Pengelompokkan Komentar Pada Media Sosial Instagram Menggunakan Metode K-Means Clustering Untuk Identifikasi Awal Cyberbullying

Ahmad Muhariya
20917003

Tesis diajukan sebagai syarat untuk meraih gelar Magister Komputer
Konsentrasi Forensik Digital
Program Studi Informatika Program Magister
Fakultas Teknologi Industri
Universitas Islam Indonesia
2022

Lembar Pengesahan Pembimbing

Pengelompokkan Komentar Pada Media Sosial Instagram Menggunakan Metode K-Means Clustering Untuk Identifikasi Awal Cyberbullying

Ahmad Muhariya
20917003

Yogyakarta, November 2022

Pembimbing 1

Dr. Imam Riadi, M.Kom

Pembimbing 2

Dr. Yudi Prayudi, S.Si., M.Kom

Lembar Pengesahan Penguji

Pengelompokkan Komentar Pada Media Sosial Instagram Menggunakan Metode K-Means Clustering Untuk Identifikasi Awal Cyberbullying

Ahmad Muhariya
20917003

Yogyakarta, November 2022

Tim Penguji,

Dr. Imam Riadi, M.Kom
Ketua

Dr. Yudi Prayudi, S.Si., M.Kom
Anggota I

Dr. Ir. Bambang Sugiantoro, S.Si., MT.
Anggota II

Mengetahui,
Ketua Program Studi Informatika Program Magister
Universitas Islam Indonesia

Irving Vitra Paputungan, S.T., M.Sc., Ph.D.

Abstrak

Pengelompokkan Komentar Pada Media Sosial Instagram Menggunakan Metode K-Means Clustering Untuk Identifikasi Awal Cyberbullying

Media sosial selain memberikan dampak positif pada masyarakat juga memiliki dampak negatif. Berdasarkan statistic, 95 persen dari pengguna internet di Indonesia menggunakan internet untuk mengakses jejaring sosial. Secara khusus untuk kalangan muda, Instgram lebih banyak digunakan dibandingkan dengan media sosialnya lainnya seperti Twitter dan Facebook. Dalam hal kasus cyberbullying, kasus yang sering terjadi adalah melalui media sosial twitter dan Instagram. Terdapat sejumlah metode yang umumnya digunakan untuk analisis terhadap kasus cyberbullying seperti SVM (Support Vector Machine), NBC (Naïve Bayes Classifier), C45, K-Nearest Neighbours. Penerapan terhadap sejumlah metode tersebut umumnya di implementasikan pada media social twitter. Sementara pengguna usia muda saat ini lebih banyak menggunakan media social Instagram dibandingkan dengan twitter. Untuk itu penelitian memberikan focus pada analisis cyberbullying pada Instagram melalui penerapamn algoritma K-Mean Clustering. Algoritma ini digunakan untuk mengelompokkan tindakan cyberbullying yang terdapat pada komentar. Dataset yang digunakan pada penelitian ini diambil dari tahun 2019 sampai 2021 dengan jumlah 650 record, terdapat 1827 kata dan sudah memiliki label. Penelitian ini telah berhasil mengelompokkan data yang diuji, dengan nilai threshold 0,5. Hasil penelitian yang didapat untuk pengelompokkan kata yang mengandung bullying pada Instagram menghasikan tingkat accuracy paling tinggi yaitu sebesar 67,38%, nilai precision 76,70%, dan nilai recall 67,48%. Data ini menunjukkan bahwa algoritma k- means dapat mengelompokkan komentar menjadi dua cluster, yaitu cluster bullying dan non-bullying.

Kata kunci

cyberbullying; media sosial; instagram; kmeans clustering;

Abstract

Grouping Comments on Instagram Social Media Using the K-Means Clustering Method for Early Identification of Cyberbullying

Social Media, in addition to having a positive impact on society, also has a negative effect. Based on statistics, 95 percent of internet users in Indonesia use the internet to access social networks. Especially for young people, Instagram is more widely used than other social media such as Twitter and Facebook. In terms of cyberbullying cases, cases often occur through social media, Twitter, and Instagram. Several methods are commonly used to analyze cyberbullying cases, such as SVM (Support Vector Machine), NBC (Naïve Bayes Classifier), C45, and K-Nearest Neighbors. Application of a number of these methods is generally implemented on Twitter social media. Meanwhile, young users currently use Instagram more social media than Twitter. For this reason, the research focuses on analyzing cyberbullying on Instagram by applying the K-Mean Clustering algorithm. This algorithm is used to classify cyberbullying actions contained in comments. The dataset used in this study was taken from 2019 to 2021 with 650 records; there were 1827 words and already had labels. This study has successfully classified the tested data with a threshold value of 0.5. The results for grouping words containing bullying on Instagram resulted in the highest accuracy, which is 67.38%, a precision value of 76.70%, and a recall value of 67.48%. These results indicate that the k-means algorithm can make a grouping of comments into two clusters: bullying and non-bullying

Keywords

cyberbullying; social media; instagram; kmeans; clustering;

Pernyataan Keaslian Tulisan

Dengan ini saya menyatakan bahwa tesis ini merupakan tulisan asli dari penulis, dan tidak berisi material yang telah diterbitkan sebelumnya atau tulisan dari penulis lain terkecuali referensi atas material tersebut telah disebutkan dalam tesis. Apabila ada kontribusi dari penulis lain dalam tesis ini, maka penulis lain tersebut secara eksplisit telah disebutkan dalam tesis ini.

Dengan ini saya juga menyatakan bahwa segala kontribusi dari pihak lain terhadap tesis ini, termasuk bantuan analisis statistik, desain survei, analisis data, prosedur teknis yang bersifat signifikan, dan segala bentuk aktivitas penelitian yang dipergunakan atau dilaporkan dalam tesis ini telah secara eksplisit disebutkan dalam tesis ini.

Segala bentuk hak cipta yang terdapat dalam material dokumen tesis ini berada dalam kepemilikan pemilik hak cipta masing-masing. Apabila dibutuhkan, penulis juga telah mendapatkan izin dari pemilik hak cipta untuk menggunakan ulang materialnya dalam tesis ini.

ober 2022

Ahmad Muhariya, S.Kom

Daftar Publikasi

Publikasi yang menjadi bagian dari tesis

Ahmad Muhariya, Imam Riadi, Yudi Prayudi (2022). Cyberbullying Analysis on Instagram Using K-Means Clustering. JUITA (Jurnal Informatika)

Sitasi publikasi 1

Kontributor	Jenis Kontribusi
Ahmad Muhariya, 2022	Mendesain eksperimen (60%) Menulis *paper* (70%)
Iman Riadi, Yudi Prayudi, 2022	Mendesain eksperimen (40%) Menulis dan mengedit *paper* (30%)

Halaman Kontribusi

“**Tidak ada kontribusi dari pihak lain**”

Halaman Persembahan

Ku persembahkan tesisku ini untuk:

"Kedua orang tua ku"

Bapak Tukimin dan ibu Siti Semiyati

"Teman-temanku"

Dan

"Pacarku atau calon pendamping hidupku"

Oktavia Purwati Ningsih

yang selalu bertanya:

"kapan kelar kuliah Ong?"

