma *training*. Pendekatan ini efisien dan cepat dalam melakukan proses komputasi. (Kumar & Minz, 2014)

Beberapa metode *feature selection* yang ada di antaranya berdasarkan *term frequency* dan juga *document frequency*. *Term frequency* adalah berapa kali kata tertentu muncul dalam sebuah dokumen sedangkan *document frequency* adalah jumlah dokumen yang memuat kata tersebut. (Azam & Yao, 2012)

2.5 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) adalah statistik numerik yang memberikan informasi mengenai seberapa penting suatu kata dalam kumpulan dokumen teks. TF-IDF sering digunakan sebagai faktor bobot dalam pencarian informasi (*information retrieval*) dan *text mining*. Nilai TF-IDF meningkat secara proporsional dengan berapa kali *i* kata yang muncul pada artikel ke-*j*, tetapi berbanding terbalik dengan frekuensi kemunculan kata *i* pada dokumen teks (Kumar & Minz, 2014). Perhitungan pembobotan TF-IDF yang digunakan untuk mengubah *text* menjadi numerik dihasilkan dengan rumus sebagai berikut.

$$w_{ij} = tf_{ij} \text{ x } idf \text{ dengan } idf = \log\left(\frac{D}{df_j}\right) \qquad (2.1)$$

Dimana,

w_{ij} : bobot dari kata *i* pada artikel ke-*j*
tf_{ij} : jumlah kemunculan kata *i* pada artikel ke-*j*
idf : *inverse document frequency*
D : jumlah seluruh dokumen
df_j : jumlah artikel *j* yang mengandung kata *i*

2.6 N-*Gram*

N-*Gram* adalah potongan sejumlah *n* karakter dalam suatu *string* tertentu atau potongan *n* kata dalam suatu kalimat tertentu (Cavnar & Trenkle, 1994). N-*Gram* merupakan sebuah metode yang diaplikasikan untuk pembangkitan kata atau karakter. Dalam penelitian ini, N-*Gram* akan dilakukan sebagai bagian dari proses *tokenizing* yang memotong *n* kata dalam suatu kalimat. Sebagai contoh: kalimat "fenomena supermoon waspadai banjir rob" dapat diuraikan ke dalam beberapa tipe N-*Gram* berbasis kata sebagai berikut.

Tabel 2.1 Contoh Pemotongan N-*Gram* Berbasis Kata

Nama	N-*Gram* Kata
Unigram	fenomena supermoon waspadai banjir rob
Bigrams	fenomena supermoon, supermoon waspadai, waspadai banjir, banjir rob
dst	

Salah-satu keunggulan menggunakan N-*Gram* dan bukan suatu kata utuh secara keseluruhan adalah bahwa N-*Gram* tidak akan terlalu sensitif terhadap kesalahan penulisan yang terdapat pada suatu dokumen (Furnkranz, 1994). Selain itu, metode N-*Gram* memiliki keunggulan lain yaitu dapat berfungsi dengan baik walaupun terdapat kesalahan tekstual serta dapat berjalan secara efisien, membutuhkan penyimpanan yang sederhana dan waktu proses yang cepat. Model N-*Gram* dokumen dibentuk berdasarkan frekuensi N-*Gram* yang muncul di dalam dokumen. Dokumen akan dibaca kata per kata, dan untuk setiap kata akan dibuat N-*Gram* dari kata tersebut. Untuk setiap N-*Gram* yang dibangkitkan, akan dicatat dalam sebuah tabel dengan N-*Gram* sebagai kunci dan jumlah sebagai isi. Apabila N-*Gram* tersebut sudah pernah muncul di dalam dokumen maka frekuensi untuk N-*Gram* itu akan ditambah satu, jika belum maka N-*Gram* tersebut akan ditambahkan ke dalam tabel dengan jumlah kemunculan satu.

2.7 *Text Clustering*

Proses *clustering* menggunakan skema *unsupervised learning* dimana pengelompokan data akan dilakukan tanpa model latihan (*learning model*). Pada konten *text mining*, dokumen-dokumen akan dikelompokkan dalam berbagai klaster berdasarkan konten isi dari dokumen (Suh, Park, & Jeon, 2010). Jenis metode *clustering* yang paling umum digunakan yaitu *partitioning-based* (partisi), *hierarchical-based* (hirarki), dan *kohonen neural network* atau yang sering disebut sebagai *Self-Organizing Map* (SOM). Pada penelitian kali ini, akan digunakan dua jenis algoritma *clustering* yaitu K-*Means clustering* yang termasuk dalam pendekatan *partitioning-based* (partisi) dan *Self-Organizing Map* (SOM).

2.7.1 K-*Means*

K-*Means* merupakan salah satu metode data *clustering* yang berusaha mempartisi *N* jumlah data ke dalam *k* jumlah kelompok/klaster. K-*Means* melakukan partisi data ke dalam kelompok/klaster sehingga data yang memiliki karakteristik yang sama akan dikelompokkan ke dalam satu klaster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain (Jiawei, Kamber, & Pei, 2012). Tujuan dari metode *clustering* ini yaitu untuk meminimalkan *objective function* yang diset dalam proses *clustering*, dimana *objective function* tersebut pada umumnya berusaha meminimalkan variasi di dalam suatu klaster dan memaksimalisasikan variasi antar klaster (Agusta, 2007). Dalam melakukan pengelompokan menggunakan K-*Means*, langkah-langkah yang harus dilakukan adalah sebagai berikut menurut (Jiawei, Kamber, & Pei, 2012).

1. Menentukan jumlah *cluster* (*k*) sekaligus pusat awal *cluster* dari sekumpulan data secara random.
2. Menghitung jarak antara data dengan pusat klaster (*centroid*). Perhitungan jarak data dengan *centroid* dilakukan dengan menghitung jarak *euclidean* melalui persamaan sebagai berikut:

$$d_{(i,j)} = \sqrt{\sum_{k=1}^{p} |v_{ik} - x_{jk}|^2} \qquad (2.2)$$

Dimana,

d_{ij} : jarak *euclidean* antara dokumen ke-*j* dengan *centroid cluster* ke-*i*

v_{ik} : *centroid* untuk kata ke-*k* pada *cluster* ke-*i*

x_{jk} : frekuensi kemunculan kata ke-*k* pada dokumen ke-*j*

p : dimensi data

3. Mengalokasikan data ke masing-masing *cluster* berdasarkan jarak *euclidean* minimum antara data dengan *centroid*.
4. Menghitung kembali *centroid cluster* ke-*i* menggunakan persamaan sebagai berikut:

$$v_{ik} = \frac{\sum_{j=i}^{n_i} x_{jk}}{n_i} \quad (2.3)$$

Dimana,

v_{ik} : *centroid* pada *cluster* ke-*i* untuk kata ke-*k*

x_{jk}: frekuensi kemunculan kata ke-*k* pada dokumen ke-*j* yang berada pada *cluster* ke-*i*

n_i : banyak dokumen yang menjadi anggota *cluster* ke-*i*

5. Kembali menghitung pusat klaster (*centroid*) seperti langkah 3 dan seterusnya secara berulang-ulang hingga tidak ada lagi perubahan atau anggota klaster yang berpindah.

Agar dapat memahami algoritma K-*Means* yang diterapkan dalam *text clustering*, berikut ini merupakan ilustrasi penerapan algoritma K-*Means* untuk jumlah *cluster* $k = 2$ dan banyaknya dokumen teks $N = 4$ sebagai berikut.

1. Menentukan *centroid* awal untuk $k = 2$

Tabel 2.2 Ilustrasi *Centroid* Awal untuk $k = 2$

***Cluster* ke-**	x_1	x_2	…	x_k	…	x_p
1	v_{11}	v_{12}	…	v_{1k}	…	v_{1p}
2	v_{21}	v_{22}	…	v_{2k}	…	v_{2p}

Dimana,

$x_1, x_2, \ldots, x_p$: *Feature* atau kata yang berukuran *p* dimensi

$v_{11}, v_{12}, \ldots, v_{1p}$: Vektor *centroid* untuk *cluster* 1

$v_{21}, v_{22}, \ldots, v_{2p}$: Vektor *centroid* untuk *cluster* 2

2. Menghitung jarak *euclidean* menggunakan persamaan (2.2)

Tabel 2.3 Ilustrasi Menghitung Jarak *Euclidean* untuk $k = 2$ dan $N = 4$

No	Jarak *Euclidean*	Keterangan
1	$d_{(1,1)} = \sqrt{\lvert v_{11} - x_{11}\rvert^2 + \cdots + \lvert v_{1p} - x_{1p}\rvert^2}$	Jarak *euclidean* antara data ke-1 dengan pusat *cluster* ke-1
2	$d_{(1,2)} = \sqrt{\lvert v_{11} - x_{21}\rvert^2 + \cdots + \lvert v_{1p} - x_{2p}\rvert^2}$	Jarak *euclidean* antara data ke-2 dengan pusat *cluster* ke-1
3	$d_{(1,3)} = \sqrt{\lvert v_{11} - x_{31}\rvert^2 + \cdots + \lvert v_{1p} - x_{3p}\rvert^2}$	Jarak *euclidean* antara data ke-3 dengan pusat *cluster* ke-1
4	$d_{(1,4)} = \sqrt{\lvert v_{11} - x_{41}\rvert^2 + \cdots + \lvert v_{1p} - x_{4p}\rvert^2}$	Jarak *euclidean* antara data ke-4 dengan pusat *cluster* ke-1
5	$d_{(2,1)} = \sqrt{\lvert v_{21} - x_{11}\rvert^2 + \cdots + \lvert v_{2p} - x_{1p}\rvert^2}$	Jarak *euclidean* antara data ke-1 dengan pusat *cluster* ke-2
6	$d_{(2,2)} = \sqrt{\lvert v_{21} - x_{21}\rvert^2 + \cdots + \lvert v_{2p} - x_{2p}\rvert^2}$	Jarak *euclidean* antara data ke-2 dengan pusat *cluster* ke-2
7	$d_{(2,3)} = \sqrt{\lvert v_{21} - x_{31}\rvert^2 + \cdots + \lvert v_{2p} - x_{3p}\rvert^2}$	Jarak *euclidean* antara data ke-3 dengan pusat *cluster* ke-2
8	$d_{(2,4)} = \sqrt{\lvert v_{21} - x_{41}\rvert^2 + \cdots + \lvert v_{2p} - x_{4p}\rvert^2}$	Jarak *euclidean* antara data ke-4 dengan pusat *cluster* ke-2

3. Mengalokasikan data berdasarkan jarak *euclidean* minimum dengan ilustrasi sebagai berikut

Tabel 2.4 Ilustrasi Alokasi Data pada Masing-Masing *Cluster*

No	Jarak *Euclidean Cluster* 1	Jarak *Euclidean Cluster* 2	Alokasi Data
1	$d_{(1,1)}$	$d_{(2,1)}$	$min(d_{(1,1)}, d_{(2,1)})$
2	$d_{(1,2)}$	$d_{(2,2)}$	$min(d_{(1,2)}, d_{(2,2)})$
3	$d_{(1,3)}$	$d_{(2,3)}$	$min(d_{(1,3)}, d_{(2,3)})$
4	$d_{(1,4)}$	$d_{(2,4)}$	$min(d_{(1,4)}, d_{(2,4)})$

Proses pengambilan keputusan penentuan anggota *cluster* dapat diilustrasikan pada Tabel 2.5 berikut.

Tabel 2.5 Ilustrasi Pengambilan Keputusan Alokasi Data pada Masing-Masing *Cluster*

Alokasi Data	Keterangan
$min(d_{(1,1)}, d_{(2,1)})$	$jika\ d_{(1,1)} < d_{(2,1)}$, maka dokumen 1 menjadi anggota *cluster* 1
$min(d_{(1,2)}, d_{(2,2)})$	$jika\ d_{(1,2)} < d_{(2,2)}$, maka dokumen 2 menjadi anggota *cluster* 1
$min(d_{(1,3)}, d_{(3,1)})$	$jika\ d_{(1,3)} < d_{(2,3)}$, maka dokumen 3 menjadi anggota *cluster* 1
$min(d_{(1,4)}, d_{(4,1)})$	$jika\ d_{(1,4)} < d_{(2,4)}$, maka dokumen 4 menjadi anggota *cluster* 1

4. Menghitung *centroid* baru untuk setiap *cluster* menggunakan persamaan (2.3) yang diilustrasikan seperti pada Tabel 2.6 berikut.

Tabel 2.6 Ilustrasi Menghitung *Centroid* Baru

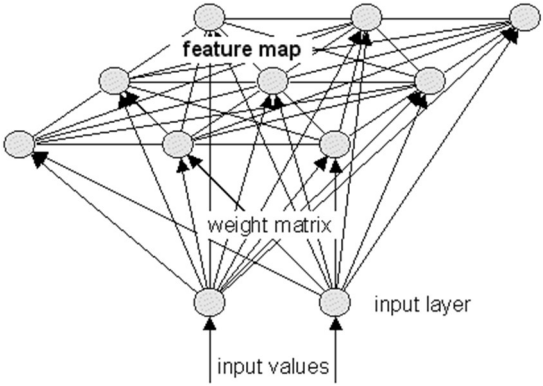
Cluster ke-	x_1	…	x_k	…	x_p
1	$v_{11}=\frac{\sum_{j=i}^{n_1} x_{j1}}{n_1}$	…	$v_{1k}=\frac{\sum_{j=i}^{n_1} x_{jk}}{n_1}$	…	$v_{1p}=\frac{\sum_{j=i}^{n_1} x_{jp}}{n_1}$
2	$v_{21}=\frac{\sum_{j=i}^{n_2} x_{j1}}{n_2}$	…	$v_{2k}=\frac{\sum_{j=i}^{n_2} x_{jk}}{n_2}$	…	$v_{2p}=\frac{\sum_{j=i}^{n_2} x_{jp}}{n_2}$

2.7.2 *Self-Organizing Maps*

Artificial Neural Network adalah suatu jaringan yang memiliki konsep mirip dengan jaringan otak manusia, dimana jaringan tersebut dapat dilatih sehingga dapat mengambil keputusan seperti yang dilakukan manusia. Pada jaringan syaraf tiruan juga dikenal metode pembelajaran yaitu *supervised learning* dan *unsupervised learning*. Contoh metode *unsupervised learning* di antaranya adalah *Kohonen Self-Organizing Maps.*

Kohonen Self-Organizing Maps atau *Self-Organizing Maps (SOM)* adalah salah satu jenis model *neural network*. SOM dikembangkan pada tahun 1982 oleh Professor Teuvo Kohonen. Dinamakan "*Self-Organizing*" karena tidak memerlukan pembelajaran (*unsupervised learning*) dan disebut "Maps" karena SOM berusaha untuk memetakan bobotnya agar sesuai dengan input data yang diberikan. SOM memungkinkan visualisasi dan proyeksi dari data berdimensi tinggi ke dimensi yang lebih rendah, paling sering menjadi bidang 2-D dengan tetap mempertahankan topologi data tersebut (Feldman & Sanger, 2007).

SOM merupakan suatu penerapan dari *neural netrwork* yang menggunakan *multiinput* dan *multioutput*. Dalam SOM tidak dikenal adanya *hidden layer*. Secara umum, bentuk kohonen SOM terdiri atas dua bentuk berdasarkan unit *output*-nya yaitu bentuk satu dimensi dan dua dimensi. Arsitektur jaringan pada SOM dua dimensi, *output layer* akan direpresentasikan sebagai suatu matriks/*array* 2 dimensi. Tiap unit pada bentuk dua dimensi akan terhubung secara lengkap dengan tiap unit *input* yang ada. Pada matriks itulan, nantinya, SOM akan terjadi. Pada *layer output*, akan terjadi *competitive learning* yang akan menyebabkan tiap unit *output* akan saling bersaing untuk menjadi unit pemenang dari *input* yang diberikan. Secara jelas, arsitektur dua dimensi pada SOM dapat dilihat pada Gambar 2.2 sebagai berikut.



Gambar 2.2 Arsitektur SOM Dua Dimensi
(Sumber: Neural Network with Java, 2004)

Tahapan algoritma SOM adalah sebagai berikut (Setiawan, 2003)

1. Neuron pada lapisan *input* (*neuron input*) sebanyak *i* dinotasikan sebagai x_1, x_2, x_3, … x_i dan *neuron* pada lapisan *output* (*neuron output*) sebanyak *j* x *l* dinotasikan sebagai y_{11}, y_{12}, y_{13}, …, y_{jl}. Bobot koneksi antara *neuron input* dan *output* dinotasikan sebagai W_{ijl}.
2. Inisialisasi bobot koneksi antara *neuron input* dan *output* (W_{ijl}) dengan bilangan random antara 0 dan 1.
3. Ulangi langkah 4 sampai dengan 7 hingga konvergen (perubahan bobot relatif kecil/lebih kecil dari batas toleransi) atau *cycle* (langkah 4 sampai dengan 7) telah dilakukan sebanyak jumlah yang telah ditentukan.
4. Pilih salah satu vektor *input x* secara acak (yang juga bilangan real random antara 0 dan 1) yang hendak diklasterkan dan di-*input*-kan ke *neuron input*.
5. Hitung jarak vektor *input* terhadap bobot koneksi d_{jl} untuk masing-masing *neuron output* dengan menggunakan rumus:

$$d_{jl} = \sum_{i=1}^{n}(w_{ijl} - x_i)^2 \qquad (2.5)$$

Dimana,

d_{jl}: Jarak *euclidean* antara *vector input* dengan *vector* bobot menuju *grid* berukuran *j x l*

w_{ijl}: Vektor bobot ke-*l* pada komponen ke-*i* dalam *vector input* ke-*j*

x_i : Komponen ke-*i*

6. Cari index *b* = *j*, *c* = *l*, dimana d_{jl} minimum, *neuron output bc* disebut *Best Matching Unit* (BMU).
7. Untuk setiap W_{ijl}, perbaharui bobot koneksi dengan mengunakan rumus:

$$w_{ijl}(t+1) = w_{ijl}(t) + \alpha(x_i(t) - w_{ijl}(t)) \quad (2.6)$$

Dimana,

$w_{ijl}(t+1)$: Vektor bobot baru ke-*l* pada komponen ke-*i* dalam *vector input* ke-*j*

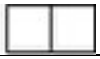
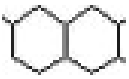
$w_{ijl}(t)$: Vektor bobot sebelumnya ke-*l* pada komponen ke-*i* dalam *vector input* ke-*j*

α : *Learning rate*

$x_i(t)$: Komponen ke-*i*

Agar dapat memahami bagaimana penerapan algoritma SOM pada *text clustering*, berikut ini akan diberikan ilustrasi yang dimulai dari penentuan topologi *map* untuk *neuron output*.

Tabel 2.7 Ilustrasi Bentuk Topologi *Map* pada SOM

Jenis Topologi *Map*	Ukuran *Grid*	Visualisasi *Map*
Rectangular	2 x 1	
	2 x 2	
Hexagonal	2 x 1	
	2 x 2	

Setelah menentukan bentuk topologi *map* untuk *neuron output*, selanjutnya menentukan inisialisasi bobot (W_{ijl}) dan *learning rate* (α). Dalam ilustrasi ini akan dijelaskan algoritma SOM untuk topologi *rectangular* dengan *grid* berukuran 2x2 dengan banyak dokumen yang akan diklasterkan N = 4, dan *learning rate* (α) = 0,5.

Tabel 2.8 Ilustrasi Bobot pada SOM

No	Notasi Bobot	Vektor Bobot
1	W_{i00}	$(w_{100}\ \ w_{200}\ \ \dots\ \ w_{i00}\ \ \cdots\ \ w_{p00})$
2	W_{i10}	$(w_{110}\ \ w_{210}\ \ \dots\ \ w_{i10}\ \ \cdots\ \ w_{p10})$
3	W_{i01}	$(w_{101}\ \ w_{201}\ \ \dots\ \ w_{i01}\ \ \cdots\ \ w_{p01})$
4	W_{i11}	$(w_{111}\ \ w_{211}\ \ \dots\ \ w_{i11}\ \ \cdots\ \ w_{p11})$

Dimana,

W_{i00} : Vektor bobot berukuran *p* dimensi antara *neuron input* ke-*i* dengan *neuron ouput* pada $x = 0$ dan $y = 0$ (Klaster 1)

W_{i10} : Vektor bobot berukuran *p* dimensi antara *neuron input* ke-*i* dengan *neuron ouput* pada $x = 1$ dan $y = 0$ (Klaster 2)

W_{i01} : Vektor bobot berukuran *p* dimensi antara *neuron input* ke-*i* dengan *neuron ouput* pada $x = 0$ dan $y = 1$ (Klaster 3)

W_{i11} : Vektor bobot berukuran *p* dimensi antara *neuron input* ke-*i* dengan *neuron ouput* pada $x = 1$ dan $y = 1$ (Klaster 4)

Selanjutnya, menghitung jarak *euclidean* antara vektor input dengan vektor bobot menuju *grid* berukuran 2 x 2 yang diilustrasikan sebagai berikut.

Tabel 2.9 Ilustrasi Perhitungan Jarak *Euclidean* untuk Dokumen 1

No	Jarak *Euclidean*	Keterangan
1	$d_{00} = \sqrt{\lvert w_{100} - x_1\rvert^2 + \cdots + \lvert w_{100} - x_1\rvert^2}$	Jarak *euclidean* antara data ke-1 dengan bobot menuju grid(0,0)
2	$d_{10} = \sqrt{\lvert w_{110} - x_1\rvert^2 + \cdots + \lvert w_{110} - x_1\rvert^2}$	Jarak *euclidean* antara data ke-1 dengan bobot menuju grid(1,0)

Tabel 2.9 Ilustrasi Perhitungan Jarak *Euclidean* untuk Dokumen 1 (lanjutan)

No	Jarak *Euclidean*	Keterangan
3	$d_{01}= \sqrt{\lvert w_{101}-x_1\rvert^2+\cdots+\lvert w_{101}-x_1\rvert^2}$	Jarak *euclidean* antara data ke-1 dengan bobot menuju grid(0,1)
4	$d_{11}= \sqrt{\lvert w_{111}-x_1\rvert^2+\cdots+\lvert w_{111}-x_1\rvert^2}$	Jarak *euclidean* antara data ke-1 dengan bobot menuju grid(1,1)

Berikutnya adalah ilustrasi jarak *Euclidean* antara vektor input dokumen ke-2 menuju *grid* berukuran 2 x 2.

Tabel 2.10 Ilustrasi Perhitungan Jarak *Euclidean* untuk Dokumen 2

No	Jarak *Euclidean*	Keterangan
1	$d_{00}= \sqrt{\lvert w_{200}-x_2\rvert^2+\cdots+\lvert w_{200}-x_2\rvert^2}$	Jarak *euclidean* antara data ke-2 dengan bobot menuju grid(0,0)
2	$d_{10}= \sqrt{\lvert w_{210}-x_2\rvert^2+\cdots+\lvert w_{210}-x_2\rvert^2}$	Jarak *euclidean* antara data ke-2 dengan bobot menuju grid(1,0)
3	$d_{01}= \sqrt{\lvert w_{201}-x_2\rvert^2+\cdots+\lvert w_{201}-x_2\rvert^2}$	Jarak *euclidean* antara data ke-2 dengan bobot menuju grid(0,1)
4	$d_{11}= \sqrt{\lvert w_{211}-x_2\rvert^2+\cdots+\lvert w_{211}-x_2\rvert^2}$	Jarak *euclidean* antara data ke-2 dengan bobot menuju grid(1,1)

Dengan langkah yang sama, dilakukan perhitungan jarak *euclidean* untuk vektor input dokumen ke-3 sebagai berikut.

Tabel 2.11 Ilustrasi Perhitungan Jarak *Euclidean* untuk Dokumen 3

No	Jarak *Euclidean*	Keterangan
1	$d_{00}= \sqrt{\lvert w_{300}-x_3\rvert^2+\cdots+\lvert w_{300}-x_3\rvert^2}$	Jarak *euclidean* antara data ke-3 dengan bobot menuju grid(0,0)
2	$d_{10}= \sqrt{\lvert w_{310}-x_3\rvert^2+\cdots+\lvert w_{310}-x_3\rvert^2}$	Jarak *euclidean* antara data ke-3 dengan bobot menuju grid(1,0)

Tabel 2.11 Ilustrasi Perhitungan Jarak *Euclidean* untuk Dokumen 3 (lanjutan)

No	Jarak *Euclidean*	Keterangan
3	$d_{01} = \sqrt{\lvert w_{301} - x_3 \rvert^2 + \cdots + \lvert w_{301} - x_3 \rvert^2}$	Jarak *euclidean* antara data ke-3 dengan bobot menuju grid(0,1)
4	$d_{11} = \sqrt{\lvert w_{311} - x_3 \rvert^2 + \cdots + \lvert w_{311} - x_3 \rvert^2}$	Jarak *euclidean* antara data ke-3 dengan bobot menuju grid(1,1)

Dan yang terakhir, dilakukan perhitungan jarak *euclidean* untuk vektor input dokumen ke-4 sebagai berikut.

Tabel 2.12 Ilustrasi Perhitungan Jarak *Euclidean* untuk Dokumen 4

No	Jarak *Euclidean*	Keterangan
1	$d_{00} = \sqrt{\lvert w_{400} - x_4 \rvert^2 + \cdots + \lvert w_{400} - x_4 \rvert^2}$	Jarak *euclidean* antara data ke-4 dengan bobot menuju grid(0,0)
2	$d_{10} = \sqrt{\lvert w_{410} - x_4 \rvert^2 + \cdots + \lvert w_{410} - x_4 \rvert^2}$	Jarak *euclidean* antara data ke-4 dengan bobot menuju grid(1,0)
3	$d_{01} = \sqrt{\lvert w_{401} - x_4 \rvert^2 + \cdots + \lvert w_{401} - x_4 \rvert^2}$	Jarak *euclidean* antara data ke-4 dengan bobot menuju grid(0,1)
4	$d_{11} = \sqrt{\lvert w_{411} - x_4 \rvert^2 + \cdots + \lvert w_{411} - x_4 \rvert^2}$	Jarak *euclidean* antara data ke-4 dengan bobot menuju grid(1,1)

Setelah itu, mengelompokkan masing-masing data pada *neuron ouput* berdasarkan jarak minimum *euclidean*.

Tabel 2.13 Ilustrasi Pengelompokkan Data dengan SOM Berdasarkan Jarak *Euclidean*

No	*Grid* (0,0)	*Grid* (0,1)	*Grid* (1,0)	*Grid* (1,1)	Alokasi Data
1	d_{00}	d_{10}	d_{01}	d_{11}	$min(d_{00}, d_{10}, d_{01}, d_{11})$
2	d_{00}	d_{10}	d_{01}	d_{11}	$min(d_{00}, d_{10}, d_{01}, d_{11})$
3	d_{00}	d_{10}	d_{01}	d_{11}	$min(d_{00}, d_{10}, d_{01}, d_{11})$
4	d_{00}	d_{10}	d_{01}	d_{11}	$min(d_{00}, d_{10}, d_{01}, d_{11})$

Dimana,
Grid(0,0) = Klaster 1
Grid(0,1) = Klaster 2
Grid(1,0) = Klaster 3
Grid(1,1) = Klaster 4

Jarak *euclidean* paling minimum akan menentukan *neuron output* pemenang yang selanjutnya vektor bobot yang menghubungkan antara *neuron input* dengan *neuron output* akan diperbarui nilainya menggunakan persamaan (2.6). Sebagai contoh, vektor bobot pada *grid* (1,1) memiliki nilai minimum sehingga akan diperbarui bobotnya yang akan diilustrasikan sebagai berikut.

Tabel 2.14 Ilustrasi Pembaruan Bobot pada Metode SOM

Bobot Baru	Perhitungan
$w_{(1,1,1)}(t+1)$	$w_{(1,1,1)}(t) + 0{,}05\big(x_1(t) - w_{(1,1,1)}(t)\big)$
$w_{(2,1,1)}(t+1)$	$w_{(2,1,1)}(t) + 0{,}05\big(x_2(t) - w_{(2,1,1)}(t)\big)$
$w_{(3,1,1)}(t+1)$	$w_{(3,1,1)}(t) + 0{,}05\big(x_3(t) - w_{(3,1,1)}(t)\big)$
$w_{(4,1,1)}(t+1)$	$w_{(4,1,1)}(t) + 0{,}05\big(x_4(t) - w_{(4,1,1)}(t)\big)$

Setelah itu dilakukan iterasi terus menerus hingga diperoleh hasil klaster yang optimal.

2.8 *Silhouette Coefficient*

Setelah membentuk *cluster* yang paling optimum, maka hasil *cluster* yang terbentuk perlu dilakukan evaluasi untuk mengukur tingkat kebaikan dalam pengelompokkan data. Salah satu metode yang digunakan untuk mengevaluasi model *clustering* adalah *silhouette coefficient*. Dengan *silhouette coefficient,* dapat diketahui seberapa baik kelompok-kelompok (*cluster*) dipisahkan dan seberapa kompak suatu *cluster*. Berikut adalah tahapan dalam *perhitungan silhouette coefficient*.

1. Menghitung *a(i)*, yaitu rata-rata dari artikel ke-*i* dengan semua artikel lain yang berada dalam satu *cluster*

$$a(i) = \frac{\sum_{j \in C_i, i \neq j} d(i,j)}{|Ci|-1} \quad (2.7)$$

Dimana,
C_i : *cluster* dimana artikel ke-*i* berada

$|C_i|$: banyak artikel dalam *cluster* C_i
J : artikel lain dalam *cluster* C_i
d(i,j) : jarak *euclidean* antara artikel ke-*i* dengan artikel ke-*j*

2. Menghitung *b(i)*, yaitu rata-rata jarak dari artikel ke-*i* dengan semua artikel yang berada di *cluster* lain dan mencari nilai yang terkecil.

$$b(i) = \min C_{k,k\neq i} \left\{ \frac{\sum_{k\in C_k} d(i,k)}{|C_k|} \right\} \quad (2.8)$$

Dimana,
C_k : *cluster* dimana artikel ke-*k* berada
$|C_k|$: banyak artikel dalam *cluster* C_i
k : artikel lain dalam *cluster* C_i
d(i,k): jarak *Euclidean* antara artikel ke-*i* dengan artikel ke-*j*

3. Mendapatkan nilai *silhouette coefficient* dengan persamaan berikut.

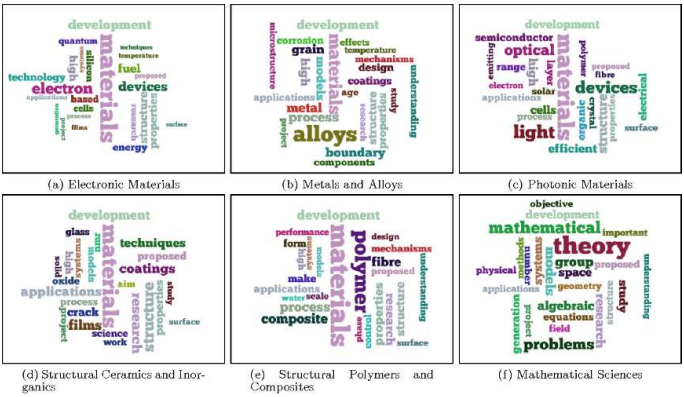
$$y(j) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (2.9)$$

Nilai dari *silhouette coefficient* berada di antara -1 dan 1. Nilai *silhouette coefficient* yang mendekati 1 menandakan *cluster* dimana *i* berada telah kompak dan *i* jauh dari *cluster* lain. Apabila *silhouette coefficient* bernilai negative (*b(i) < a(i)*), artinya *i* lebih dekat ke objek-obyek *cluster* lain dibandingkan obyek yang berada dalam *cluster* yang sama. Hal ini merupakan situasi yang buruk dan harus dihindari. Rata-rata nilai *silhouette coefficient* dari semua obyek yang berada dalam kumpulan data atau *dataset* dapat digunakan untuk mengukur kualitas dari *clustering* yang telah dilakukan.

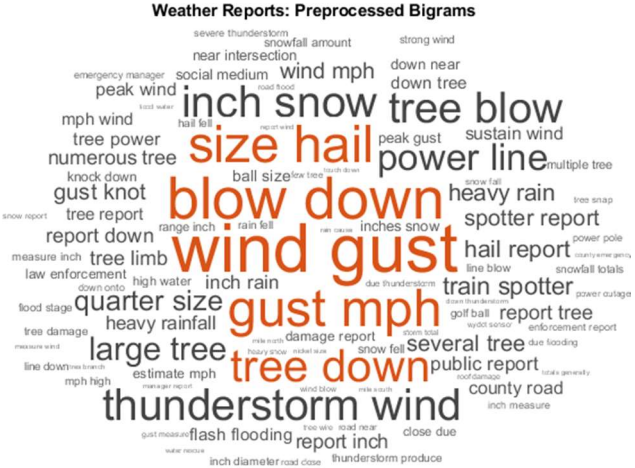
2.9 *Word Cloud*

Word cloud merupakan salah satu metode yang sering digunakan untuk penggambaran data teks. *Word cloud* dapat melakukan representasi sebuah data teks dengan cara membuat plot kata-kata yang sering muncul. Semakin sering kata itu muncul

maka huruf kata tersebut semakin besar, begitu juga apabila suatu kata jarang muncul maka ukuran kata itu akan lebih kecil dari yang lainnya. *Word cloud* juga dapat menampirkan frasa (gabungan lebih dari satu kata) yang sering muncul. Berikut ini contoh dari penggambaran data teks dengan *word cloud* ditunjukkan pada Gambar 2.2 dan Gambar 2.3.



Gambar 2.3 Contoh *Word Cloud* dengan Uni-*Gram*
(Sumber: Castella & Sutton, 2014)



Gambar 2.4 Contoh *Word Cloud* dengan Bi-*Gram*
(Sumber: Analyze Text Data Using Multiword Phrases, 2019)

2.10 Berita

Berita bagi seseorang adalah keterangan mengenai suatu peristiwa atau isi pernyataan seseorang yang menurutnya perlu diketahui untuk mewujudkan filsafat hidupnya. Jadi dapat di simpulkan bahwa berita merupakan sebuah pemberitahuan yang mengungkap tentang sebuah kejadian atau hal yang terjadi pada waktu tertentu (Soehoet, 2006). Berita disusun menurut bagian-bagian tertentu seperti:

1. *Headline*

   Headline atau biasa disebut judul. Sering juga dilengkapi dengan anak judul yang berguna untuk menolong pembaca agar segera mengetahui peristiwa yang akan diberitakan dan menonjolkan satu berita dengan dukungan teknik grafika.

2. *Deadline*

   Ada yang terdiri atas nama media massa, tempat kejadian dan tanggal kejadian. Tujuannya adalah untuk menunjukkan tempat kejadian dan inisial media.