Kata Pengantar

Assalamu'alaikum warahmatullahi wabarakatuh

Alhamdulillah, puji dan syukur kehadirat Allah SWT yang telah melimpahkan rahmat, hidayah, serta kasih sayang-Nya. Sholawat dan salam senantiasa tercurah kepada junjungan dan uswah hasanah kita, Nabi Muhammad SAW, yang telah membimbing kita dari zaman jahiliyah gelap gulita menuju zaman islam yang terang benderang seperti sekarang ini. Atas karunia Allah SWT yang Maha Pemurah, penulis dapat menyelesaikan thesis dengan judul "Analisis Cyberbullying Pada Layanan Sosial Media Instagram Menggunakan K-Means Clustering". Thesis ini disusun untuk memenuhi persyaratan guna mencapai derajat Magister (Strata II/ S2) Program Studi Magister Informatika di Universitas Islam Indonesia. Penulis menyadari sepenuhnya jika tanpa adanya bantuan dari berbagai pihak, maka penulis tidak akan menyelesaikan thesis ini dengan baik. Oleh karena itu, selain rasa syukur yang tiada henti tercurah, izinkan penulis dengan tulisan ini untuk menyampaikan segenap terimakasih kepada :

1. Allah SWT, atas segala kesempatan, kekuatan dan seluruh hal yang diberikan kepada penulis hingga detik ini.
2. Kedua orangtua tercinta serta seluruh keluarga kami. Terimakasih atas do'a, kasih sayang dan support yang tiada pernah ada habisnya.
3. Bapak Dr. Imam Riadi, M.Kom dan Dr. Yudi Prayudi, S.Si., M.Kom selaku dosen pembimbing sekaligus Dewan Penguji yang selalu memberikan arahan, motivasi semangat dan do'a selama kegiatan penelitian dan penyusunan thesis.
4. Bapak Dr. Ahmad Luthfi, M.Kom selaku Anggota Dewan Penguji yang telah memberikan kritik dan saran yang membangun, dukungan, serta nasihat yang amat berarti. Terimakasih atas ilmu dan pengetahuan yang telah diberikan.
5. Penulis mengucapkan terima kasih kepada Direktorat Penelitian dan Pengabdian kepada Masyarakat (DPPM) dan Program Magister Informatika Universitas Islam Indonesia atas dukungan administrasi dan fasilitas penelitian yang telah diberikan.
6. Bapak dan Ibu Dosen Program Studi Magister Informatika Universitas Islam Indonesia yang telah mendidik dan memberikan bekal pengetahuan selama x kegiatan perkuliahan atau praktikum. Semoga menjadi amal jariyah di akhirat kelak.
7. Seluruh keluarga besar haji pokol dan mbah basuri atas semua dukungan dan doa.

8. Seluruh teman teman seperjuangan Keluarga Besar Mahasiswa Program Studi Magister Informatika Universitas Islam Indonesia atas semua dukungan dan doa.
9. Last, kepada seluruh sahabat, teman dan orang-orang yang mengenal penulis - baik hanya singgah maupun bertahan hingga sekarang yang tidak bisa kami sebutkan satu persatu. Terimakasih sudah hadir dan memberikan banyak pelajaran dalam kehidupan.

Penulis ucapkan terima kasih. Semoga Allah SWT membalas kebaikan dari berbagai pihak yang kami sebutkan dengan berkah-Nya. Penulis berharap skripsi ini dapat bermanfaat untuk para pembaca sekalian. Apabila adakekurangan dalam penulisan, penulis menyampaikan permohonan maaf yang sebesar besarnya. Oleh karena tidak sempurnanya tesis ini, penulis berharap kritik dan saran yang membangun untuk perbaikan penulis selanjutnya.