3. *Lead*

   Lead biasa disebut dengan beranda berita. Biasanya ditulis pada paragraf pertama sebuah berita. Beranda berita merupakan unsur yang paling penting dari sebuah berita, yang menentukan apakah isi berita akan dibaca atau tidak. Selain merupakan sari pati sebuah berita, teras berita juga melukiskan seluruh berita secara singkat.

4. *Body*

   Body atau tubuh berita. Isinya menceritakan peristiwa yang dilaporkan dengan bahasa yang singkat, padat, dan jelas. Dengan demikian, *body* merupakan perkembangan berita.

2.11 Portal Tribunnews Surabaya

Salah satu situs berita *online* yang aktif memberikan informasi terkini mengenai Kota Surabaya dan Jawa Timur adalah *http://www.surya.co.id* atau *http://surabaya.tribunnews .com/.* Situs berita yang ber-*domain* tribunnews.com ini merupakan bagian dari koran-koran Tribun Network. Saat ini, portal induk yaitu TRIBUNnews.com yang dikelola oleh PT Tribun Digital Online di

bawah Divisi Koran Daerah Kompas Gramedia (*Group of Regional Newspaper*) menyajikan halaman *digital paper* dari koran-koran Tribun Network. *Digital paper* merupakan koran yang hanya terbit secara *online* dalam format digital. Dalam sebulan, jutaan netizen mengunjungi portal berita tersebut. Situs berita *online* ini berawal dari terbitnya surat kabar harian di Surabaya yaitu Harian Surya. Meski media berbasis kertas, Surya bersama divisi pemberitaan *online* terus berupaya aktif mengembangkan situs berita *online*-nya.

Menurut Pieter P. Gero dalam Laporan Wajah PERS Malang yang ditulis oleh Hughes, sejarah berdirinya Harian Surya diawali pada tahun 1986 di kota Surabaya dalam bentuk mingguan yang bernama Mingguan Surya. Pada saat itu, tiras Mingguan Surya sekitar 50.000, dan diterbitkan oleh PT Antar Surya Jaya yang dipayungi Pos Kota Grup, yang berpusat di Jakarta. Pada tahun 1989, Mingguan Surya diambil alih oleh Kompas - Gramedia Grup, kemudian berubah menjadi Harian Pagi Surya dan langsung masuk pasar di daerah Malang. Pada akhir tahun 2001, tirasnya sebanyak 150.000 eksemplar per edisi, termasuk 22.500 eksemplar yang diredar di daerah Malang. Sekarang Surya dikenal sebagai surat kabar nomor dua di Surabaya, setelah Jawa Pos (Hughes, 2001).

Gambar 2.5 Situs Berita http://surabaya.tribunnews.com/
(PT Tribun Digital Online, 2019)

(*Halaman ini sengaja dikosongkan*)

BAB III
METODOLOGI PENELITIAN

3.1 Sumber Data

Data yang digunakan dalam penelitian ini adalah 6.027 *headline* atau judul berita *online* yang diterbitkan melalui situs *http://surabaya.tribunnews.com/* pada 1 Januari 2018 hingga 31 Desember 2018 yang diambil pada hari Rabu, 6 Maret 2019 pukul 17.45 WIB. Judul berita *online* yang dikumpulkan merupakan judul berita yang membahas kejadian atau peristiwa yang terjadi dan berkaitan dengan Kota Surabaya. Teknik pengumpulan berita *online* menggunakan metode *web scraping*. Selanjutnya, kumpulan artikel berita tersebut akan diolah dan membentuk *cluster* untuk menghasilkan topik berita secara khusus.

3.2 Variabel Penelitian dan Struktur Data

Berikut merupakan contoh data yang diperoleh melalui situs *http://surabaya.tribunnews .com/* dengan filter "Berita Surabaya".

Tabel 3.1 Contoh Data

Dokumen ke-	Tanggal	Judul Berita
1	01/01/2018	Fenomena Supermoon 2 Januari 2018, Surabaya Sebaiknya Waspadai Banjir Rob
2	01/01/2018	Besok, Dispendukcapil Surabaya Gelar Yustisi untuk Antisipasi Urbanisasi
.	.	.
.	.	.
.	.	.
6.026	31/12/2018	Daftar Kasus Menonjol di Surabaya Versi Polisi, Peledakan Bom hingga Perdagangan Bayi di Instagram
6.027	31/12/2018	Malam Pergantian Tahun Baru, Cuaca Kota Surabaya Diprediksi Cerah Berawan

Setelah melakukan pengumpulan data, selanjutnya melakukan tahap *pre-processing* teks untuk mengolah data tekstual sehingga menjadi data numerik yang siap diolah. Sehingga, pada penelitian ini variabel yang digunakan dalam analisis *text clustering* adalah sebagai berikut.

Tabel 3.2 Variabel Penelitian

Variabel	Keterangan	Skala
a_{ij}	Frekuensi kemunculan kata	Rasio

Sementara itu, berikut merupakan struktur data yang digunakan dalam penelitian ini.

Tabel 3.3 Struktur Data Penelitian

Berita ke-	Kata				
	Kata ke-*1*	**Kata ke-*2***	**Kata ke-*3***	**…**	**Kata ke-*n***
1	a_{11}	a_{12}	a_{13}	…	a_{1n}
2	a_{21}	a_{22}	a_{23}	…	a_{2n}
3	a_{31}	a_{32}	a_{33}	…	a_{3n}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
m	a_{m1}	a_{m2}	a_{m3}	…	a_{mn}

Keterangan:

a_{ij} : banyak kata ke-*j* yang muncul pada dokumen berita ke-*i*, *i* = 1, 2, …, *m*

m : banyak artikel berita *online*

n : banyak kata atau *feature*

3.3 Langkah Analisis

Langkah analisis yang dilakukan dalam penelitian ini adalah sebagai berikut.

1. Melakukan *web scraping* artikel berita *online*

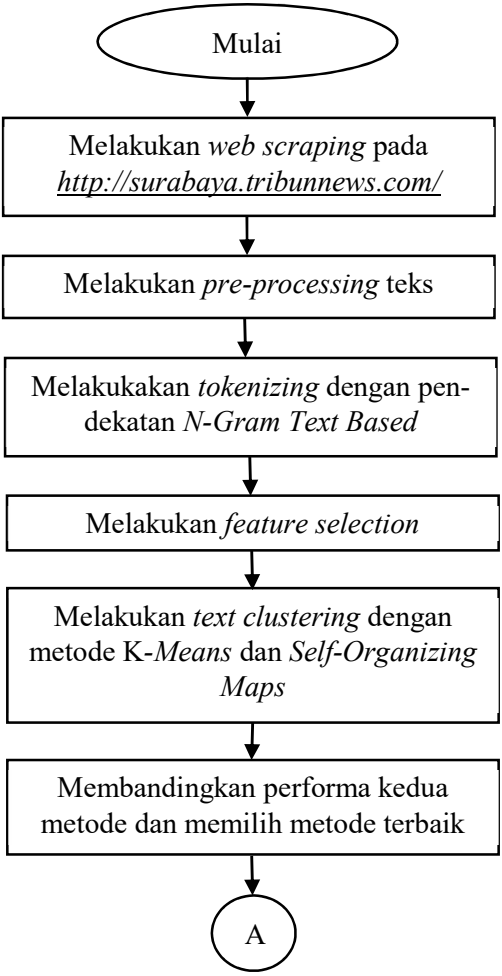
   Artikel berita *online* dikumpulkan dengan melakukan *web scraping* menggunakan bantuan *software* R pada situs *http://surabaya.tribunnews.com/* dengan *filter* Berita Surabaya. Kemudian, simpan hasil data tersebut ke dalam suatu *database* dengan format .csv

2. Melakukan *text pre-processing* dengan langkah-langkah sebagai berikut.
   a. Melakukan *case folding* yaitu mengubah semua huruf dalam dokumen berita menjadi huruf non kapital.
   b. Menghapus tanda baca, bilangan, dan karakter selain *a-z* pada dokumen artikel berita.
   c. Menghapus kata yang tidak bermakna pada dokumen berita yang terdapat pada daftar *stopwords*. Kamus *stopwords* yang digunakan dalam penelitian ini berasal dari daftar *stopwords* oleh Z. Tala ditambah daftar nama atau kata ganti orang, kata keterangan waktu, kata keterangan kuantitas, kata hubung, dan kata sifat.
   d. Menghapus spasi ganda dan karakter HTML pada dokumen berita.
3. Melakukan *tokenizing* untuk memecah dokumen judul berita *online* menjadi kata per kata menggunakan pendekatan *unigram* dan *bigrams*.
4. Melakukan *feature selection* berbasis *term* dan *document frequency* yang muncul dalam dokumen berita.
5. Melakukan konversi data teks menjadi numerik berdasarkan frekuensi kemunculan kata pada setiap artikel serta menentukan pembobotan data teks dengan menggunakan TF-IDF.
6. Melakukan *text clustering* dengan algoritma K-*Means* sebagai berikut.
   a. Menentukan *k* jumlah klaster/kelompok. Pada penelitian ini kombinasi *k* yang digunakan adalah 2 hingga 10.
   b. Menentukan *centroid* awal secara acak sebanyak jumlah *cluster*.
   c. Menghitung jarak *euclidean* antara korpus berita dengan *centroid*.
   d. Mengalokasikan data ke dalam klaster terdekat.
   e. Kembali menghitung pusat klaster (*centroid*) seperti langkah ke-2 dan seterusnya secara berulang-ulang hingga tidak ada lagi perubahan atau anggota klaster yang berpindah.

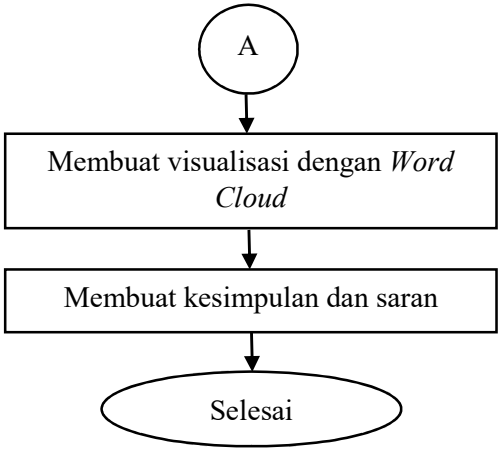
7. Melakukan *text clustering* dengan algoritman *Self-Organizing Maps* sebagai berikut.
   a. Menyiapkan *grid* berukuran dua dimensi
   b. Inisialisasi bobot koneksi antara *neuron input* dan *output* (W_{ijl}) dengan bilangan random antara 0 dan 1.
   c. Pilih salah satu vektor *input* x secara acak (yang juga bilangan real random antara 0 dan 1) yang hendak diklasterkan dan di-*input*-kan ke *neuron input*.
   d. Hitung jarak vektor *input* terhadap bobot koneksi d_{jl} untuk masing-masing *neuron output*.
   e. Hitung jarak vektor *input* terhadap bobot koneksi d_{jl} untuk masing-masing *neuron output* dengan menggunakan persamaan (2.5).
   f. Cari index $b = j$, $c = l$, dimana d_{jl} minimum, *neuron output bc* disebut *Best Matching Unit* (BMU).
   g. Untuk setiap W_{ijl}, perbaharui bobot koneksi dengan mengunakan persamaan (2.6).
   h. Ulangi langkah c sampai dengan g hingga konvergen (perubahan bobot relatif kecil/lebih kecil dari batas toleransi) atau *cycle* (langkah 4 sampai dengan 7) telah dilakukan sebanyak jumlah yang telah ditentukan.
8. Membandingkan performa kedua metode dan memilih metode terbaik dengan melihat nilai *average width* terbesar
9. Melakukan visualisasi dengan *word cloud* untuk mengetahui frekuensi kemunculan kata yang paling tinggi sebanyak k *cluster* optimum serta menggali informasi pada setiap *cluster* yang terbentuk.
10. Menentukan topik serta membuat kesimpulan dan saran bagi pemerintah dan masyarakat; manajemen Tribunnews Surabaya; dan juga penelitian selanjutnya.

3.4 Diagram Alir

Diagram alir penelitian ini digambarkan sebagai berikut.



Gambar 3.1. Diagram Alir Penelitian



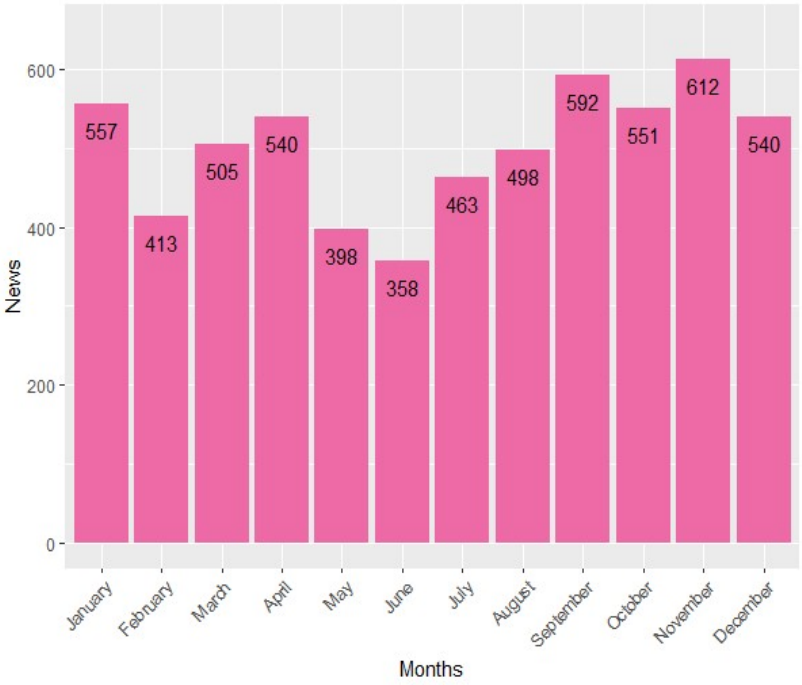
Gambar 3.1. Diagram Alir Penelitian (lanjutan)

BAB IV
ANALISIS DAN PEMBAHASAN

Pada bab ini akan dijelaskan mengenai analisis teks pada dokumen berita yang diterbitkan pada tahun 2018 di situs digital Harian Surya Surabaya yang beralamatkan *http://surabaya.tribunnews.com/*. Sebanyak 6.027 dokumen berita akan dianalisis pada bagian ini yang dimulai dari eksplorasi data awal untuk mengetahui karakteristik dokumen berita. Kemudian dilanjutkan pada tahap *text pre-processing*, *feature selection*, pembobotan *feature* dengan TF-IDF, menentukan jumlah klaster optimum untuk metode *clustering* dengan K-*Means* dan *Self-Organizing Maps* (SOM) beserta evaluasinya menggunakan metode *average silhouette width*. Setelah memperoleh jumlah *cluster* optimum selanjutnya membentuk *text clustering* yang dilanjutkan membuat visualisasi kata yang sering muncul menggunakan *word cloud* untuk pencarian informasi berdasarkan topik berita mengenai Kota Surabaya.

4.1 Karakteristik Data

Portal berita *http://surabaya.tribunnews.com/* merupakan salah satu media berita *online* yang aktif menerbitkan berita khususnya mengenai Kota Surabaya dan sekitarnya setiap harinya. Berbagai peristiwa, kejadian, dan kegiatan yang dilaksanakan di Kota Surabaya diliput untuk kemudian diberitakan kepada masyarakat sebagai sarana penyampaian informasi mengenai Kota Surabaya terkini. Dengan bantuan program komputer, dilakukan pengumpulan judul berita mengenai Kota Surabaya yang diterbitkan melalui portal berita *online* Tribunnews Surabaya sepanjang tahun 2018. Sebanyak 6.027 judul berita telah dikumpulkan untuk dilakukan analisis lebih lanjut. Eksplorasi data awal dilakukan dalam penelitian ini untuk mengetahui karakteristik data berita secara umum. Berikut grafik yang menggambarkan jumlah berita yang diterbitkan oleh Triibunnews Surabaya setiap bulannya selama tahun 2018.



Gambar 4.1 Jumlah Berita Surabaya yang Dipublikasikan Tahun 2018

Gambar 4.1 menunjukkan bahwa jumlah berita mengenai Kota Surabaya pada tahun 2018 banyak diterbitkan pada bulan November, yakni sebanyak 612 berita. Sedangkan pada bulan Juni portal berita Tribunnews Surabaya menerbitkan berita mengenai Kota Surabaya paling sedikit di sepanjang tahun 2018 yakni 358 artikel berita. Hal ini menunjukkan banyak peristiwa dan kejadian penting yang terjadi selama bulan November dibandingkan bulan-bulan lainnya sehingga banyak berita yang diterbitkan oleh portal berita Tribunnnews Surabaya melalui *website*-nya.

Berikut merupakan karakteristik mengenai jumlah berita yang diterbitkan setiap bulan oleh Tribunnews Surabaya berdasarkan ukuran pemusatan dan persebaran data.