Wassalamu’alaikum warahmatullahi wabarakatuh.
Yogyakarta, Oktober 2022

Ahmad Muhariya, S.Kom

Daftar Isi

Daftar Tabel

Daftar Gambar

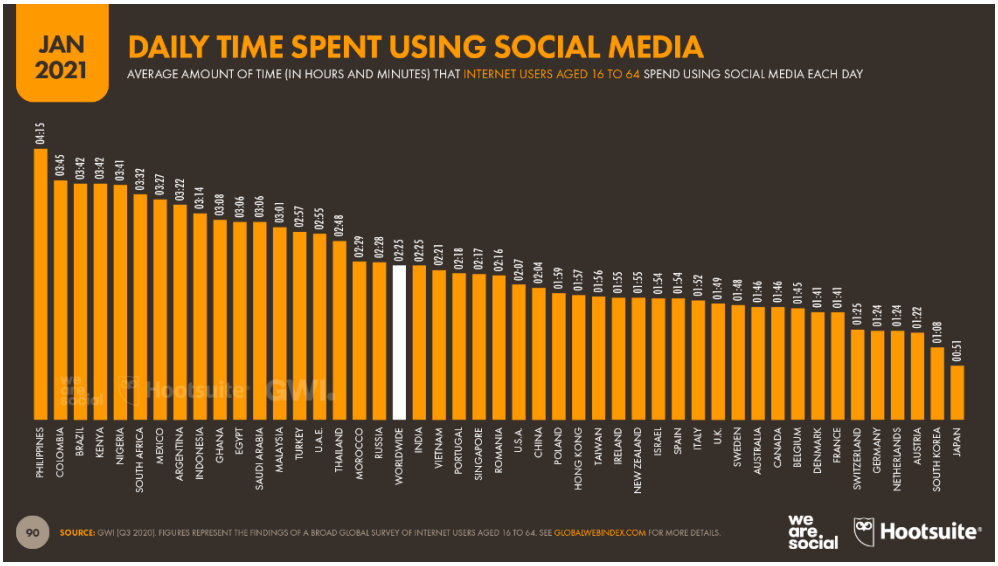
BAB 1

Pendahuluan

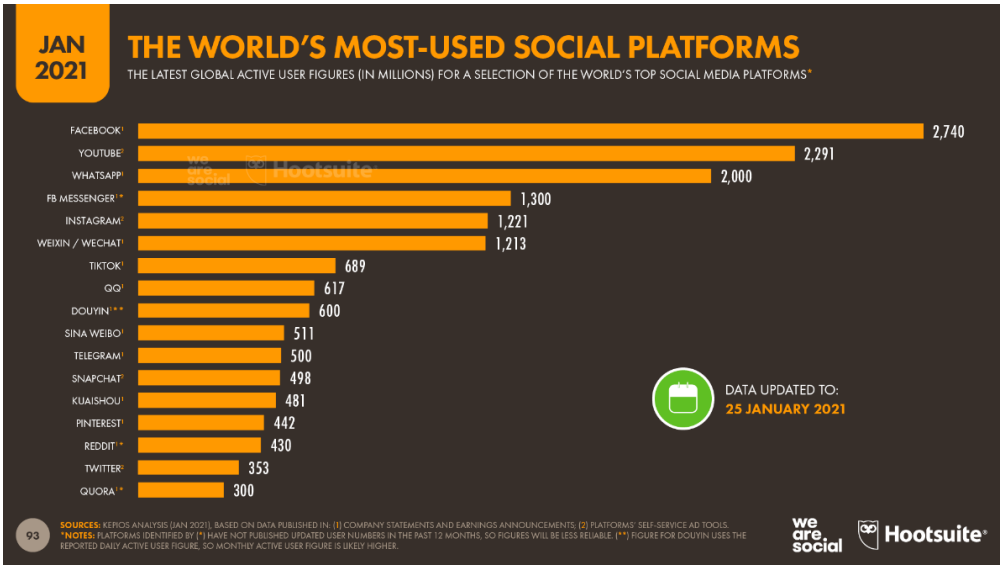
1.1 Latar Belakang

Perkembangan teknologi yang semakin pesat memberikan warna yang positif bagi masyarakat. Adanya kemudahan teknologi ini tentunya memudahkan masyarakat dalam menjalani segala aktivitas. Selain memberikan dampak positif, teknologi juga memberikan warna yang negatif. Hal ini dapat dilihat dari banyaknya kejahatan yang dilakukan dengan memanfaatkan teknologi internet. Sejalan dengan hal tersebut, dapat dipastikan bahwa semakin berkembangnya teknologi maka pola kejahatan juga semakin beragam. Beragamnya masyarakat dalam menghadapi kemajuan teknologi dan komunikasi bisa dilihat dari informasi yang disebarkan secara cepat melalui internet terutama melalui Media Sosial.

Media sosial bagian bentuk dari kemajuan teknologi yang mudah mempengaruhi masyarakat dalam cara pandang, gaya hidup dan budaya. Selain itu media sosial bisa bercakap, berinteraksi, mengasah ketajaman nalar, hal ini bisa membuat masyarakat berdampak negatif maupun positif (Fazry and Cipta Apsari 2021). Kementrian Komunikasi Dan Informatika menyatakan bahwa penggunaan internet di indonesia mencapai 175,5 juta orang, 95 persen menggunakan internet untuk mengakses jejaring sosial, situs yang sering diakses adalah *Instagram, Facebook* dan *Twiter*. Sebanyak 120 juta orang di indonesia menggunakan perangkat *mobile*, seperti *smartphone* atau tablet untuk mengakses media sosial, dengan penetrasi 45 persen. Dalam sepekan, aktivitas online dimedia sosial melalui *smartphone* mencapat 37 persen menunjukkan statistik penggunaan media sosial dari berbagai negara indonesia menepati peringkat ketiga (Kementerian Komunikasi dan Informatika 2021). Dalam hal ini statistic juga menyebutkan bahwa dikalangan anak muda, media social Instagram lebih dominan jika dibandingkan dengan Twitter. Karena itu, Instagram menjadi media sosial yang paling sering digunakan sebagai media social untuk postingan yang mengarah pada kasus cyberbullying. Presentase statistik dari negara ditunjukkan pada gambar 1.1 statistik penggunaan media sosial berbagai negara.



Gambar 1. 1 Statistik Media Sosial Berbagai Negara



Gambar 1. 2 Penggunaan media sosial yang paling popular

Instagram menjadi media sosial yang paling umum *cyberbullying* di internet. Menurut hasil survei dari lembaga antibullying, Ditch The Label. *Cyberbullying* yang di maksud di sini adalah komentar negatif pada postingan, pesan personal yang tak baik, serta mengolok-olok. Sepuluh ribu Remaja yang berusia 12 tahun sampai 20 tahun yang berdomisili di Inggris di jadikan sumber survei. Hasil yang didapat lebih dari 42 persen korban mengaku mendapatkannya di instagram, sehingga dibutuhkannya keahlian di bidang forensika digital. Tujuannya untuk mendukung langkah investigasi dan pencarian barang bukti pada kejahatan *cyberbullying*. *Cyberbullying* ternyata lebih menyakitkan dibandingkan dengan kekerasan fisik (Tapia, Aguinaga, and Luje 2019), dampak *cyberbullying* secara tidak langsung berakibat tindakan-tindakan kriminal seperti pencemaran nama baik,

peneroran dll (Chukwuere, Chukwuere, and Adom 2021). Sedangkan definisi dari cyberbullying dapat disimpulkan sebagai perilaku yang melecehkan, menghina, mengancam, merendahkan, atau membahayakan seseorang terus menerus dengan memanfaatkan teknologi dan internet (Suryanto 2017).

Dari banyaknya literatur yang didapat, ada banyak metode dan algoritma yang digunakan untuk analisis cyberbullying, diantaranya adalah: SVM (Support Vector Machine), NBC (Naïve Bayes Classifier), C45, K-Nearest Neighbours. Akan tetapi penerapan algoritma tersebut kebanyakan diimplementasikan pada Twitter, bukan Instagram. Penelitian sejenis dengan objek Instargram telah berhasil mengelompokkan pengguna Instagram berdasarkan kesesuaian hashtag tertentu pada teks Instagram. Dalam hal ini algoritma k-means dan TF-IDF digunakan sebagai fitur utama untuk pengelompokkan tersebut (Habibi and Cahyo 2019). Namun penerapan algoritma k-means clustering untuk analisis cyberbullying di Instagram belum pernah dilakukan (Naf'an et al. 2019). Metode k-means digunakan untuk mengelompokkan tindakan cyberbullying yang ada pada komentar. K-means merupakan metode yang bersifat partitional data. K-means bekerja dengan cara membagi himpunan data ke dalam cluster yang tidak overlap sehingga setiap data berada tepat berada dalam satu cluster. Penentuan hasil nilai cluster dilihat dari jarak terdekat antara data dengan centroid (Suryanto 2017). K-Means clustering bertujuan untuk mengoptimalkan suatu fungsi untuk menghitung jarak ruang antara objek dengan centroid (titik tengah) cluster (Adiya and Desnelita 2019). Dengan demikian, diharapkan penggunaan K-means clustering juga dapat dilakukan untuk mengelompokkan komentar pada Instagram yang memuat unsur bullying dan non-bullying. Kontribusi dari penelitian ini adalah dalam hal penerapen algoritma kmeans clustering yang dipadukan dengan pembobotan tf-idf untuk identifikasi awal cyberbullying sebagai bukti digital untuk keperluan persidangan dan implementasi algoritma tersebut untuk analisis cyberbullying di Instagram belum pernah dilakukan.

K-Means merupakan salah satu metode clustering data yang dapat digunakan untuk membagi data menjadi dua cluster atau lebih. Dengan menggunakan algoritma K-Means akan lebih memudahkan clustering karena kesederhanaan dan efisiensi sehingga komentar di Instagram dapat dikelompokkan menjadi beberapa cluster dalam kasus Cyberbullying (Adiya and Desnelita 2019).

1.2 Rumusan Masalah

Berdasarkan uraian pada latar belakang maka dapat dirumuskan suatu permasalahan yaitu :

1. Mengimplementasikan k-means clustering yang dipadukan dengan pembobotan TF-IDF untuk pengelompokkan komentar bullying dan non-bullying.
2. Mengidentifikasi tindakan awal *cyberbullying* sebagai bukti digital untuk keperluan persidangan

1.3 Batasan Penelitian

Untuk menjaga focus dalam penelitian maka ada beberapa batasan yang diberikan dalam penelitian adalah :

1. Penelitian ini hanya berfokus pada komentar yang berbahasa indonesia KBBI ataupun KBBA.
2. Penelitian ini hanya berfokus pada komentar atau text yang mengandung *bullying* bukan pada suara, gambar dan video yang mengandung *bullying.*

1.4 Tujuan Penelitian

Tujuan dalam penelitian ini adalah :

1. Menganalisis komentar instagram dengan teknik pengelompokan menggunakan k-means clustering yang dipadukan dengan pembobotan TF-IDF.
2. Mengembangkan teknik dalam mengidentifikasi tindakan awal *cyberbullying* dan *Non-cyberbullying* pada komentar Instagram.

1.5 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah :

1. Menambah pengetahuan dalam menganalisis bukti digital *Cyberbullying* dengan teknik pengelompokan menggunakan *K-Means Clustering.*
2. Memberikan informasi *cyberbullying* di Indonesia kepada pihak-pihak yang membutuhkan untuk dijadikan referensi khususnya orang tua ataupun penyidik. Informasi tersebut dapat memberikan edukasi yang benar bagi pelaku maupun korban.

1.6 Luaran Penelitian

Luaran penelitian ini adalah : Luaran penelitian ini adalah publikasi di jurnal lokal yang terakreditasi Sinta.

BAB 2
Tinjauan Pustaka

2.1 Penelitian Terdahulu

Penelitian ini difokuskan pada analisis cyberbullying di media sosial Instagram menggunakan metode k-means clustering. Deteksi cyberbullying berguna untuk mengembangkan teknologi untuk melindungi seseorang di situs jejaring sosial. Beberapa teknik telah digunakan dalam beberapa tahun terakhir, untuk mengidentifikasi konten yang melanggar di platform media sosial seperti Instagram, Facebook dan Twitter. Banyak peneliti telah mengusulkan deteksi kata-kata pengganggu dengan menggunakan teknik dan metode yang berbeda pada bahasa yang berbeda.

Cyberbullying terdiri dari dua individu yang terlibat, yaitu pelaku bullying dan korban bullying sebagai target bullying (Al-rahmi et al. 2018) menyelidiki faktor-faktor yang mempengaruhi cyber bullying pada pelecehan cyber, cyberstalking di kalangan mahasiswa. Saat ini kasus cyberbullying sudah banyak terjadi, target bullying yang paling sering mendapatkan komentar berisi cyberbullying biasanya adalah seseorang *public figure* dengan *followers* lebih dari 1 juta di akun sosial medianya (Yoannes Romando, Sulistyowati, and Wibisono 2019). Tidak berbeda jauh dengan *public figure,* seseorang yang populer di media sosial seperti selebgram juga kerap mendapatkan komentar buruk yang mengandung cyberbullying di akun Instagramnya (Andriansyah et al. 2018). Para *public figure* kerap kali memposting sesuatu yang kontroversial sehingga menyebabkan pro dan kontra di masyarakat (Nurrahmi 2018). Representasi pesan yang kuat dan diskriminatif sangat penting untuk sistem deteksi yang efektif. Selain itu, penyisipan kata telah digunakan secara otomatis memperluas dan menyempurnakan daftar kata intimidasi yang diinisialisasi (Zhao and Mao 2017). Tindakan yang dilakukan seseorang yang akan melakukan cyberbullying dengan tingkat yang berbeda-beda (Imam Riadi, Sunardi 2021) dan kemiripan kata atau text yang ada pada bukti yang ada pada (Widiandana and Riadi 2019) sentiment memiliki bobot yang berbeda ada sentiment positif dan negative sehingga hasilnya dapat diketahui polaritas dari setiap sentimentnya (Luqyana, Cholissodin, and Perdana 2018).