Tabel 4.1 Ukuran Statistik Jumlah Berita Kota Surabaya yang Diterbitkan Tahun 2018

Min	Q1	Median	Q3	Maks	*Mean*	*Varians*
358	451	523	553	612	502	6.321,1

Tabel 4.1 menampilkan bahwa rata-rata jumlah berita yang diterbitkan tiap bulannya adalah sebanyak 502 artikel berita. Dimana, tingkat keragaman jumlah berita yang diterbitkan setiap bulannya mengenai Kota Surabaya adalah sebesar 6.321. Tingkat keragaman yang dihasilkan cukup tinggi. Hal ini dapat disimpulkan bahwa tidak ada ketentuan khusus yang diterapkan oleh pihak manajemen Tribunnews Surabaya dalam menentukan jumlah berita yang harus dipublikasikan setiap bulannya sehingga keragaman jumlah berita yang diterbitkan cukup tinggi. Kesimpulan ini juga didukung dari *range* jumlah berita yang diterbitkan antara bulan Juni (358 berita) dan November (612 berita) cukup besar. Kebijakan ini wajar diterapkan oleh pihak manajemen perusahaan yang bergerak di bidang jurnalistik untuk tidak memberi batasan terhadap jumlah berita yang diliput karena media bekerja untuk melakukan pemberitaan terhadap peristiwa yang terjadi kepada masyarakat, dimana suatu peristiwa biasanya terjadi di luar perkiraan manusia.

Setelah mengetahui karakteristik data awal mengenai berita Kota Surabaya tahun 2018, selanjutnya data berita akan memasuki tahap pra proses. Sekumpulan data teks judul berita merupakan data yang diperoleh dengan bantuan program *web scraper* sehingga masih mengandung karakter HTML dan karakter-karakter lainnya (*noise*) sehingga dapat menghambat proses analisis pada data teks judul berita. Data teks judul berita atau yang selanjutnya disebut korpus berita, perlu melewati beberapa tahap pra proses sebelum melakukan analisis *text clustering*. Tahapan pra proses yang dilakukan dalam penelitian ini adalah mengubah semua teks menjadi huruf non-kapital (*case folding*), menghapus tanda baca (*removePunct*), angka (*removeNum*), *stopwords* (*removeStopwords*), spasi ganda, dan karakter HTML (*removeWhiteSpace*). Hasil korpus berita setelah melewati masing-masing tahap pra proses disajikan pada Tabel 4.2.

Tabel 4.2 Tahap Pra Proses pada Korpus Berita

No	Pra Proses	Korpus Berita	Keterangan
1	Sebelum pra proses	"t\t\Fenomena Supermoon 2 Januari 2018, Surabaya Sebaiknya Waspadai Banjir Rob\t\t\t\t\t\t\t"	Contoh data awal hasil *scraping* pada *website* Trbunnews.
2	*Case folding*	"\t\t\tfenomena supermoon 2 januari 2018, surabaya sebaiknya waspadai banjir rob\t\t\t\t\t\t\t"	Setiap huruf kapital berubah menjadi huruf non kapital.
3	*removePunct*	"\t\t\tfenomena supermoon 2 januari 2018 surabaya sebaiknya waspadai banjir rob\t\t\t\t\t\t\t"	Menghapus tanda baca koma (,).
4	*removeNum*	"\t\t\tfenomena supermoon januari surabaya sebaiknya waspadai banjir rob\t\t\t\t\t\t\t"	Menghapus bilangan, pada dokumen ini bilangan "2" dan "2018" terhapus.
5	*removeStopwords*	"\t\t\tfenomena supermoon waspadai banjir rob\t\t\t\t\t\t\t"	Menghapus *stopwords*, pada dokumen ini kata "januari", "surabaya", dan "sebaik-nya" terhapus
6	*removeStripWhiteSpace*	"fenomena supermoon waspadai banjir rob"	Menghapus spasi ganda dan karakter HTML, pada dokumen ini karakter "\t\t\t", "\t\t\t \t\t\t", dan spasi ganda terhapus.

Case folding merupakan tahap pertama dalam pra proses pada penelitian ini yang bertujuan untuk mengubah semua karakter pada korpus berita menjadi huruf non-kapital. Selanjutnya, perlu dilakukan pembersihan terhadap karakter berupa tanda baca dan angka yang dilanjutkan dengan menghapus kata-kata yang sering

muncul namun tidak bermakna seperti kata hubung (*stopwords*). Daftar kata *stopwords* pada penelitian ini bersumber dari "Stopwords List oleh Z. Tala". Selain itu, *stopwords* pada penelitian ini ditambah dengan daftar nama orang yang muncul dalam korpus berita dan juga kata-kata yang dianggap *noise* sehingga mempengaruhi hasil validasi klaster. Kata-kata *noise* tersebut seperti kata ganti atau nama orang, keterangan waktu, keterangan kuantitas, kata hubung, dan kata sifat yang tidak bermakna dalam pembentukan *cluster*. Penghapusan karakter tanda baca, angka, *stopwords* menyebabkan spasi ganda korpus berita. Oleh karena itu, perintah *removeStripWhiteSpace* perlu dijalankan untuk menghapus spasi ganda pada korpus berita sekaligus menghilangkan karakter HTML.

Tahap selanjutnya setelah membersihkan korpus berita adalah memecah kalimat menjadi kata per kata atau *tokenizing*. Informasi mengenai frekuensi kemunculan kata yang tersebar dalam korpus berita dapat digali lebih lanjut melalui *tokenizing*. Teknik *tokenizing* yang dilakukan dalam penelitian ini adalah dengan menggunakan *unigram* dan *bigrams*.

Teknik *tokenizing* dengan *unigram* memecah kalimat pada seluruh korpus berita menjadi kata per kata. Ilustrasi *tokenizing* dengan *unigram* ditunjukkan pada tabel berikut.

Tabel 4.3 *Tokenizing* dengan *Unigram*

Korpus Berita	***Unigram***
"fenomena supermoon waspadai banjir rob"	"fenomena" "supermoon" "waspadai" "banjir" "rob"

Dari ilustrasi pada Tabel 4.3 di atas, dapat dilihat bahwa kalimat pada korpus berita yang bertuliskan "fenomena supermoon waspadai banjir rob" dipecah menjadi 5 kata yaitu "fenomena", "supermoon", "waspadai", "banjir", dan "rob". Sementara itu, pada penelitian ini juga akan dilakukan *tokenizing* dengan *bigrams*. Pemecahan kalimat dengan *bigrams* akan membentuk 1 *feature*

yang berasal dari dua kata saling berurutan pada korpus berita. Ilustrasi *tokenizing* dengan *bigrams* ditunjukkan pada Tabel 4.4.

Tabel 4.4 *Tokenizing* dengan *Bigrams*

Korpus Berita	***Bigrams***
"fenomena supermoon waspadai banjir rob"	"fenomena supermoon" "supermoon waspadai" "waspadai banjir" "banjir rob"

Pada Tabel 4.4 dapat dilihat bahwa apabila dilakukan *tokenizing* dengan *bigrams*, kalimat "fenomena supermoon waspadai banjir rob" akan dipecah berdasarkan dua kata berurutan yang menyusunnya. Sehingga, pada contoh kalimat tersebut diperoleh hasil *tokenizing* dengan *bigrams* menjadi frasa-frasa seperti "fenomena supermoon", "supermoon waspadai", "waspadai banjir" dan "banjir rob".

Selanjutnya, kata-kata yang telah melalui tahap *tokenizing* akan menjadi variabel atau *feature* dalam melakukan pengelompokkan berita dengan metode *text clustering*. Perbandingan jumlah *feature* yang terbentuk setelah menghilangkan *stopwords* dan dilanjutkan dengan *tokenizing unigram* dan *bigrams* dapat dilihat pada Tabel 4.5 berikut.

Tabel 4.5 Jumlah *Feature* Setelah *Tokenizing*

Unigram	***Bigrams***
8.338	26.875

Tabel 4.5 menunjukkan bahwa jumlah *feature* yang dihasilkan setelah melalui tahap *tokenizing* dengan *unigram* diperoleh 8.338 *features* sementara pada *bigrams* diperoleh 26.875 *features*. Jumlah *feature* ini tergolong masih sangat banyak untuk selanjutnya diolah menggunakan metode *text clustering*. Oleh karena itu, perlu dilakukan reduksi jumlah *feature* dengan cara memilih *feature* yang penting dan dianggap memiliki kontribusi besar dalam melakukan analisis *clustering*. Pemilihan *feature* atau *feature selection* pada penelitian ini dilakukan berdasarkan frekuensi kemunculan *feature* pada tiap korpus berita. Hasil dari *feature selection* pada *unigram* akan dijelaskan pada Tabel 4.6.

Tabel 4.6 Jumlah *Feature* pada *Unigram*

Metode	Jumlah *Feature*
Sebelum *feature selection*	8.338
Feature selection	37

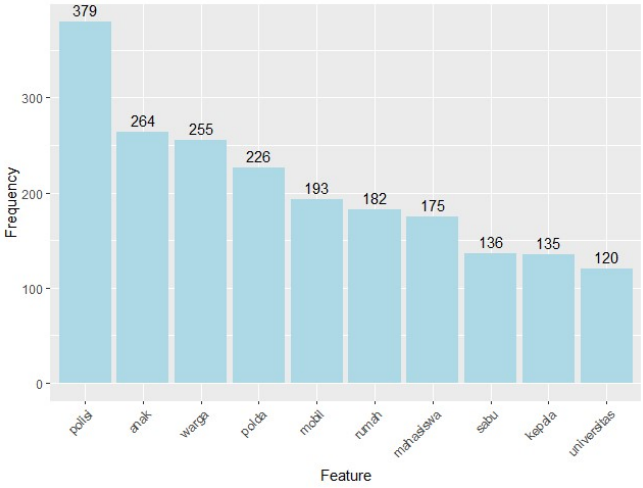
Teknik *feature selection* yang digunakan pada *unigram* adalah dengan memilih *feature* penting dengan *sparsity* = 0,99. Artinya, dari 8.338 *feature* akan dilakukan penghapusan pada *feature* yang proporsi kemunculannya di bawah 1% dari keseluruhan korpus berita. Dengan *sparsity* = 0,99 maka *feature* terpilih pada *unigram* sebanyak 37 *features*. Metode pemilihan *feature selection* dengan *sparsity* sebesar 0,99 memiliki makna yang sama dengan memilih *feature* yang penting dan disebutkan lebih dari 60 kali dalam korpus berita. Selanjutnya, hasil *feature selection* pada *bigrams* dapat dilihat pada Tabel 4.7.

Tabel 4.7 Jumlah *Feature* pada *Bigrams*

Metode	Jumlah *Feature*
Sebelum *feature selection*	26.875
Feature selection	40

Pemilihan *feature* pada *bigrams* menggunakan batas minimum kemunculan *feature* pada 2 korpus berita yang berbeda dan memiliki jumlah frekuensi minimal sebesar 15. Artinya, *feature* yang muncul pada 1 korpus berita dan disebutkan kurang dari 15 kali pada keseluruhan korpus berita akan dihapus. Hasil *feature selection* pada *bigrams* diperoleh 40 *features.*

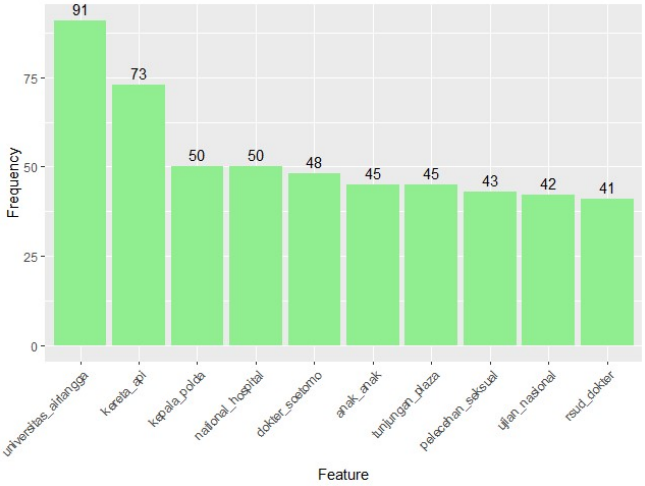
Dari hasil *features* yang telah terpilih, selanjutnya akan dibuat visualisasi mengenai sepuluh *features* pada masing-masing N-*Grams* dengan frekuensi tertinggi. Sepuluh *features* dengan frekuensi tertinggi setelah melalui tahap *tokenizing* dengan *unigram* dan *feature selection* dapat dilihat pada Gambar 4.2.



Gambar 4.2 Sepuluh *Features* dengan Frekuensi Tertinggi (*Unigram*)

Gambar 4.2 menunjukkan bahwa sepuluh *features* dengan frekuensi tertinggi setelah dilakukan pra proses dengan *unigram* adalah "polisi", "anak", "warga", "polda", "mobil", "rumah", "mahasiswa", "sabu", "kepala", dan "universitas". *Feature* "polisi" memiliki frekuensi tertinggi yakni sebanyak 379 yang tersebar di antara 6.027 korpus berita. Artinya, pada tahun 2018 kata "polisi" paling sering digunakan untuk memberikan informasi mengenai kejadian yang ada di Kota Surabaya oleh Tribunnews. Selanjutnya kata "anak" disebutkan sebanyak 264 kali, kata "warga" disebutkan 255 kali, kata "polda" disebutkan 226 kali, kata "mobil" disebutkan 193 kali, dan seterusnya. Banyaknya *feature* yang dihasilkan dari *tokenizing* dengan *unigram* adalah sebanyak 8.339 yang kemudian akan dipilih untuk pembentukan *cluster.* Hasil *feature selection* untuk *unigram* selangkapnya ditampilkan pada Lampiran 12.

Berikut merupakan sepuluh *features* dengan frekuensi tertinggi setelah melalui tahap *tokenizing* dengan *bigrams*.



Gambar 4.3 Sepuluh *Features* dengan Frekuensi Tertinggi (*Bigrams*)

Gambar 4.3 menunjukkan bahwa sepuluh *features* dengan frekuensi tertinggi setelah dilakukan pra proses adalah "universitas_airlangga", "kereta_api", "kepala_polda", "national_ hospital", "dokter_soetomo", "anak_anak", "tunjungan_plaza", "pelecehan_seksual", "ujian_nasional", dan "rsud_dokter". *Feature* "universitas_airlangga" memiliki frekuensi tertinggi yakni sebanyak 91 yang tersebar di antara 6.027 korpus berita. "universitas_airlangga" merupakan nama salah satu perguruan tinggi negeri yang ada di Surabaya. Nama perguruan tinggi negeri ini disebutkan sebanyak 91 kali di sepanjang tahun 2018 untuk memberikan banyak informasi mengenai peristiwa di Kota Surabaya yang berkaitan dengan Universitas Airlangga. Sellain itu, kata "kereta_api" disebutkan sebanyak 73 kali, kata "kepala_polda" dan "national_hospital" disebutkan 50 kali, "dokter_soetomo" disebutkan 48 kali, dan seterusnya. 26.875 *features* hasil *bigrams* akan dipilih untuk *text clustering* dan hasil *feature selection* pada *bigrams* terlampir pada Lampiran 13.

4.2 *Document Feature Matrices* dan TF-IDF

Feature yang terpilih selanjutnya akan dibentuk *document feature matrices* dan TF-IDF untuk mengubah struktur data teks menjadi numerik. Perhitungan TF-IDF diawali dengan menghitung frekuensi kemunculan setiap *feature* pada keseluruhan korpus berita atau *Document Frequency* (DF) seperti yang ditampilkan pada Tabel 4.8. Frekuensi kemunculan setiap *feature* kemudian distandardisasi menjadi *Inverse Document Frequency* (IDF). Berikut merupakan ilustrasi perhitungan DF-IDF untuk *tokenizing* dengan *unigram*.

Tabel 4.8 Ilustrasi Perhitungan DF-IDF (*Unigram*)

Docs	***Feature***				
	polisi	**warga**	**…**	**rumah**	**…**
1	0	0	…	0	…
…	…	….	…	…	…
700	1	0	…	0	…
701	0	1		1	…
…	…	…	…	…	…
2.071	1	0	…	0	…
2.072	0	0	…	1	…
…	…	…	…	…	…
2.093	0	0	…	1	…
2.094	0	1	…	0	…
…	…	…	…	…	…
DF	379	255	…	182	…
IDF	$log\left(\frac{6.027}{379}\right)$ = 1,201462	$log\left(\frac{6.027}{255}\right)$ = 1,373561	…	$log\left(\frac{6.027}{182}\right)$ = 1,52003	…

Dapat dilihat pada Tabel 4.8, DF untuk kata “polisi” adalah 379. Artinya, terdapat 379 dokumen berita yang memuat kata

"polisi". Selain itu, DF kata "warga" berjumlah 255 artinya terdapat 255 dokumen yang memuat kata "warga" serta DF kata "rumah adalah 182 yang berarti terdapat 182 dokumen berita yang memuat kata "rumah". Kemudian, pada setiap *feature* dilakukan perhitungan IDF menggunakan persamaan (2.1). Perhitungan DF-IDF yang sama juga dilakukan untuk DF-IDF hasil *tokenizing* dengan *bigrams* ditampilkan pada Tabel 4.9.

Tabel 4.9 Ilustrasi Perhitungan DF-IDF (*Bigrams*)

Docs	***Feature***				
	dinas_ pendidikan	**satpol_ pp**	**kebun_ binatang**	**…**	**apartemen_ educity**
1	0	0	0	…	0
…	…	…	…	…	…
5	1	0	0	…	1
…	…	…	…	…	…
11	0	1	0	…	1
…	…	…	…	…	…
20	0	0	1	…	1
21	0	0	0	…	0
22	0	0	1	…	1
…	…	…	…	…	…
5.993	1	0	0	…	1
…	…	…	…	…	…
DF	19	24	18	…	17
IDF	$log\left(\frac{6.027}{19}\right)$ = 2,501348	$log\left(\frac{6.027}{24}\right)$ = 2,39989	$log\left(\frac{6.027}{18}\right)$ = 2,524829	…	$log\left(\frac{6.027}{17}\right)$ = 2,549652

Nilai DF untuk kata "dinas_pendidikan" adalah sebesar 19. Artinya, terdapat 19 dokumen berita yang memuat kata "di-

nas_pendidikan". Kemudian IDF juga dihitung menggunakan persamaan (2.1) dan memberikan hasil sebesar 2,501348. Setelah menghitung DF-IDF, selanjutnya setiap *feature* diberi bobot TF-IDF dengan melakukan perhitungan menggunakan persamaan (2.1). Yaitu mengalikan TF setiap kata pada masing-masing korpus berita dengan nilai IDF dari *feature* yang akan diberi bobot. Hasil perhitungan TF-IDF pada masing-masing N-*Gram* yang disajikan dalam bentuk *data frame* pada Lampiran 14 dan Lampiran 15 diubah ke dalam bentuk matriks sebagai berikut.

Tabel 4.10 Matriks TF-IDF tiap N-*Grams*

Tokenizing	**Matriks TF-IDF**
Unigram	$\mathbf{U} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 1{,}960557 & \dots & 0 & \dots & 0 \\ 1{,}20376 & 0 & 0 & \dots & 1.840582 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \end{bmatrix}$
Bigrams	$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 2{,}146633 & \dots & 2{,}081131 & \dots & 0 \\ 0 & 2{,}39989 & 0 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 2{,}348737 & \dots & 0 \end{bmatrix}$

Matriks TF-IDF pada *tokenizing unigrams* berukuran 6.027 x 37, sedangkan untuk *tokenizing bigrams* memiliki matriks yang berukuran 6.027 x 40. Pada Tabel 4.11 dapat dilihat bahwa bobot TF-IDF pada *feature* "polisi" di korpus berita ke-4 adalah 1,20376. Sedangkan, bobot TF-IDF pada *feature* "kereta_api" adalah sebesar 2,39989 pada korpus berita ke-5. Selanjutnya, matriks TF-IDF yang dihasilkan akan menjadi input dalam *text clustering*.

4.3 Analisis Hasil *Text Clustering*

Text clustering pada dasarnya adalah mengelompokkan sekumpulan data teks yang memiliki karakteristik yang sama berdasarkan kriteria variabel atau *feature* tertentu. Pada sub bab ini akan dibahas hasil evaluasi pembentukan *cluster* pada korpus berita menggunakan metode K-*Means* dan *Self-Organizing Maps* (SOM). Ukuran evaluasi pembentukan *cluster* yang digunakan adalah *average silhouette width* pada masing-masing *k* dan N-*Grams*.

4.3.1 K-*Means*

Algoritma pembentukan *cluster* dengan metode K-*Means* dimulai dari menentukan jumlah *cluster* yang ingin dibentuk (*k*). Pada penelitian ini, jumlah *cluster* optimum ditentukan berdasarkan hasil evaluasi terbaik untuk jumlah *cluster* $k = 2$ hingga $k = 10$. Untuk jumlah $k = 2$ hingga $k = 10$, akan dievaluasi hasilnya berdasarkan nilai *average silhouette width*. Tabel 4.11 menampilkan hasil nilai *average silhouette width* untuk mengevaluasi *cluster* yang terbentuk pada *unigram* dan *bigrams*.

Tabel 4.11 Nilai *Average Silhouette Width* untuk Evaluasi Hasil *Clustering* dengan Metode K-*Means*

k	*Unigram*	*Bigrams*
2	0,4073249	0,7942249
3	0,4191991	0,7557231
4	0,4302132	0,7664045
5	0,4425845	0,7735838
6	0,4304674	0,7859037
7	0,3990316	0,786297
8	0,4315266	0,8040836
9	0,4434383	0,7974604
10	**0,4600947**	**0,8132585**

Setelah dilakukan percobaan pembentukan 2 hingga 10 klaster, nilai *average silhouette width* dibandingkan untuk dicari nilai yang paling besar. Semakin nilai *average silhouette width*

mendekati 1, menunjukkan bahwa semakin baik *cluster* yang terbentuk untuk mengelompokkan korpus berita. Dari Tabel 4.11, dapat dilihat bahwa nilai *average silhouette width* tertinggi pada *tokenizing unigram* dan *bigrams* adalah ketika $k = 10$ dengan nilai *average silhouette width* masing-masing sebesar 0,4600947 dan 0,8132585. Semakin tinggi N pada *tokenizing* akan meningkatkan nilai *average silhouette width* hampir dua kali lipat. Secara keseluruhan, nilai *average silhouette width* tertinggi adalah sebesar 0,8132585. Nilai ini dihasilkan untuk jumlah *cluster* optimum (k) sebanyak 10 *cluster* dengan input nilai TF-IDF ketika melakukan *tokenizing* dengan *bigrams*. Sehingga, dapat disimpulkan bahwa jumlah *cluster* optimum yang dapat dihasilkan dengan metode K-*Means* adalah sebanyak 10 *cluster* dengan *tokenizing bigrams*.

Setelah menentukan jumlah *cluster* yang akan dibentuk, selanjutnya menentukan inisiasi *centroids* awal. Pada penelitian ini, nilai *centroid* awal diambil dari 10 baris pertama pada matriks TF-IDF. Berikut ilustrasinya.

Tabel 4.12 Ilustrasi *Centroid* Awal

***Cluster* ke-**	**dinas_ pendidikan**	**satpol_ pp**	**…**	**tol_ sumo**	**apartemen_ educity**
1	0	0	…	0	0
2	0	0	…	0	0
3	0	0	…	0	0
4	0	0	…	0	0
5	2,5013476	0	…	0	0
…	…	…	…	…	…
10	0	0	…	0	0

Dari *centroid* awal tersebut, selanjutnya akan dilakukan perhitungan jarak *euclidean* setiap dokumen (menggunakan input matriks TF-IDF) dengan *centroid* pada *cluster* 1 hingga *cluster* 10 dengan menggunakan persamaan (2.2). Berikut ini adalah contoh perhitungan jarak *euclidean* pada *cluster* 1.

Tabel 4.13 Ilustrasi Perhitungan Jarak *Euclidean* (K-*Means*)

Jarak *Euclidean*	Perhitungan
$d_{(1,1)}$	$\sqrt{\lvert 0-0\rvert^2+\cdots+\lvert 0-0\rvert^2}$=0
...	...
$d_{(1,5)}$	$\sqrt{\lvert 0-2{,}5013476\rvert^2+\cdots+\lvert 0-0\rvert^2}$=2,501
...	...
$d_{(1,6.027)}$	$\sqrt{\lvert 0-0\rvert^2+\cdots+\lvert 0-0\rvert^2}$=0

Hasil perhitungan jarak *euclidean* pada setiap *cluster* dapat dilihat pada Lampiran 16. Selanjutnya adalah menentukan pengelompokkan data berdasarkan jarak *euclidean* minimum seperti yang ditampilkan pada Tabel 4.14.

Tabel 4.14 Pengelompokkan Korpus Berita

Docs	*Cluster* ke-					Keterangan	
	1	**…**	**5**	**…**	**10**	**Min $d_{(i,j)}$**	***Cluster* ke-**
1	0	…	2,501	…	0	0	1
2	0	…	2,501	…	0	0	1
3	0	…	2,501	…	0	0	1
4	0	…	2,501	…	0	0	1
5	2,501	…	0	…	2,501	0	5
6	0	…	2,501	…	0	0	1
…	…	…	…	…	…	…	…
6.027	0	…	2,501	…	0	0	1

Setelah mengelompokkan setiap korpus berita menurut jarak *euclidean*, selanjutnya adalah menghitung *centroid* baru untuk melakukan iterasi *clustering* selanjutnya. Dengan menggunakan persamaan (2.3), diperoleh nilai *centroid* baru sebagai berikut.

Tabel 4.15 Ilustrasi Hasil *Centroid* Baru untuk Iterasi ke-2

Cluster ke-	dinas_ pendidikan	satpol_ pp	…	tol_ sumo	apartemen_ educity
1	0	0.0095868	…	0.0072144	0.0072144

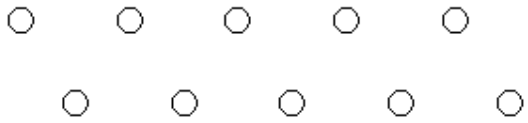
Tabel 4.15 Ilustrasi Hasil *Centroid* Baru untuk Iterasi ke-2 (lanjutan)

***Cluster* ke-**	**dinas_ pendidikan**	**satpol_ pp**	**…**	**tol_ sumo**	**apartemen_ educity**
2	0	0	…	0	0
…	…	…	…	…	…
5	2.5013476	0	…	0	0
…	…	…	…	…	…
10	0	0	…	0	0

Nilai *centroid* pada iterasi ke-2 selanjutnya digunakan untuk mengukur jarak *euclidean* antara korpus berita dengan *centroid* pada masing-masing *cluster*. Iterasi pada K-*Means* akan berhenti jika sudah mencapai fungsi objektif yang optimal ditandai dengan hasil pengelompokkan korpus berita (*clustering*) jika dibandingkan hasil pada iterasi sebelumnya tidak mengalami perubahan. Hasil *cluster* yang optimum inilah yang selanjutnya akan diukur evaluasinya seperti yang sudah ditampilkan pada Tabel 4.12.

4.3.2 *Self-Organizing Maps* (SOM)

Pada penelitian ini, *unsupervised learning* dengan metode SOM diterapkan dengan pendekatan dua dimensi. Oleh karena itu, topologi *grid* dua dimensi akan disiapkan terlebih dahulu. Pada penelitian ini akan digunakan topologi *grid* "*hexagonal*" dengan jumlah *grid* tidak memiliki ukuran lebih dari 10. Berikut adalah ilustrasi salah satu *grid* yang digunakan dalam *clustering* yaitu berukuran 5x2, artinya dimensi pada sumbu-*x* berukuran 5 satuan sedangkan dimensi pada sumbu-*y* berukuran 2 satuan.



Gambar 4.4 Insialisasi Grid pada SOM Dimensi 5 x 2

Pada SOM, setiap *vector input* akan dikelompokkan pada masing-masing *grid* berdasarkan bobot yang menghubungkan *vector input* dan *vector output*. Oleh karena itu, *clustering* pada SOM

diawali dengan menentukan inisiasi bobot. Bobot awal yang menghubungkan antara *neuron input* dengan *output* yaitu w_{jl} ditetapkan secara random antara 0 hingga 1. Selanjutnya, akan dipilih *vector input x* secara acak yang hendak diklasterkan dan di-*input*-kan ke *neuron input*. Sebagai contoh, *vector* input berupa korpus berita ke-1 akan diklasterkan menggunakan SOM. Berikut ilustrasinya.

Tabel 4.16 Ilustrasi Perhitungan Jarak *Euclidean* (SOM)

Jarak *Euclidean*	**Perhitungan**
$d_{(0,0)}$	$(-0{,}056-(0{,}722))^2+\cdots+\left(-0{,}053-(0{,}758)\right)^2 = 16{,}977$
$d_{(1,0)}$	$(-0{,}056-(0{,}119))^2+\cdots+\left(-0{,}053-(0{,}823)\right)^2 = 16{,}064$
$d_{(2,0)}$	$(-0{,}047-(0{,}162))^2+\cdots+\left(-0{,}053-(0{,}174)\right)^2 = 12{,}884$
…	…
$d_{(5,2)}$	$(-0{,}056-(-0{,}062))^2+\cdots+\left(-0{,}053-(0{,}232)\right)^2 = 13{,}371$

Dari perhitungan jarak *euclidean* pada Tabel 4.16 dan Lampiran 17, selanjutnya dapat diketahui bahwa jarak *Euclidean* minimum adalah sebesar 0,255 yang dihasilkan dari vektor bobot ke-2. Artinya, *best matching unit* untuk korpus berita pertama adalah pada *cluster* ke-2. Kemudian, dilakukan iterasi dengan cara memperbarui bobot koneksi yang ke-2 menggunakan persamaan 2.6. Berikut ilustrasi perhitungannya.

Tabel 4.17 Ilustrasi Perhitungan Pembaruan Bobot (SOM)

Bobot Baru	**Perhitungan**
$w_{(1,1,2)}(t+1)$	$0{,}163+0{,}05\left(-0{,}056-(0{,}163)\right)=0{,}152$
$w_{(2,1,2)}(t+1)$	$0{,}240+0{,}05\left(-0{,}063-(0{,}240)\right)=0{,}225$
…	…
$w_{(40,1,2)}(t+1)$	$0{,}174+0{,}05\left(-0{,}053-(0{,}174)\right)=0{,}163$

Langkah perhitungan seperti yang diuraikan pada Tabel 4.16 dan 4.17 dilakukan secara berulang dan iterasi akan berhenti apabila bobot sudah mencapai konvergensi. Hasil *clustering* pada setiap ukuran *grid* dan N-*Grams* selanjutnya akan diukur nilai *average silhouette width*-nya seperti pada Tabel 4.18 sebagai berikut.

Tabel 4.18 Nilai *Average Silhouette Width* untuk Evaluasi Hasil *Clustering* dengan Metode SOM

Ukuran Grid	*Unigram*	*Bigrams*
2x1	0,4591605	0,7387875
2x2	0,4006103	0,7186294
2x3	0,3718146	0,7336159
2x4	0,3823941	0,7381319
2x5	0,3901797	0,759079
3x1	0,4700456	0,7470631
3x2	0,4241915	0,7353732
3x3	0,3767859	0,7722931
4x1	**0,5018657**	0,7166075
4x2	0,3572944	0,7366435
5x1	0,349109	0,758765
5x2	0,4174652	**0,7919696**

Kriteria penentuan *cluster* optimum dengan metode SOM sama seperti metode K-*Means* yaitu berdasarkan nilai *average silhouette width*. Pada metode SOM, *tokenizing* dengan *unigram* memberikan nilai *average silhouette width* tertinggi pada ukuran *grid* 4x1 yaitu sebesar 0,5018657, sedangkan *tokenizing* dengan *bigrams* memberikan nilai *average silhouette width* tertinggi pada ukuran *grid* 5x2 yaitu sebesar 0,7919696. Berdasarkan hasil *average silhouette width* pada tiap *grid* yang diteliti, nilai *average silhouette width* tertinggi pada *grid* berukuran 5x2. Jadi, dapat disimpulkan bahwa jumlah *cluster* optimum yang dihasilkan dengan metode SOM adalah sebanyak 10 *cluster*.

4.4 Pemilihan Metode *Clustering* Optimum

Pembentukan *cluster* dengan metode K-*Means* dan SOM telah dibahas pada sub bab sebelumnya. Selanjutnya, hasil evalusasi kedua metode *clustering* tersebut dibandingkan berdasarkan nilai *average silhouette width*. Berikut adalah nilai *average silhouette width* terbaik yang diperoleh setelah membentuk *cluster* dengan metode K-*Means* dan SOM untuk masing-masing N-*Grams*.

Tabel 4.19 Perbandingan Hasil Evaluasi *Clustering* Metode K-*Means* dan SOM

Metode	*Tokenizing*	*k* Optimum	*Average Silhouette Width*
K-*Means*	*Bigrams*	10	**0,8132585**
SOM	*Bigrams*	10	0,7919696

Berdasarkan Tabel 4.20, nilai *average silhouette width* tertinggi yaitu sebesar 0,8132585 dihasilkan apabila pembentukan *cluster* dengan metode K-*Means* menggunakan pendekatan *tokenizing* dengan *bigrams*. Jumlah klaster optimum yang dihasilkan adalah 10 klaster. Oleh karena itu, metode *text clustering* yang digunakan untuk menentukan topik berita terbaik mengenai Kota Surabaya adalah dengan menggunakan metode K-*Means*.

4.5 Karakteristik *Cluster* yang Terbentuk

Setelah memperoleh jumlah *cluster* optimum, selanjutnya akan dilakukan pembentukan 10 *cluster* pada korpus berita dengan menggunakan metode K-*Means*. Berdasarkan pengelompokan korpus berita dengan menggunakan metode K-*Means* dan visualisasi *word cloud*, diperoleh rangkuman topik berita pada masing-masing *cluster* sebagai berikut.

Tabel 4.20 Topik Berita Hasil *Clustering* dengan K-*Means*

Cluster ke-	Frekuensi Berita	Topik Berita
1	65	Pelecehan Seksual
2	71	Kereta Api
3	86	Universitas Airlangga
4	17	Kepolisian Kabupaten/Kota
5	43	Narkoba

Tabel 4.20 Topik Berita Hasil *Clustering* dengan K-*Means* (lanjutan)

***Cluster* ke-**	**Frekuensi Berita**	**Topik Berita**
6	5.515	Hiburan, Kriminalitas, Peristiwa Penting, dan lain-lain.
7	50	RSUD dr. Soetomo
8	64	Kepolisian Daerah
9	33	Pelabuhan Tanjung Perak
10	83	Pendidikan

Berdasarkan rangkuman hasil penentuan topik berita menggunakan *text clustering* dengan K-*Means* pada Tabel 4.21 dapat dilihat bahwa korpus berita paling banyak terkelompokkan ke dalam klaster ke-6 yaitu membahas mengenai hiburan, kriminalitas, peristiwa penting, dan lain-lain. Topik berita yang paling banyak dibahas selanjutnya adalah berita yang berkaitan dengan Universitas Airlangga, yaitu sebanyak 86 berita. Sementara itu, topik yang paling sedikit dibahas yaitu mengenai pendidikan di Kota Surabaya yang hanya terdiri dari 17 judul berita. Selanjutnya, akan disajikan *word cloud* pada masing-masing *cluster* untuk mengetahui karakteristik pada masing-masing *cluster*.