Cyberbullying dikategorikan kedalam beberapa bentuk seperti flaming, pelecehan, cyber stalking, fitnah, pengecualian, trolling, peniruan identitas, dan tipu daya (Hang and

Dahlan 2019). Sedangkan pada Konferensi Internasional ke-3 2018 (Gorro et al. 2018) Pengelompokan Cyberbullying mengkategorikan menjadi 3 :

1. Ancaman : Ungkapan yang mengandung ancaman fisik atau psikologis, atau indikasi pemerasan
2. Kutukan : Ungkapan keinginan bahwa suatu bentuk kesengsaraan atau kemalangan akan menimpa korbannya
3. Seksual : Ekspresi dengan makna seksual yang menyebutkan bagian tubuh pribadi

Dari 3 kelompok di atas, Menurut (Gorro et al. 2018) Tahapan-tahapan yang perlu dilakukan dalam investigasi bukti digital cyberbullying :

1. Membangun corpus dalam hal ini Proses Scraping menggunakan Selenium.
2. Pra pemrosesan data data yang tidak relevan seperti symbol, karakter, gambar maka dilakukannya proses penghapusan karakter, URL, gambar, symbol.
3. Identifikasi Fitur setiap postingan dipriksa dengan TF-IDF dan countvectorizer untuk mengidentifikasi kata yang sering digunakan dalam postingan yang mengandung unsur cyberbullying seperti Bodoh, tolol, bobo, gago, gaga, tanga, bobo, Gunggong, bajingan, bercinta, kantot, magkantot, iyot, brengsek, seks, pecundang, kalog, bigaun, pelacur, kontol, pantat, bajingan, vagina, kokang, mati, kontol, membunuh, mati, kutu buku, aneh, jelek, gendut, sial, sialan, jalang, sial, kencing, bajingan.
4. Pelatihan model dalam ini setiap postingan diberi label secara manual sesuai dengan jenis cyberbullying yang sesuai, postingan yang tidak mengandung unsur dilabeli dengan “lainnya”
5. Klasifikasi Cyberbullying, setiap kata yang diidentifikasi, postingan dikumpulkan dan diklasifikasikan sebagai postingan cyberbullying
6. Corpus Cyberbullying, kumpulan postingan disimpan kedalam database lalu dilakukan analisis lebih lanjut.

Keenam tahapan yang dikemukakan tersebut, ternyata belum menjadi solusi terhadap kesenjangan dalam hal mekanisme penanganan bukti digital, perkataan yang mengandung kebencian atau penindasan sulit untuk dideteksi oleh manusia dikarenakan banyaknya bahasa yang dipakai (Ishara Amali and Jayalal 2020). Kebanyakan yang dilakukan peneliti sebelumnya hanya melakukan pendekatan satu Bahasa dan pengumpulan data yang dilakukan secara manual. Solusi di berikan dengan menggunakan dua Bahasa dan penggunaan API dalam pengumpulan data (Pawar and Raje 2019). Lalu tahapan-tahapannya yang pertama adalah mengumpulkan data, kemudian dilakukan pra-pemrosesan

pembersihan data, selanjutnya mengklasifikasikan data berguna membandingan data sentiment (Tapia and Aguinaga 2018). Selain itu, permasalahan yang sering ditemukan dalam cyberbullying yaitu sulitnya mengidentifikasi korban dan pelaku yang melakukan cyberbullying karena pemeriksaan dilakukan secara kasat mata dan tidak adanya referensi yang kuat sebagai bukti kasus cyberbullying yang dilakukan oleh pelaku cyberbullying (Riadi, Sunardi, and Widiandana 2020). Penelitian ini diharapkan dapat menambah referensi penyidik untuk mendapatkan atau mengidentifikasi tindakan cyberbullying yang selama ini semakin tinggi. Metode yang dipakai untuk mengidentifikasi tindakan cyberbullying mengunakan metode K-Means clustering dan metode TF-IDF untuk mencari kata dalam komentar dengan mempersiapkannya sebelum mencari kata yang ada di dalam komentar. Biasa dilihat tabel 2.1 Ulasan Kritis Tema.

Tabel 2. 1 Ulasan Kritis Tema

No	Peneliti	Tujuan	Pendekatan/Metode Penelitian	Domain Penelitian	Hasil Penelitian
1.	(Al-rahmi et al. 2018)	Pengukur masalah etika terkait penggunaan media sosial pada kalangan mahasiswa	Menggunakan kuesioner sebagai metode pengumpulan data dan metode penelitian kuantitatif	Semua situs sosial media	Penggunaan media sosial dan keterlibatan dunia maya memperkuat intimidasi dunia maya, pelecehan dunia maya, dan cyberstalking
2.	(Yoannes Romando, Sulistyowati, and Wibisono 2019)	Menerapkan model investigasi digital forensic kusus untuk identifikasi komentar negative pada social media instagram dengan pengelompokan yang memiliki bobot yang berbeda	Metode yang dipakai meliputi : - Case folding - Tokenizing - Filtering - Ekstraksi - Klasifikasi	Media sosial Instagram	Metode yang digunakan pada penelitian ini mampu mendeteksi komentar negatif dari instagram berbahasa Indonesia
3.	(Andriansyah et al. 2018)	Mengklasifikasikan dan mengkategorikan komentar	Metode yang di pakai dengan menggunakan data training dan data set	Media sosial Instagram	Model ini dapat diterapkan pada Instagram, misalnya jika

No	Peneliti	Tujuan	Pendekatan/Metode Penelitian	Domain Penelitian	Hasil Penelitian
		instagram yang mengandung cyberbullying.	yang diambil dari komentar lalu dilakukan pra-prosessing dan diberi label lalu dimasukkan kedalam model SVM		komentar pada postingan Instagram terindikasi sebagai komentar bullying, maka dapat dihilangkan dari postingan tersebut.
4.	(Nurrahmi 2018)	Deteksi teks cyberbullying menggunakan teknik text mining, deteksi aktor cyberbullying menggunakan metode komputasi untuk pengukuran kredibilitas pengguna.	Metode yang di pakai dengan menggunakan data training dan data set yang diambil dari komentar lalu dilakukan pra-prosessing dan diberi label lalu dimasukkan kedalam model SVM	Media sosial Twitter	Alat pelabelan berbasis web untuk mengklasifikasikan tweet menjadi tweet cyberbullying dan non-cyberbullying
5.	(Pawar and Raje 2019).	Mendeteksi cyberbullying dalam bahasa Inggris, Hindi dan Marathi	Metode yang dipakai meliputi : Pengumpulan data menggunakan API, Pra-pemrosesan Data, Pembuatan Data Sintetis	Tweet, ulasan surat kabar, dan ulasan wisatawan.	Mengembangkan prototipe yang beroperasi di seluruh kumpulan data yang dibuat untuk dua bahasa
6.	(Hang and Dahlan 2019)	Mengusulkan suatu leksikon cyberbullying yang dapat digunakan untuk membantu deteksi cyberbullying di media sosial.	Metode yang dipakai meliputi : Memahami konsep eksklusi cyberbullying, pemilihan daftar kata, identifikasi kata kunci, identifikasi kelas dan subclass, dan pengembangan ontologi dan	Semua situs sosial media	Leksikon dapat digunakan sebagai kamus bagi pengguna di media sosial, karena sangat berguna bagi pengguna untuk mengetahui kata-kata mana yang dianggap sebagai kata-kata cyberbullying.

No	Peneliti	Tujuan	Pendekatan/Metode Penelitian	Domain Penelitian	Hasil Penelitian
			leksikon cyberbullying		
7.	(Ishara Amali and Jayalal 2020)	Pendekatan otomatis untuk mendeteksi komentar media sosial	Model yang dilakukan dengan menerapkan model aturan yang dibuat Komentar yang mengandung kata yang buruk, kombinasi kata ganti pertama/ kedua/ kata buruk, kombinasi kata ganti kedua dengan kata yang buruk, kombinasi kata ganti ketiga dengan kata yang buruk, kombinasi kata ganti orang pertama/kata buruk/ kata ganti orang ketiga.	Twiter	Membuat analisis teks menggunakan aturan dan algoritma pembelajaran mesin. Untuk mengidentifikasi pelecehan dalam teks media sosial, kami menggunakan lima aturan
8.	(Tapia and Aguinaga 2018)	Menemukan konten yang berbau cyberbullying, dengan analisis sentimen	Pendekatan yang dilakukan dengan model : pengumpulan data, Tuan Tuit/Api yang berada di sosmed, Pra pemrosesan, Klasifikasi manual, Klasifikasi perbandingan otomatis, dan Analisa pola perilaku	Semua situs sosial media	Tuan Tuit memberikan hasil yang lebih baik dalam klasifikasi perasaan negatif sehubungan dengan Sentimen.

No	Peneliti	Tujuan	Pendekatan/Metode Penelitian	Domain Penelitian	Hasil Penelitian
9.	(Imam Riadi, Sunardi 2021)	Menerapkan model investigasi forensic kusus untuk social media kedalam sebuah aplikasi yang akan digunakan untuk penyelidikan, dokumentasi, dan laporan	Metode investigasi digital forensic pada sosail media, yakni dengan DFRWS dengan tahapan Identifikasi, preservation, collection, examination, analisis, dan presentasi	Whatsapp Grub	Metode investigasi digital forensic dengan DFRWS membantu proses akuisisi sehingga terjaga kelestariannya
10.	(Widiandana and Riadi 2019)	Mengungkap bukti digital cyberbullying berbentuk teks.	Metode investigasi digital forensic pada sosail media, dengan tahapan, yakni Perancangan, Simulasi cyberbullying, akusisi dengan metode NIST, analisis cyberbullying, hasil identifikasi	Whatsapp	Dengan menggunakan metode investigasi NIST membantu pengangkatan barang bukti digital, dengan memanfaatkan metode cosine similarity dengan membandingkan antara kata yang ada dalam barang bukti dengan keyword yang mengarah pada tindakan cyberbullying sehingga dapat mengetahui pelaku telah melakukan cyberbullying

2.2 Konsep Pengetahuan

2.2.1 Digital Forensik

Digital forensic merupakan sebuah keahlian, seni dan keterampilan dalam menganalisa dan memulihkan data dari perangkat digital seperti computer, smartphone, leptop, dan lainnya (Prasad & Pandey, 2016). Tujuannya untuk mendapatkan temuan-temuan dari penyelidikan yang nantinya berguna untuk menjawab dipersidangan (Quick & Choo, 2016). Pada umumnya, proses digital forensic dibagi menjadi 4 (empat) tahap (Mckemmish, 1999) bisa dilihat pada gambar 2.1, antara lain:

Gambar 2. 1 Proses Digital Forensik

1. Proses ini adalah tahapan awal (Identifikasi bukti digital). Perlu diketahui apa saja jenis bukti digitalnya, dimana disimpan, dan bagaimana penyimpanannya. Proses ini sangatlah penting dimana bukti digital yang ditemukan akan mendukung proses penyelidikan selanjutnya.
2. Proses ini adalah tahapan kedua (Preservasi bukti digital). Proses ini meliputi penyimpanan dan penyiapan bukti digital. Bukti digital memiliki sifat mudah rusak, terjadi perubahan dan bisa saja terhapus (volatile), sedikit saja terjadinya perubahan pada bukti digital maka tidak dapat di ajukan ke pengadilan.
3. Proses ini adalah tahapan ketiga (Analysis bukti digital). Setelah proses kedua ditemukan maka perlu dilakukan proses extraksi dan dilakukan proses selanjutnya yaitu analisis bukti digital. Pemeriksaan ini untuk mendapatkan suatu informasi yang digunakan untuk menjawab yang berhubungan dengan investigasi seperti siapa pelaku, apa yang terjadi, apa saja aplikasi yang digunakan, siapa pelakunya dan kapan waktunya.
4. Proses ini adalah proses terakhir (Presentation). Temuan yang didapatkan dalam proses pemeriksaan sampai analisis perlu dipresentasikan ke pihak terkait seperti penyidik ataupun pengadilan.

2.2.2 Forensik Media Sosial

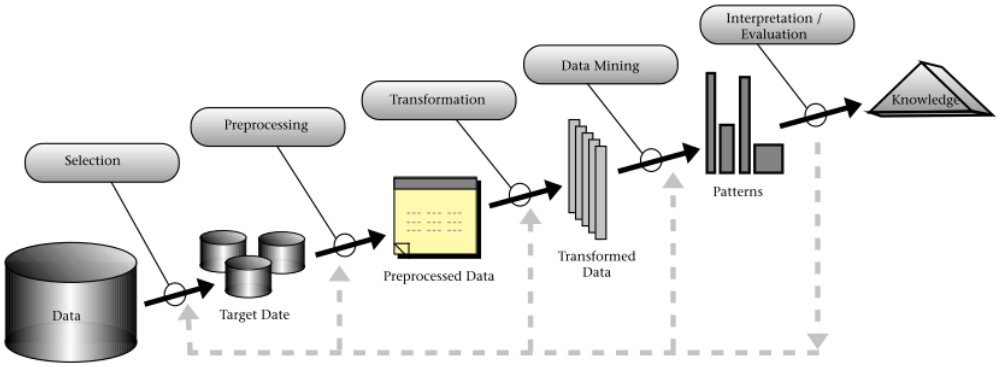
Investigasi forensic media social biasanya dilakukan dengan melihat bukti awal yang ditemukan, lalu melihat dimana buti tersebut terjadi. Forensik media social merupakan bagian dari forensic digital yang mana penerapan teknik investigasi dan analisis dilakukan untuk mengumpulkan bukti digital yang nantinya digunakan dipersidangan (Mohan and Venkataraman 2017), Untuk mendapatkan data dan informasi dari media social, dapat

dilakukan dengan beberapa cara seperti mengekstraksi data dari perangkat digital, mengekstraksi data menggunakan web crawler, dan mengekstraksi data menggunakan API yang disediakan oleh 23 media social . Situs media social yang popular biasanya menyediakan API untuk dimanfaatkan oleh para software developer, penyelidikan oleh apparat penegak hukum, dan lainnya. Dalam mengumpulkan bukti dari media social, harus menggunakan prinsip-prinsip dalam digital forensic, dikarenakan bukti digital yang memiliki sifat yang rapuh atau gampang berubah atau terhapus (Casey, 2011).