Berikut merupakan visualisasi *word cloud* pada *cluster* 1

Gambar 4.5 *Word Cloud* Klaster 1

Gambar 4.5 menampilkan *word cloud* yang dihasilkan *cluster* 1. Menurut ukurannya, pada *word cloud* tersebut

didominasi oleh kata “national hospital” dan “pelecehan seksual”. Selain itu, terdapat pula kata-kata “seksual pasien”, “rs national”, “seksual national”, “mantan perawat”, “perawat national”, dan lain-lain. Dari kumpulan kata-kata tersebut dapat diambil kesimpulan bahwa kata-kata pada *cluster* 1 banyak berasal dari berita mengenai pelecehan seksual yang terjadi di rumah sakit di tahun 2018. Berita ini pertama kali dimuat di media pada tanggal 25 Januari 2018 yaitu tentang seorang perawat di Rumah Sakit National Hospital yang melecehkan pasiennya. Berita ini juga dimuat dalam media berita nasional seperti BBC.com, Tempo.co, Kompas.com, dan lain-lain.

Selanjutnya, akan ditampilkan *word cloud* klaster 2 pada Gambar 4.6 sebagai berikut.

Gambar 4.6 *Word Cloud* Klaster 2

Pada Gambar 4.6 dapat dilihat bahwa kata “kereta api” sering ditemukan pada berita yang terkelompokkan pada klaster 2. Selain itu, terdapat pula kata-kata seperti “perlintasan kereta”, “tiket kereta”, “tertabrak kereta”, “ditabrak kereta”, “kecelakaan

kereta", dan lain-lain. Ditemukan pula kata-kata yang menyebutkan secara spesifik jenis-jenis kereta yang ada di Indonesia seperti "api logawa", "sri tanjung", dan "api mutiara". Nama produk mobil juga ditampilkan pada *cluster* tersebut yaitu "pajero sport". Dari kumpulan kata-kata tersebut banyak mendeskripsikan berita yang berkaitan dengan perkeretaapian di Indonesia seperti diskon tiket kereta api dan kecelakaan antara kereta api dengan mobil *pajero sport*. Setiap bulannya, Tribunnews Surabaya selalu menerbitkan berita mengenai kereta api di Kota Surabaya, dimana berita mengenai kecelakaan banyak diberitakan pada bulan Oktober 2018. Berita mengenai kecelakaan kereta api ini juga sempat diberitakan di media berita *online* lainnya seperti Kompas.com dan Jawa Pos. Sehingga, dapat disimpulkan bahwa klaster dua membahas topik berita mengenai kereta api.

Selanjutnya, visualisasi *cluster* 3 akan ditampilkan pada Gambar 4.7.

Gambar 4.7 *Word Cloud* Klaster 3

Pada Gambar 4.7 banyak menyebutkan kata "universitas airlangga", "mahasiswa universitas", "rektor universitas", "dosen

universitas", "airlangga terima", dan lain-lain. Berdasarkan *word cloud* tersebut, dapat disimpulkan bahwa subjek yang sering disebutkan merupakan *civitas akademika* perguruan tinggi. Berita yang berkaitan dengan profil *civitas akademika* banyak diterbitkan di sepanjang tahun 2018 terutama yang berkaitan dengan salah satu perguruan tinggi negeri di Surabaya, Universitas Airlangga. Berita yang dibahas di antaranya adalah meninggalnya dosen Universitas Airlangga dan kasus bunuh diri dokter alumnus Universitas Airlangga pada bulan April 2018. Selain itu, banyak pula diberitakan mengenai hasil penelitian dan penemuan oleh mahasiswa Universitas Airlangga. Topik berita yang cocok untuk klaster 2 adalah mengenai Universitas Airlangga, dimana topik ini juga digunakan oleh portal berita *online* lainnya seperti detik.com, liputan6.com, dan lain-lain.

Berikut pada gambar 4.8 merupakan visualisasi *word cloud* dari *cluster* 4.

Gambar 4.8 *Word Cloud* Klaster 4

Klaster 4 terdiri dari 17 korpus berita yang banyak membahas Kapolres Tulungagung. Hal ini dapat ditunjukkan dari kata-

kata yang sering muncul yaitu “kepala polres”, “polres tulungagung”, “tol sumo”, dan lain-lain. Pada bulan September 2018, terjadi kecelakaan mobil yang berisikan keluarga Kapolres Tulungagung di Tol Sumo. Selain berita mengenai kecelakaan, berita terkait kepala polres lainnya adalah penipuan yang mengatasnamakan Kapolres Perak dan juga sertijab Kapolres pada bulan November 2018. Sehingga, topik berita yang cocok untuk menggambarkan klaster 4 adalah kepolisian kota/kabupaten. Berita yang berkaitan dengan kapolres juga banyak diliput oleh media berita lain seperti Kompas dan detik.com.

Berikut ini akan ditampilkan visualisasi *word cloud* untuk klaster 5.

Gambar 4.9 *Word Cloud* Klaster 5

Dari visualisasi *word cloud* pada Gambar 4.9 dapat dilihat bahwa kata-kata yang ditampilkan di antaranya adalah “sabu sabu”, “divonis penjara”, “gagalkan penyelundupan”, “hukuman penjara”, dan lain-lain. Kumpulan kata-kata tersebut mewakili 44 berita mengenai kasus narkoba dan sabu-sabu yang ada di Kota Surabaya selama tahun 2018. Sehingga, topik yang cocok untuk menggambarkan klaster 5 adalah mengenai narkoba. Berita mengenai narkoba di Surabaya juga aktif diberitakan oleh detik.com.

Selanjutnya, akan ditampilkan visualisasi topik pada klaster 6 sebagai berikut.

Gambar 4.10 *Word Cloud* Klaster 6

Gambar 4.10 menunjukkan *word cloud* dengan kata-kata yang memiliki frekuensi tertinggi pada klaster 6 yang terdiri dari 5.515 korpus berita. Hal ini berarti 91,5% korpus berita terkelompokkan ke dalam klaster 6. Dengan anggota klaster yang sangat dominan daripada klaster lain menjadikan klaster 6 bersifat unik daripada klaster lainnya sehingga diperlukan analisis deskriptif lebih mendalam mengenai karakteristiknya. Hasil analisis pada klaster 6 menunjukkan bahwa sebanyak 96 dari 104 judul berita atau sekitar 92,30% judul berita yang diawali kata "Breaking News" terkelompokkan pada klaster 6. Menurut Simbiosa dalam Kamus Jurnalistik yang diterbitkan tahun 2012, definisi "Breaking News" pada media *online* adalah berita penting dan baru yang diterima redaksi dan masih dapat dimuat di halaman muka, disisipkan, atau mengganti naskah yang sudah ada. Hal ini berarti pada klaster 6 memuat 96 judul berita yang bersifat sangat penting atau memiliki urgensi yang tinggi untuk segera diberitakan kepada masyarakat. Judul atau *headline* berita yang memuat kata "Breaking News" di antaranya adalah "BREAKING NEWS - Maling

Brankas di Tunjungan Plaza Surabaya Ditangkap" yang diterbitkan pada 30 Januari 2018, "BREAKING NEWS Mobil Toyota Innova Milik Pejabat Pemkot Surabaya Ditembak" pada 14 Maret 2018, serta "Breaking News - Wali Murid Demo Dindik Jatim Tuntut Transparansi Pagu PPDB SMA/SMK" yang diterbitkan pada 4 Juli 2018, dan lain-lain Sementara itu, frekuensi dari kata-kata yang terkandung dalam judul beriita dapat dilihat pada Tabel 4.21 sebagai berikut.

Tabel 4.21 Sepuluh *Feature* Berita yang Sering Dibahas pada Klaster 6

Feature	**Frekuensi Berita**	***Feature***	**Frekuensi Berita**
tunjungan plaza	40	harian surya	22
mobil pejabat	33	taman bungkul	21
bandara juanda	32	ditangkap polisi	20
satpol pp	27	asian games	19
royal plaza	24	pt kai	19

Dari kumpulan kata-kata pada *word cloud* dan Tabel 4.21 di atas dapat dilihat bahwa kata-kata tersebut seringkali berkaitan dengan lokasi atau keterangan tempat terjadinya berbagai peristiwa dan kejadian yang ada di Surabaya. Seperti kata "tunjungan plaza", "royal plaza", "kebun binatang", "taman bungkul", dan lain-lain. Kata "tunjungan plaza" disebutkan pada 40 berita, mobil pejabat disebutkan pada 33 berita, bandara juanda disebutkan pada 27 berita, dan lain-lain seperti yang terlampir pada Lampiran 23. Kata-kata pada klaster 6 didominasi oleh nama-nama tempat yang identik dengan Kota Surabaya seperti Tunjungan Plaza, Royal Plaza, Taman Bungkul, dan lain-lain. Nama tempat seringkali disebutkan dalam judul berita sebagai salah satu unsur 5W+1H yaitu "where" atau tempat terjadinya suatu peristiwa. Untuk mengetahui karakteristik lebih dalam lagi mengenai klaster 6 berikut akan disajikan judul berita yang berkaitan dengan dua pusat perbelanjaan di Surabaya yaitu Tunjungan Plaza dan Roya Plaza.

Tabel 4.22 Contoh Judul Berita yang Berkaitan dengan Dua Pusat Perbelanjaan Terbesar di Surabaya

***Feature* ata Frasa**	**Judul Berita**	**Tanggal**
Tunjungan Plaza	Kebakaran di Tunjungan Plaza 3 Surabaya, Diduga Akibat Konsleting Listrik	17 Januari 2018
	BREAKING NEWS - Maling Brankas di Tunjungan Plaza Surabaya Ditangkap	30 Januari 2018
	Libur Natal, Tunjungan Plaza Surabaya Ramai Pengunjung	24 Desember 2018
Royal Plaza	PT KAI Sudah Serahkan Spesifikasi Block Rel untuk Pecah Bottleneck di Royal Plaza Suraba	24 Januari 2018
	Terkait Kemacetan di Frontage Depan Royal Plaza, PT KAI: Silakan Pemkot Koordinasi dengan Kami	9 Februari 2018
	Proyek Bottle Neck Samping Royal Plaza 3 Kali Gagal Lelang, PT KAI: Sudah Sepenuhnya Wewenang Pemkot	23 Mei 2018

Dapat dilihat pada Tabel 4.22 di atas bahwa kata "Tunjungan Plaza" digunakan sebagai *headline* berita untuk menggambarkan peristiwa kebakaran, pencurian, dan juga hiburan. Selain itu, pada kata "Royal Plaza" ditemukan topik berita yang unik yang berkaitan dengan "PT KAI" yaitu proyek rekayasa lalu linas oleh PT. KAI di sekitar Royal Plaza. Hal ini menunjukkan bahwa frekuensi berita milik kata "Royal Plaza" dan "PT. KAI" saling beririsan sehingga dapat memperngaruhi pembentukan klaster. Oleh karena itu, karena klaster 6 bersifat unik, maka topik yang cocok untuk klaster 6 adalah hiburan, kriminalitas, peristiwa penting, dan lain-lain. Selain itu, banyak pula kejadian yang berkaitan dengan Kota Surabaya yang selain diliput oleh Tribun News, juga aktif diberitakan oleh media lain seperti Kompas dan detik.com. Selanjutnya akan ditampilkan *word cloud* klaster 7 sebagai berikut.

Gambar 4.11 *Word Cloud* Klaster 7

Kata-kata yang muncul pada Gambar 4.11 di antaranya adalah "dokter soetomo", "rsud dokter", "bayi kembar", "kembar siam", "dirawat rsud", dan lain-lain. Apabila melihat dari kata-kata yang dominan, dapat disimpulkan bahwa topik berita yang banyak dibahas berlokasi di Rumah Sakit Umum Daerah (RSUD) Dokter Soetomo. Banyak kejadian yang diberitakan berlokasi di RSUD dr. Soetomo. Di antaranya kejadian bayi kembar siam pada bulan Januari, April dan Mei 2018; pasien korban kecelakaan dan miras oplosan yang dirawat di rumah sakit tersebut pada bulan April dan Agustus 2018; serta berbagai kegiatan edukasi kesehatan yang dilaksanakan oleh pihak manajemen RSUD dr. Soetomo. Berita mengenai RSUD dr. Soetomo dapat dikatakan tidak terlalu sering diliput oleh media berita nasional lainnya. Tetapi Tribun News aktif memberikan berita mengenai rumah sakit rujukan bagi kawasan di Pulau Jawa wilayah Timur ini. Untuk mengetahui topik berita pada klaster 8, berikut ini akan disajikan *word cloud* pada Gambar 4.12.

Gambar 4.12 *Word Cloud* Klaster 8

Gambar 4.12 menampilkan *word cloud* untuk *cluster* 8. *Word cloud* tersebut berisi kata-kata di antaranya "kepala polda", "tabrak lari", "pengemudi tabrak", "polda irjen", "polda pangdam", dan lain-lain. Kata-kata ini banyak menyebutkan sebuah jabatan di instansi kepolisian yaitu kepala polda. Dalam hal ini, kepala polda mengacu pada Kepala Polisi Daerah Jawa Timur. Sebagai orang nomor satu di kepolisian Provinsi Jawa Timur, kapolda tidak luput dari bahan pemberitaan. Terbukti selama tahun 2018, sebanyak 64 berita yang diterbitkan selama tahun 2018 dikelompokkan ke dalam *cluster* yang sama dengan menyebut kata kapolda. Sehingga dapat disimpulkan bahwa topik yang tepat untuk klaster 8 adalah kepolisian daerah. Topik berita mengenai kepolisian daerah khususnya Kapolda Jawa Timur juga sering diberitakan oleh Kompas.com dan detik.com. Selanjutnya, *word cloud* klaster 9 akan ditampilkan pada Gambar 4.13.

Gambar 4.13 *Word Cloud* Klaster 9

Word cloud di atas menujukkan kata "tanjung perak", "pelabuhan tanjung", "polres pelabuhan", "kepala polres", dan lain-lain. Kata yang paling dominan adalah "tanjung perak" yaitu nama sebuah pelabuhan yang terletak di Surabaya Utara. Pelabuhan Tanjung Perak adalah pelabuhan domestik dan juga internasional. Berita yang melibatkan Pelabuhan Tanjung Perak diantaranya adalah kasus penyelundupan sabu pada Januari 2018, kasus imigrasi yang tidak sesuai prosedur, dan juga masalah pelayaran. Terdapat 33 berita dengan kata kunci "tanjung perak" yang diterbitkan selama tahun 2018. Selain Tribunnews Surabaya, berita mengenai Pelabuhan Tanjung Perak juga sering dipublikasi-kan oleh Kompas.com, liputan6.com, detik.com, dan lain-lain. Terakhir, *word cloud* untuk klaster 10 akan ditampilkan sebagai berikut.

Gambar 4.14 *Word Cloud* Klaster 10

Dari Gambar 4.14 dapat dilihat bahwa kata-kata yang sering muncul di antaranya adalah "ujian nasional", "anak anak", "kecurangan ujian", "pelaksanaan ujian", "dinas pendidikan", dan lain-lain. Pada klaster 10, berita yang paling unggul dibicarakan adalah mengenai ujian nasional. Pada ujian nasional tahun 2018, terbukti mantan kepala sekolah sebuah SMP Negeri di Surabaya melakukan kecurangan dengan cara membocorkan naskah soal. Selain itu, berita mengenai kegiatan yang dilakukan anak-anak di Surabaya juga banyak dibahas terutama kegiatan anak-anak panti asuhan. Sehingga dapat disimpulkan bahwa topik klaster 10 membahas mengenai pendidikan di Kota Surabaya. Selain Tribunnews Surabaya, berita mengenai pendidikan khususnya mengenai anak-anak dan ujian nasional juga diliput oleh Kompas.com dan liputan 6.com.

(*Halaman ini sengaja dikosongkan*)

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis dan pembahasan yang telah diuraikan pada bab sebelumnya, diperoleh kesimpulan sebagai berikut.

1. Eksplorasi awal terhadap data berita memberikan hasil bahwa jumlah berita yang diterbitkan oleh Tribunnews Surabaya pada setiap bulannya di tahun 2018 mengalami fluktuatif, dimana jumlah berita paling banyak diterbitkan adalah pada bulan November yaitu 612 berita. Selain itu, rata-rata jumlah berita yang diterbitkan setiap bulannya adalah sebanyak 502 artikel berita. Kemudian, setelah melalui pra proses hingga *feature selection* diperoleh hasil jumlah *feature* untuk *unigram* adalah 37 dan *bigram* adalah 40 *features* yang selanjutnya akan dibentuk matriks TF-IDF sebagai input untuk pembentukan *text clustering*.
2. *Tokenizing* dengan *bigrams* terbukti dapat meningkatkan nilai *average silhouette width* hampir dua kali lipat dibandingkan dengan *unigram*. Berdasarkan hasil *text clustering* dengan metode K-*Means* pada $k = 2$ hingga $k = 10$ dan SOM pada ukuran *grid* tidak lebih dari 10, diperoleh jumlah *cluster* optimum adalah sebanyak 10 klaster dengan nilai *average silhouette width* tertinggi sebesar 0,8132585 yang dihasilkan dari metode K-*Means* menggunakan *tokenizing bigrams*.
3. Jumlah anggota *cluster* terbanyak adalah pada *cluster* 6 yang terdiri dari 5.515 korpus berita. Klaster ini membahas berita mengenai hiburan, kriminalitas, peristiwa penting, dan lain-lain yang terjadi di sekitar Surabaya. Topik berita yang paling banyak dibahas selanjutnya adalah mengenai Universitas Airlangga, yaitu sebanyak 86 berita. Sementara itu, topik yang paling sedikit dibahas yaitu sebanyak 17 judul berita yang mengenai kepolisian kabupaten/kota.