2.2.3 Data Mining

Data mining adalah suatu algoritma yang digunakan untuk mengekstrak pola data yang berguna digunakan dalam pengambilan keputusan. Data mining memiliki beberapa proses diantaranya forensic data *(association)*, mengidentifikasi kelompok tertentu *(clustering)*, menemukan data untuk memprediksinya *(forecasting)*, menyetor dan menemukan data kedalam kelompok tertentu *(classification)*. Data mining tidak terbatas pada digital forensic, model data mining yang dikembangkan bisa berguna dan membantu investigator untuk menemukan bukti digital lebih cepat dan efesien (Lei Xun et al. 2014).

Istilah data mining dan knowledge discovery in database (KDD) sering digunakan secara bersamaan dan bergantian untuk menjelaskan proses pencarian informasi tersembunyi dalam suatu big data. Kedua konsep tersebut memiliki konsep yang berbeda akan tetapi saling berkaitan satu sama lain. Salah satu tahapan dalam keseluruhan proses KDD adalah data mining (Usama Fayyad, Gregory Piatetsky-Shapiro 1996). Proses KDD memiliki beberapa tahapan dapat dijelaskan pada gambar 2.2, antara lain :



Gambar 2. 2 Fase-fase Dalam Data Mining

1. Data selection

Pemilihan data (seleksi) dari sekumpulan data perlu dilakukan pemilihan sebelum tahapan pencarian dan penggalian informasi dalam KDD dimulai. Data hasil pemilihan (seleksi) yang nantinya akan digunakan untuk proses data mining disimpan disuatu berkas dan terpisah.

2. Pre-prosessing/ Cleaning

Sebelum proses data mining dilakukan proses Cleaning data yang nantinya menjadi focus KDD. Proses cleaning mencakup diantaranya : membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data. Kesalahan data seperti kesalahan ketik alias (tipografi). Juga perlu dilakukan proses Enrichment yaitu proses "memperkaya" data yang ada dengan data maupun informasi lain yang releven dan diperlukan untuk KDD.

3. Transformation

Coding merupakan proses transformasi pada data yang dipilih, sehingga data tersebut sesuai untuk proses selanjutnya. Proses coding dalam KDD merupakan proses yang inisiatif, kreatif dan sangat bergantung pada pola ataupun jenis informasi yang akan dicari dalam basis data.

4. Data mining

Data mining adalah proses pencarian pola, jenis informasi yang dibutuhkan dalam data terpilih dengan menggunakan teknik maupun metode tertentu. Metode, teknik dan algoritma dalam data mining bermacam-macam. Pemilihan metode yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. Interpretation/ Evaluation

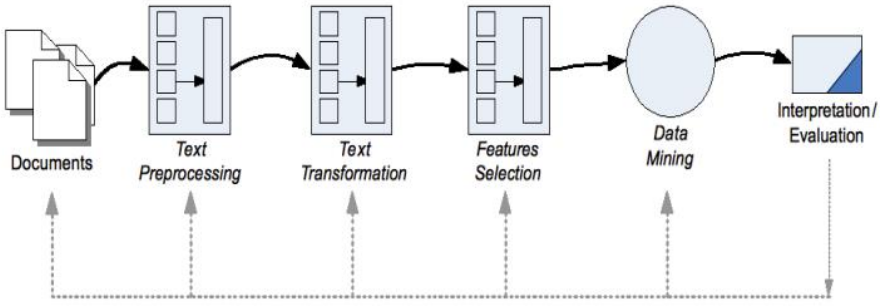
Pola yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang diinginkan agar mempermudah dan dimengerti oleh pihak lain. Tahapan ini adalah bagian dari proses KDD yang disebut dengan Interpretation tahapan ini merupakan tahapan yang mencangkup pemeriksaan apakah pola maupun jenis informasi yang ditemukan bertentangan dengan fakta maupun hipotesa yang ada.

2.2.4 Text Mining

Text mining (disebut juga dengan text data mining) merupakan suatu proses untuk mengambil informasi dari teks yang ada. Text mining digunakan untuk mencari pola-pola yang ada di teks dalam bahasa natural yang tidak terstuktur seperti buku, email, artikel, halaman web, dll. Berdasarkan ketidakteraturan struktur data teks, maka proses text mining memerlukan beberapa tahap awal yang pada intinya adalah mempersiapkan agar teks dapat diubah menjadi lebih terstruktur. Text mining dapat diartikan juga sebagai penemuan

informasi yang baru dan tidak diketahui sebelumnya oleh komputer, dengan begitu secara otomatis dapat mengekstrak informasi dari sumber – sumber yang berbeda (Hamzah 2012).

Text mining memiliki definisi menambang data yang berupa teks dimana sumber data biasanya di dapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen (Bai and Li 2009). Tujuan dari text mining adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan pada text mining adalah kumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari text mining antara lain yaitu pengkategorisasian teks (text categorization) dan pengelompokan teks (text clustering) (Zhang et al. 2016).



Sumber: Lokesh Kumar, 2013

Gambar 2. 3 Tahapan dalam Text Mining

Pada text mining, informasi yang akan digali berisi informasi-informasi yang strukturnya tidak beraturan. Oleh karena itu, diperlukan proses pengubahan bentuk menjadi data yang terstruktur sesuai kebutuhannya untuk proses dalam data mining, yang biasanya akan menjadi nilai-nilai numerik (You, Guo, and Peng 2017). Proses ini sering disebut Text Preprocessing. Teks yang ada distrukturkan dengan proses seperti parsing, dan dimasukkan ke dalam sebuah database. Dilakukan dengan melakukan tokenisasi terhadap dokumen yang menghasilkan kumpulan token kemudian disaring dengan membuang token-token yang ada dalam daftar stopword.

Hasil dari penyaringan token akan dilakukan proses stemming, yaitu proses menghilangkan imbuhan sehingga akan diperoleh kata dasar yang selanjutnya menjadi term-term atau fitur – fitur yang hanya berupa sekumpulan kata. Setelah itu, dilakukan penyusunan dokumen dimana setiap term hasil pemisahan tersebut muncul didalamnya. Hasil penyusunan dokumen berupa frekuensi kemunculan setiap term pada dokumen tersebut.