5.2 Saran

Berdasarkan penelitian yang telah dilakukan, saran yang dapat diberikan bagi manajemen Tribunnews Surabaya berupa algortima beserta topik berita yang telah dikelompokkan agar masyarakat semakin mudah memperoleh informasi yang ingin dicari. Saran yang dapat diberikan kepada pemerintah harapannya dapat mengevaluasi kebijakan yang sudah ada terkait peristiwa yang sering terjadi di masyarakat. Selain itu, pada penelitian selanjutnya yang berkaitan dengan *text clustering* diharapkan lebih memperhatikan lagi mengenai *feature selection* agar diperoleh hasil *cluster* yang lebih baik.

DAFTAR PUSTAKA

Agusta, Y. (2007). K-Means - Penerapan, Permasalahan dan Metode Terkait. *Jurnal Sistem dan Informatika, 3*, 47-60.

Ambarwati dan Winarko, E. (2014). Pengelompokan Berita Indonesia Berdasarkan Histogram Kata Menggunakan Self-Organizing Map. *IJCCS, 8*(1), 101-110.

Asosiasi Penyelenggara Jasa Internet Indonesia. (2018). *Survei APJII: Penetrasi di Indonesia Capai 143 Juta Jiwa.* Jakarta: APJII.

Azam, N. dan Yao, J. T. (2012). Comparison of Term Frequency and Document Frequency Based Feature Selection Metrics in Text Categorization. *Expert Systems with Applications, 39*(5), 4760-4768.

Castella, Q. dan Sutton, C. (2014). Word Storms: Multiples of Word Clouds for Visual Comparison of Documents. Seoul: International World Wide Web Conference Committee (IW3C2). Diakses pada tanggal 19 Februari 2019.

Cavnar, W. B. dan Trenkle, J. M. (1994). N-Gram-Based Text Categorization. *Proceedings of SDAIR-94, 3rd Annual Symposium an Document Analysis and Information Retrieval.* TREC-2.

CRAN. (2018). *Package 'tm'.* Diakses pada tanggal 19 Februari 2019 dari Cran-r-project: http://tm.r-forge.r-project.org/

Dumbill, E. (2014). *Planning For Big Data.* Sebastopol: O'Reilly Media, Inc.

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

Furnkranz, J. (1994). *A Study Using n-gram Features for Text Categorization.* Technical Report OEFAI-TR-98-30, Austrian Research Institute, Artificial Intelligence.

Gullo, F. (2015). From Patterns in Data to Knowledge Discovery: What Data Mining Can Do. *Physics Procedia, 62*, 18-22.

Guthikonda, S. M. (2009). *Kohonen Self-Organizing Maps.* Springfield: Wittenberg University. Diakses pada tanggal 18 Februari 2019.

Hughes, K. (2001). *Wajah PERS Malang.* Malang: Fakultas Ilmu Sosial dan Ilmu Politik Universitas Muhammadiyah Malang.

Internet World Statistics. (2017). *Internet World Statistics*. Diakses pada tanggal 21 Februari 2019 dari Internet World Statistics: http://www.internetworldstats.com

Jiawei, H., Kamber, M. dan Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Waltham, MA: Morgan Kaufmann.

Johnson, F. dan Gupta, S. (2012). Web Content Mining Techniques: A Survey. *Int. J. Comput. Appl., 47*(11), 44-50.

Koloway, B. C. (2019). *Tribunnews Malang*. Diakses pada tanggal 15 Juni 2019 dari Harian Surya Raih Silver Winner Surat Kabar Terbaik Regional Jawa IPMA 2019: https://suryamalang.tribunnews.com/2019/02/08/harian-surya-raih-silver-winner-surat-kabar-terbaik-regional-jawa-ipma-2019

Kumar, V. dan Minz, S. (2014). Feature Selection: A Literarure Review. *Smart Computing Review, 4*(3), 211-229.

Langgeni, D. P., Baizal, Z. A. dan W., Yanuar Firadus A. (2010). Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection. *semnasIF2010*.

Liao, S.-H., Chu, P.-H. dan Hsiao, P.-Y. (2012). Data mining techniques and applications-a decade review from 2000 to 2001. *Expert Systems with Applications, 39*(12), 11303-11311.

Linoff, G. S. dan Berry, M. J. (2011). *Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management* (3rd ed.). Indianapolis: Wiley Publishing, Inc.

Lubis, M. (2017). *Nielsen.* Diakses pada tanggal 21 Februari 2019 dari Media Cetak Mampu Mempertahankan Posisinya: https://www.nielsen.com/id/en/press-room/2017/media-cetak-mampu-mempertahankan-posisinya.print.html

Miner, G., Delen, D., Elder, J., Fast, A., Hill, T. dan Nisbet, R. (2012). *Practical Text Mining and Statistical Analaysis for Non-Structured Text Data Applications.* Oxford: Elseiver.

PT Tribun Digital Online. (2019). *SURYA.co.id*. Diakses pada tanggal 1 Maret 2019 dari Halaman Depan SURYA.co.id: http://surabaya.tribunnews.com/.

Putri, M. M. dan Fithriasari, K. (2015). Pengelompokkan Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Kesehatan Masyarakat Menggunakan Metode Kohonen SOM dan K-Means. *Jurnal Sains dan Seni ITS, 4*(1), D13 - D18.

Riyadi, M. A., Pratiwi, D. S., Irawan, A. R. dan Fithriasari, K. (2017). Clustering Stationary and Non-Stationary Time Series Based on Autocorrelation Distance of Hierarchical and K-Means Algorithms. *International Journal of Advances in Intelligent Informatics, 3*(3), 154-160.

Setiawan, K. (2003). *Paradigma Sistem Cerdas.* Malang: Bayumedia Publishing.

Soehoet, D. A. (2006). *Dasar-dasar Jurnalistik.* Jakarta: Yayasan Kampus Tercita-IISIP.

Statsoft. (2015). *Text Mining (Big Data, Unstructured Data).* Diakses pada tanggal 15 Februari 2019 dari Text Mining: http://www.statsoft.com/textbook/text-mining

Suh, J. H., Park, C. H. dan Jeon, S. H. (2010). Applying Text and Data Mining Techniques to Forecasting the Trend of Petitions Filed to e-People. *Expert Systems with Applications*(37), 7255-7268.

Sumadiria, H. (2011). *Jurnalistik Indonesia, Menulis Berita Dan Feature, Panduan Praktis Jurnalis Profesional.* Jakarta: Simbiosa Rekatama Media.

The MathWorks, Inc. (2019). *Analyze Text Data Using Multiword Phrases.* Diakses pada tanggal 20 Februari 2019 dari MathWorks: https://www.mathworks.com /help/textanalytics/ug/analyze-text-data-using-multi-word-phrases.html

Tribunnews. (2016). *Tribunnews.com*. Diakses pada tanggal 17 Februari dari Halaman Depan Tribunnews.com: http://www.tribunnews.com/

Turland, M. (2010). *PHP-Architect's Guide to Web Scraping.* Canada: Marco Tabini & Associates.

LAMPIRAN

Lampiran 1. *Syntax Crawling* Data Berita Tribunnews Surabaya Tahun 2018

```
#Scraping
library(rvest)
library(purrr)

url_base <- "http://surabaya.tribunnews.com/topic/berita-sura-
baya?&page=%d"
map_df(40:241, function(i) {
  cat(".")
  pg <- read_html(sprintf(url_base, i))
  data.frame(Title = gsub("\n\t\t\t\t\t\t",
                          "",
                          html_text(html_nodes(pg, ".fbo"))),
                          stringsAsFactors=FALSE)
}) -> Judul

dplyr::glimpse(Judul)

#Simpan Data Scraping
write.csv(Judul, file = "D:/Judul Berita.csv")

#Cleaning HTML
Judul <- read.csv("D:/Judul Berita.csv", header=TRUE)
str(Judul)
seq()
n <- nrow(Judul)
a <- seq(seq(1:3), n, 30)
dat[ toDelete ,]
```

Lampiran 2 Data Judul Berita Tribunnews Surabaya Tahun 2018

Judul ke-	Tanggal	Judul Berita
1	01/01/2018	Fenomena Supermoon 2 Januari 2018, Surabaya Sebaiknya Waspadai Banjir Rob
2	01/01/2018	Besok, Dispendukcapil Surabaya Gelar Yustisi untuk Antisipasi Urbanisasi
3	01/01/2018	Perayaan Malam Tahun Baru di Surabaya Hasilkan Sampah 6 Ton
4	01/01/2018	Polisi Tilang 311 Pengendara Motor di Malam Tahun Baru di Surabaya
5	01/01/2018	Klarifikasi Dinas Pendidikan Jatim Soal Surat Siswa Lamongan ke Ahok yang Jadi Viral
6	01/01/2018	Libur Tahun Baru, KBS Dipadati 52 Ribu Pengunjung
.	.	.
.	.	.
.	.	.
6021	31/12/2018	Alasan Risma Tak Hidupkan Air Mancur Menari di Jembatan Surabaya pada Malam Tahun Baru 2019
6022	31/12/2018	Mahasiswa Unair Lakukan Penelitian Bersama Mahasiswa Master dan PhD di Taiwan
6023	31/12/2018	Camat Jambangan Lapor Polisi Setelah Dtuduh Pungut Uang Lingkungan Pabrik Kertas di Surabaya
6024	31/12/2018	Tahun Baru, 9 Pelaku Kasus Aneka Perjudian Beromzet Ratusan Juta di Jawa Timur Ditangkap
6025	31/12/2018	Hingga Akhir Tahun 2018, Seperti Ini Progress Proyek MERR Gununganyar Tembus Tol Juanda
6026	31/12/2018	Daftar Kasus Menonjol di Surabaya Versi Polisi, Peledakan Bom hingga Perdagangan Bayi di Instagram
6027	31/12/2018	Malam Pergantian Tahun Baru, Cuaca Kota Surabaya Diprediksi Cerah Berawan

Lampiran 3 Syntax Pre-processing Text

```
#Menyiapkan Libary
library("tm")
library("stringr")
library("quanteda")
library("RWeka")
library("kohonen")
library("som")
library("cluster")
library("fpc")
library("factoextra")
library("cluster")
library("kohonen")
library("som")
library("readtext")
library("corpus")
library("Matrix")
library("textreg")
library("grid")
library("RColorBrewer")
library("ggplot2")
library("wordcloud")

#Memanggil Data
docs <- read.csv("D:/Database Fix","", header=TRUE)
docs <- data.frame(docs)
docs$Date <- as.character(docs$Date)
docs$Date <- as.Date.character(docs$Date, format="%d/%m/%Y")
docs <- as.vector(as.character(unlist(docs[,3])))

#PRE PROCESSING
docs <- VCorpus(VectorSource(docs))

docs <- tm_map(docs, content_transformer(tolower))

removePunct <- function(x) gsub("[[:punct:]]+", " ", x)
docs <- tm_map(docs, content_transformer(removePunct))
```

Lampiran 3 *Pre-processing Text* (lanjutan)

```
removeNum <- function(x) gsub("[[:digit:]]+", "", x)
docs <- tm_map(docs, content_transformer(removeNum))

stopwords_id <- read.table('D:/stopwords-id.txt', header = FALSE)
docs <- tm_map(docs, removeWords, stopwords_id$V1)

stopwords_id2 <- read.table('D:/stopwordsfff2.txt', header = FALSE)
docs <- tm_map(docs, removeWords, stopwords_id2$V1)

docs <- tm_map(docs, stripWhitespace)

docs.char <- convert.tm.to.character(docs) #character

docs.char <- str_replace_all(
  docs.char, # column we want to search
  c( " bocah "=" anak ",
     " kapolres "=" kepala polres ",
     " kapolda "=" kepala polda ",
     " dr "=" dokter ",
     " unbk "=" ujian nasional ",
     " unair "=" universitas airlangga ",
     " ka "=" kereta api ",
     " jl "=" jalan ",
     " e "=" elektronik "
     )
)

docs <- VCorpus(VectorSource(docs.char)) #corpus

docs.char <- convert.tm.to.character(docs) #character
```

Lampiran 4 *Tokenizing* dan *Feature Selection* (*Unigram*)

```
#Tokenizing
unigram.dfm <- dfm(docs.char, ngrams = 1)

#Feature Selection
unigram.dfm <- dfm_trim(unigram.dfm, sparsity = 0.99)

#Characteristics
head(unigram.dfm, n = 5, nf = 15)
str(unigram.dfm)
length(featnames(unigram.dfm))

#TFIDF
unigram.tfidf <- dfm_tfidf(unigram.dfm, scheme_tf = "count",
                           scheme_df = "inverse", base = 10, force = FALSE)
unigram.tfidf <- as.matrix(unigram.tfidf)

#Simpan
write.table(convert(unigram.dfm, to ="data.frame"), "D:/DFM unigram.txt")
write.table(unigram.tfidf, "D:/TFIDF unigram.txt")
```

Lampiran 5 *Tokenizing* dan *Feature Selection* (*Bigrams*)

```
#Tokenizing
bigram.dfm <- dfm(docs.char, ngrams = 2)

#Feature Selection
bigram.dfm <- dfm_trim(bigram.dfm, min_termfreq = 15, min_docfreq = 2)

#Characteristics
head(bigram.dfm, n = 2, nf = 4)
str(bigram.dfm)
length(featnames(bigram.dfm))
```

Lampiran 5 *Tokenizing* dan *Feature Selection* (*Bigrams*) (lanjutan)

```
#TFIDF
bigram.tfidf <- dfm_tfidf(bigram.dfm,
                          scheme_tf = "count",
                          scheme_df = "inverse",
                          base = 10, force = FALSE)
bigram.tfidf <- as.matrix(bigram.tfidf)

#Simpan
write.table(convert(bigram.dfm, to ="data.frame"), "D:/DFM bigram.txt")
```

Lampiran 6 *Text Clustering* dengan K-*Means*

```
#Unigram
Avg.Sil <- list()
for (ncl in 2:10){
KMC <- kmeans(unigram.tfidf, centers = ncl, nstart = 1000)
Coef.Sil <- silhouette(KMC$cluster, dist(unigram.tfidf))
Avg.Sil[ncl] <- summary(Coef.Sil)$avg.width
}
Avg.Sil

#Bigrams
Avg.Sil2 <- list()
For (ncl in 2:10){
KMC.Bigram <- kmeans(bigram.tfidf, centers = ncl, nstart = 1000)
Coef.Sil2 <- silhouette(KMC.Bigram$cluster, dist(bigram.tfidf))
Avg.Sil2[ncl] <- summary(Coef.Sil2)$avg.width
}
Avg.Sil2
```

Lampiran 7 *Text Clustering* dengan SOM

```
#Unigram
unigram.tfidf <- as.matrix(scale(unigram.tfidf))

som_grid <- somgrid(xdim = 2, ydim=1, topo="hexagonal")
som.model1 <- supersom(unigram.tfidf, grid=som_grid, rlen=100,
alpha=c(0.05, 0.01), keep.data=TRUE)
Coef.Sil.SOM1 <- silhouette(som.model1$unit.classif,
dist(unigram.tfidf))
summary(Coef.Sil.SOM1)$avg.width


#Bigram
bigram.tfidf <- as.matrix(scale(bigram.tfidf))

som_grid <- somgrid(xdim = 2, ydim = 1, topo ="hexagonal")
som.model2 <- supersom(bigram.tfidf,
grid=som_grid, rlen=100,
alpha=c(0.05, 0.01),
keep.data=TRUE)
Coef.Sil.SOM2 <- silhouette(som.model2$unit.classif,
dist(bigram.tfidf))
summary(Coef.Sil.SOM2)$avg.width
```

Lampiran 8 Visualisasi Karakteristik Data Awal

```
freq.months <- table(months(docs$Date))
z <- as.data.frame(freq.months)
ggplot(z,
       aes(x = Var1, y = Freq)) +
       geom_bar(stat="identity", fill = "hotpink2") +
       xlab("Months") +
       ylab("News") +
       geom_text(aes(x = Var1, y = Freq, label = Freq), vjust = 2) +
       ylim(0, 800) +
       theme(axis.text.x=element_text(angle=45, hjust=1))
summary(z)

dev.off()

freq.unigram <- textstat_frequency(unigram.dfm, n = 10)
freq.unigram$feature <- with(freq.unigram, reorder(feature, -frequency))
ggplot(freq.unigram, aes(x = feature, y = frequency)) +
  geom_bar(stat="identity", fill = "lightblue") +
           xlab("Feature") +
           ylab("Frequency") +
  geom_text(aes(x = feature, y = frequency, label = frequency), vjust = -0.5) +
  theme(axis.text.x=element_text(angle=45, hjust=1))
```

Lampiran 9 Visualisasi Karakteristik Data Awal (lanjutan)

```
freq.bigram <- textstat_frequency(bigram.dfm, n = 10)
freq.bigram$feature <- with(freq.bigram, reorder(feature, -frequency))
ggplot(freq.bigram, aes(x = feature, y = frequency)) +
  geom_bar(stat="identity", fill = "lightgreen") + xlab("Feature") +
ylab("Frequency") +
  geom_text(aes(x = feature, y = frequency, label = frequency), vjust = -0.5)+
  theme(axis.text.x=element_text(angle=45, hjust=1))
```

Lampiran 10 Visualisasi *Word Cloud*

```
clust.bigram <- as.data.frame(KMC.Bigram$cluster)

# Creating list of clusters with their files

cl2 = list()
for (i in 1:ncl) {
  cl2[paste("cl_",i, sep = "")] = list(rownames(subset(clust.bigram,
                                                       clust.bigram == i)))
}
cl2

# Creating corpuses for each cluster
for (i in 1:ncl) {
  name = paste("cl_corp_", i, sep = "")
  assign(name, docs[match(cl2[[i]], names(docs))])
}
```

Lampiran 11 Visualisasi *Word Cloud* (lanjutan)

```
dtm.bigram = list()

# Creating a list of DTMs for each cluster
for (i in 1:ncl) {
  BigramTokenizer <- function(x) NGramTokenizer(x, Weka_control(min = 2,
                                                                max = 2))
  bigram_dtm_i = DocumentTermMatrix(get(paste("cl_corp_",i,sep="")),
                                    control = list(tokenize =
                                    BigramTokenizer))
  dtmb_i <- as.matrix(bigram_dtm_i)
  dtm.bigram[paste("cluster_",i,sep="")] = list(dtmb_i)
}

dtm.bigram

# Ploting most common words in each cluster

for (i in 1:ncl) {
  cl_mb <- as.matrix(dtm.bigram[[i]])
  freq.term <- sort(colSums(cl_mb), decreasing=TRUE)
  wordcloud(words = names(freq.term),
            freq = freq.term,
            min.freq = 2,
            height = 15,
            width = 15,
            res = 300,
            random.order = T,
            colors = brewer.pal(8, "Dark2"),
            main = paste("Most common words for cluster", i, sep = " "))
}
```

Lampiran 12 *Feature* Terpilih pada *Unigram*

polisi	plaza	terbakar
warga	rumah	anak
polrestabes	hotel	pt
polda	pasar	mahasiswa
pelajar	its	sd
sma	gratis	mantan
mobil	kereta	universitas
pemuda	api	kebakaran
sabu	kepala	pasien
ditangkap	bus	tersangka
tewas	nasional	airlangga
dokter	dprd	
tunjungan	smp	

Lampiran 13 *Feature* Terpilih pada *Bigrams*

dinas_pendidikan	rsud_dokter	anak_muda
satpol_pp	sma_smk	ketua_dprd
kebun_binatang	kembar_siam	taman_bungkul
sabu_sabu	rumah_sakit	ujian_nasional
ditangkap_polisi	tanjung_perak	pelecehan_seksual
dokter_soetomo	royal_plaza	national_hospital
asian_games	mahasiswa_universitas	uk_petra
tunjungan_plaza	kepala_polres	tabrak_lari
bandara_juanda	pelabuhan_tanjung	meninggal_dunia
anak_anak	pt_kai	mahasiswa_its
kereta_api	kamar_kos	tol_sumo
harian_surya	kepala_polda	apartemen_educity
universitas_airlangga	mobil_pejabat	
divonis_penjara	kecanduan_seks	

Lampiran 14 TF-IDF pada *Unigram*

Docs	**polisi**	**warga**	**…**	**pasien**	**…**
1	0	0	…	0	…
…	…	…	…	…	…
628	1,20376	0	…	1,98771	…
629	0	0	…	0	…
630	0	1,376981	…	0	…
631	0	0	…	0	…
632	0	0	…	0	…
633	1,20376	0	…	0	…
634	0	0	…	0	…
635	1,20376	0	…	0	…
…	…	…	…	…	…
787	0	1,376981	…	0	…
788	1,20376	0	…	0	…
789	0	0	…	0	…
790	1,20376	0	…	0	…
…	…	…	…	…	…
1.255	0	0	…	1,98771	…
1.256	0	0	…	1,98771	…
1.257	0	0	…	0	…
1.258	0	0	…	1,98771	…
…	0	0	…	0	…
6.026	1,20376	0	…	0	…
6.027	0	0	…	0	…

Lampiran 15 TF-IDF pada *Bigrams*

Docs	**dinas_ pendidikan**	**dokter_ soetomo**	**…**	**ujian_ nasional**	**…**
1	0	0	…	0	…
…	…	…	…	…	…
5	2,501348	0	…	0	…
…	…	…	…	…	…
1.929	0	2,09886	…	0	…
1.930	0	2,09886	…	0	…
1.931	0	0	…	0	…
1.932	0	0	…	0	…
1.933	0	0	…	0	…
1.934	0	0	…	0	…
1.935	0	2,09886	…	0	…
…	…	…	…	…	…
2.042	2,501348	0	…	2,15685	…
2.043	0	0	…	0	…
2.044	0	0	…	0	…
2.045	0	0	…	0	…
2.046	2,501348	0	…	2,15685	…
…	…	…	…	…	…
5.532	2,501348	0	…	2,15685	
…	…	…	…	…	…
5.560	2,501348	0	0	0	
…	…	…	…	…	…
6.027	0	0	…	0	…

Lampiran 16 Jarak *Euclidean* Iterasi 1 (K-*Means*)

Docs	d1	d2	…	d5	…	d9	d10
1	0	0	…	2,501	…	0	0
2	0	0	…	2,501	…	0	0
3	0	0	…	2,501	…	0	0
4	0	0	…	2,501	…	0	0
5	2,501	2,501	…	0	…	2,501	2,501
6	0	0	…	2,501	…	0	0
7	0	0	…	2,501	…	0	0
8	0	0	…	2,501	…	0	0
9	0	0	…	2,501	…	0	0
10	0	0	…	2,501	…	0	0
11	2,400	2,400	…	3.466	…	2,400	2,400
12	0	0	…	2,501	…	0	0
13	0	0	…	2,501	…	0	0
14	0	0	…	2,501	…	0	0
15	0	0	…	2,501	…	0	0
…	…	…	…	…	…	…	…
6.020	3,967	3,967	…	4,690	…	3,967	3,967
6.021	0	0	…	2,501	…	0	0
6.022	3,192	3,192	…	4.055	…	3,192	3,192
6.023	0	0	…	2,501	…	0	0
6.024	0	0	…	2,501	…	0	0
6.025	0	0	…	2,501	…	0	0
6.026	0	0	…	2,501	…	0	0
6.027	0	0	…	2,5013476	…	0	0

Lampiran 17 Jarak *Euclidean* Iterasi 1 (SOM)

	dinas_ pendidi-kan	sat-pol_ pp	…	maha-siswa_ its	…	aparte-men_ educity	Total
	-0,06	-0,06	…	-0,05	…	-0,05	
	Distance						
d1	0,00	252,09	…	0,00	…	0,00	252,32
d2	0,00	0,00	…	0,00	…	0,04	0,26
d3	0,00	0,00	…	30,53	…	0,00	119,76
d4	0,00	0,00	…	0,00	…	0,00	150,07
d5	0,00	0,00	…	0,00	…	0,00	268,26
d6	0,00	0,00	…	0,00	…	0,00	249,67
d7	0,00	0,00	…	0,00	…	0,00	121,32
d8	0,00	0,00	…	0,00	…	0,00	129,78
d9	6,84	0,00	…	0,00	…	0,00	150,79
d10	0,00	0,00	…	0,00	…	0,00	77,83

Lampiran 18 *Feature* pada Klaster 1

No	*Feature*	Frekuensi Berita
1	national hospital	49
2	pelecehan seksual	42
3	perawat national	12
4	mantan perawat	11
5	seksual pasien	11
6	rs national	10
7	seksual national	10
8	dilaporkan pelecehan	6
9	pasien national	6
10	pasien rs	6
11	perawat pelecehan	6
12	seksual perawat	6
13	pasien cantik	5
14	pelecehan pasien	5
15	perawat rs	5
16	oknum perawat	4
17	absen sidang	3
18	hospital pelecehan	3
19	lecehkan pasien	3
20	perawat dilaporkan	3
21	perawat pasien	3
…	…	…
233	vonis mantan	1

Lampiran 19 *Feature* pada Klaster 2

No	*Feature*	Frekuensi Berita
1	kereta api	71
2	perlintasan kereta	8
3	tiket kereta	8
4	ditabrak kereta	7
5	pajero sport	6
6	tertabrak kereta	6
7	disambar kereta	5
8	kecelakaan kereta	5
9	api perlintasan	4
10	api sri	4
11	palang pintu	4
12	perlintasan margorejo	4
13	rel kereta	4
14	sri tanjung	4
15	api diskon	3
16	api lebaran	3
17	api mutiara	3
18	penonton membara	3
19	pt kai	3
20	royal plaza	3
21	tersambar kereta	3
…	…	…
313	wonokromo tewas	1

Lampiran 20 *Feature* pada Klaster 3

No	*Feature*	Frekuensi Berita
1	universitas airlangga	86
2	mahasiswa universitas	12
3	rektor universitas	11
4	dosen universitas	4
5	mahasiswi universitas	4
6	tunjungan plaza	4
7	airlangga meninggal	3
8	airlangga tenggelam	3
9	airlangga terima	3
10	meninggal dunia	3
11	wisuda universitas	3
12	aesthetic center	2
13	airlangga aesthetic	2
14	airlangga bunuh	2
15	airlangga ubah	2
16	airlangga universitas	2
17	alumnus universitas	2
18	audit bpk	2
19	bem universitas	2
20	bpk peneliti	2
21	bunuh tunjungan	2
…	…	…
480	wujud bakti	1

Lampiran 21 *Feature* pada Klaster 4

No	*Feature*	Frekuensi Berita
1	kepala polres	17
2	polres tulungagung	12
3	tol sumo	8
4	istri ajudan	5
5	ajudan kepala	4
6	tulungagung tewas	4
7	jenazah istri	3
8	renggut nyawa	3
9	dinas kepala	2
10	humas polda	2
11	kecelakaan maut	2
12	kecelakaan tol	2
13	mobil dinas	2
14	nyawa istri	2
15	polres perak	2
16	sumo dimakamkan	2
17	tewas kecelakaan	2
18	tewas tol	2
19	tulungagung kecelakaan	2
20	tulungagung tol	2
21	ajudan ganteng	1
…	…	…
91	wilayah polres	1

Lampiran 22 *Feature* pada Klaster 5

No	*Feature*	Frekuensi Berita
1	sabu sabu	27
2	divonis penjara	17
3	bandara juanda	3
4	kurir sabu	3
5	pengedar sabu	3
6	selundupkan sabu	3
7	beli sabu	2
8	dprd divonis	2
9	gagalkan penyelundupan	2
10	gerebek rumah	2
11	hukuman penjara	2
12	malaysia selundupkan	2
13	penjara terdakwa	2
14	penyelundupan sabu	2
15	sabu malaysia	2
16	sabu pemuda	2
17	wna malaysia	2
18	anak divonis	1
19	anak janda	1
20	anak umur	1
21	bandara batam	1
…	…	…
181	warung kopi	1

Lampiran 23 *Feature* pada Klaster 6

No	*Feature*	Frekuensi Berita
1	tunjungan plaza	40
2	mobil pejabat	33
3	bandara juanda	32
4	satpol pp	27
5	royal plaza	24
6	harian surya	22
7	taman bungkul	21
8	ditangkap polisi	20
9	asian games	19
10	pt kai	19
11	sma smk	19
12	apartemen educity	17
13	kebun binatang	17
14	kamar kos	16
15	kecanduan seks	16
16	anak muda	15
17	ketua dprd	15
18	mahasiswa its	15
19	uk petra	15
20	dinas pendidikan	14
21	kebakaran rumah	14
…	…	…
24.850	zoo target	1

Lampiran 24 *Feature* pada Klaster 7

No	*Feature*	Frekuensi Berita
1	dokter soetomo	48
2	rsud dokter	41
3	kembar siam	8
4	bayi kembar	7
5	dirawat rsud	7
6	miras oplosan	6
7	dokter spesialis	5
8	dokter rsud	4
9	diponegoro dokter	3
10	siam ternate	3
11	spesialis tht	3
12	tht rsud	3
13	anak rsud	2
14	cacar air	2
15	dokter rs	2
16	fakta dokter	2
17	libur lebaran	2
18	maut gresik	2
19	museum dokter	2
20	oplosan dirawat	2
21	oplosan maut	2
…	…	…
239	warga gresik	1

Lampiran 25 *Feature* pada Klaster 8

No	*Feature*	Frekuensi Berita
1	kepala polda	50
2	tabrak lari	15
3	pengemudi tabrak	8
4	polda irjen	6
5	polda pangdam	5
6	pangdam brawijaya	4
7	apel pilar	2
8	bayi instagram	2
9	diamuk warga	2
10	gubernur kepala	2
11	pilar kepala	2
12	polda apresiasi	2
13	polda jenguk	2
14	polda kunjungi	2
15	polda mahasiswa	2
16	polda perintahkan	2
17	polisi lamongan	2
18	aduan perumahan	1
19	ajang mendekatkan	1
20	ancam kepala	1
21	anggotanya sikat	1
…	…	…
272	zakat diimbau	1

Lampiran 26 *Feature* pada Klaster 9

No	*Feature*	Frekuensi Berita
1	tanjung perak	33
2	pelabuhan tanjung	15
3	polres pelabuhan	10
4	kejari tanjung	3
5	kepala polres	3
6	imigrasi tanjung	2
7	perak amankan	2
8	perak kejahatan	2
9	polres tanjung	2
10	akbp wakasat	1
11	amankan lembar	1
12	amankan wna	1
13	anggotanya terciduk	1
14	api polres	1
15	ayam semburat	1
16	balai karangtina	1
17	balai karantina	1
18	banjarmasin gerbang	1
…	…	…
172	zona integritas	1

Lampiran 27 *Feature* pada Klaster 10

No	*Feature*	Frekuensi Berita
1	anak anak	42
2	ujian nasional	42
3	nasional smp	7
4	kecurangan ujian	6
5	pelaksanaan ujian	6
6	dinas pendidikan	5
7	bocornya ujian	4
8	nasional smpn	4
9	anak panti	3
10	kepala smpn	3
11	panti asuhan	3
12	anak bom	2
13	anak remaja	2
14	belajar bahasa	2
15	bertemu anak	2
16	dipantau cctv	2
17	jelang pelaksanaan	2
18	jelang ujian	2
19	kasek smpn	2
20	kebocoran ujian	2
21	little pony	1
…	…	…
417	ypac beberkan	1

Lampiran 28 Lampiran Surat Pernyataan Data

SURAT PERNYATAAN

Saya yang bertanda tangan di bawah ini, mahasiswa Departemen Statistika FMKSD ITS,

Nama : Fonda Leviany
NRP : 062115 4000 0015

menyatakan bahwa data yang digunakan dalam Tugas Akhir ini merupakan data sekunder yang diambil dari ~~penelitian / buku / Tugas Akhir / Thesis / Publikasi /~~ lainnya yaitu :

Sumber : *Website* Tribunnews Surabaya
Keterangan: Data mengenai Berita Surabaya diambil dari *website* (http://surabaya.tribunnews.com) pada 6 Maret 2019.

Surat pernyataan ini dibuat dengan sebenarnya. Apabila terdapat pemalsuan data maka saya siap menerima sanksi sesuai aturan yang berlaku.

Surabaya, Juli 2019

Mengetahui,
Pembimbing Tugas Akhir

Dr. Kartika Fithriasari, M. Si.
NIP. 19691212 199303 2 002

Mahasiswa

Fonda Leviany
NRP. 062115 4000 0015

BIODATA PENULIS

Penulis yang bernama lengkap Fonda Leviany ini, akrab disapa Fonda oleh koleganya. Putri sulung dari pasangan Bapak Daniel dan Ibu Farida pada tanggal 28 September 1997 lalu dilahirkan di Kabupaten Klungkung, Bali. Pendidikan formal yang pernah ditempuh penulis adalah SD Muhammadiyah 1 Denpasar, SMP Albanna Denpasar, SMA Negeri 1 Gresik, dan melanjutkan pendidikan tingkat tingginya di Institut Teknologi Sepuluh Nopember (ITS) di Kota Surabaya mengambil Program Studi Sarjana di Departemen Statistika, Fakultas Matematika, Komputasi, dan Sains Data (FMKSD) melalui jalur SNMPTN. Selama menjalani 4 tahun masa perkuliahan, penulis juga aktif di organisasi UKM Koperasi Mahasiswa ITS sebagai Staff Bidang Bisnis (2016/2017), Asisten Bidang Bisnis (2017/2018), dan Ketua Bidang Keuangan (2018/2019). Selain itu, penulis juga aktif di organisasi Professional Statistics (PSt) HIMASTA-ITS sebagai Staf Divisi Public Relation (2016/2017) dan Kepala Divisi Public Relation (2017/2018). Di bidang akademik, penulis juga berkesempatan menjadi semifinalis dalam National Statistics Competition yang diselenggarakan di Universitas Brawijaya dan menjadi finalis di Indonesian Research Competition 3rd ISCO. Penulis pernah mendapatkan job survei dan analisis data serta mendapatkan pengalaman melakukan kerja praktik di PT. Pelindo III Cabang Tanjung Perak pada Divisi IT bagian manajemen data operasional. Bagi pembaca yang ingin menyampaikan kritik, saran, atau ingin berdiskusi dengan penulis dapat menghubungi melalui fondalevi@gmail.com.

(*Halaman ini sengaja dikosongkan*)