2.2.5 Media Sosial

Seperti halnya di dunia nyata, perkembangan teknologi memberikan warna yang berbeda di masyarakat, menjalin hubungan juga bisa dilakukan sama siapa saja melalui internet. Di internet seseorang bisa berkenalan dengan siapa pun, kapan pun, latar belakang yang berbeda, suku yang berbeda dan bahkan negara yang berbeda. Intinya jejaring social diciptakan dengan tujuan menduplikasi pola jejaring di dunia nyata akan tetapi dalam lingkup yang sangat luas. Lihat saja Facebook, Twitter, Instagram dan lain sebagainya membuat manusia saling terhubung satu dengan lainnya (Fahrimal 2018). Kebebasan ini memungkinkan penggunaan internet atau jejaring social lebih private untuk setiap akunnya yang mereka punya. Mereka dapat mengupload, menulis, dan berbagi apa saja yang mereka kehendaki. Intinya hasil perkembangan teknologi (jejaring social) menimbulkan dua sisi, sisi yang pertama kehadirannya datap memberikan dampak positif seperti membantu dan memudahkan masyarakat saling terhubung, akan tetapi di sisi lain, internet dan media sosial memiliki dampak negatif ketika berhadapan dengan aspek etika dan moral. Sopan santun dan tata krama dalam kehidupan sehari-hari memiliki aturan yang ketat dan tidak dapat ditoleransi jika dilanggar.

2.2.6 Cyberbullying

Istilah Cyberbullying ditambahkan kedalam kamus OED ((Oxford English Dictionary) pada tahun 2010, merupakan penggunaan teknologi untuk melakukan perilaku yang melecehkan, menghina, mengancam, merendahkan, atau membahayakan seseorang terus menerus dengan memanfaatkan teknologi dan internet dan media sosial (Hidajat et al., 2015). Professor Dan Olweus pada tahun 1993 telah mendefinisikan bullying yang mengandung tiga unsur mendasar perilaku bullying antara lain:

1. Bersifat menyerang dan negative.
2. Dilakukan secara berulang-ulang.
3. Adanya ketidakseimbangan kekuatan antara pihak yang terlibat.

2.2.7 Instagram

Instagram merupakan situs jejaring sosial yang populer di seluruh dunia, Instagram memungkinkan orang untuk berbagi foto, video, dan text dengan orang lain (Ting, 2014). Berkaitan dengan cyberbullying di Instagram, Instagram tidak hanya digunakan untuk berbagi foto, video, dan text dengan orang lain. Tetapi Instagram juga digunakan untuk membully seseorang melalui foto, video maupun berkomentar berupa text (Corliss, 2017). Karakteristik yang dimiliki intagram diantaranya:

1. Length

Hal yang membedakan intagram dengan media social lainnya adalah adanya batasan karakter teks dalam memposting yaitu maksimal sebanyak 2.200 karakter, maka dari itu adanya cukup ruang untuk menambahkan konteks ke postingan melalui teks akan tetapi perlu dicatat juga teks dalam Instagram dipotong menjadi 125 karakter.

2. Data Availability

Adanya Instagram API memberikan kemudahan dalam mengumpulkan jutaan postingan atau lebih untuk training.

3. Languange Model

Model Bahasa yang ada atau disampaikan dalam Instagram berbagai macam, sehingga kemungkinan banyak terdapat ejaan yang salah maupun penggunaan Bahasa 'alay' didalamnya.

4. Hashtag

Pengguna Instagram mengirimkan pesan tentang beberapa topik yang digunakan untuk mengingat percakapan dari seluruh pengguna yang ada kedalam satu aliran.

5. *Domain*

Pengguna Instagram mengirimpan pesan singkat yang digunakan untuk menandai seseorang di dalamnya.

2.2.8 K-Means Clustering

K-Means Clustering adalah suatu metode penganalisaan data atau metode Data Mining yang melakukan proses pemodelan tanpa supervise (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Metode K-means berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok lain. Dengan kata lain, metode ini berusaha untuk meminimalkan variasi antar data yang ada di dalam suatu cluster dan memaksimalkan variasi dengan data yang ada di cluster lainnya (Agusta, 2007).

Rumus Euclidian Distance:

$$d(x,y) = |x - y| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad \textbf{.....(1)}$$

Menurut (Agusta, 2007), langkah-langkah melakukan clustering dengan metode K-means adalah sebagai berikut :

1. Pilih jumlah cluster k.
2. Inisialisasi k pusat cluster ini bisa dilakukan dengan berbagai cara, namun yang paling sering dilakukan adalah dengan cara random. Pusat-pusat cluster diberi nilai awal dengan angka-angka random.
3. Alokasikan semua data atau objek ke cluster terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek tersebut. Demikian juga kedekatan suatu data ke cluster tertentu ditentukan jarak antara data dengan pusat cluster. Dalam tahap ini perlu dihitung jarak tiap data ke tiap pusat cluster. Jarak paling antara satu data dengan satu cluster tertentu akan menentukan suatu data masuk dalam cluster mana. Untuk menghitung jarak semua data ke setiap tiitk pusat cluster dapat menggunakan teori jarak Euclidean yang dirumuskan sebagai berikut:

   D (i,j) = $\sqrt{(X1i - X1j)2 + (X2i - X2j)2 + \cdots + (Xki - Xkj)2}$ **…. (2)**

   Dimana :

   D (i,j) = Jarak data ke i ke pusat cluster j

   Xki = Data ke i pada atribut data ke k

   Xkj = Titik pusat ke j pada atribut ke k
4. Hitung kembali pusat cluster dengan keanggotaan cluster yang sekarang. Pusat cluster adalah rata-rata dari semua data/ objek dalam cluster tertentu. Jika dikehendaki bisa juga menggunakan median dari cluster tersebut. Jadi rata-rata (mean) bukan satu-satunya ukuran yang bisa dipakai.
5. Tugaskan lagi setiap objek memakai pusat cluster yang baru. Jika pusat cluster tidak berubah lagi maka proses clustering selesai. Atau, kembali ke langkah nomor 3 sampai pusat cluster tidak berubah lagi.
6. Hasil cluster dengan metode k-means sangat bergantung pada nilai pusat kelompok awal yang diberikan. Pemberian nilai awal yang berbeda bisa menghasilkan kelompok yang berbeda. Ada beberapa cara memberi nilai awal secara random, menentukan nilai awalnya atau menggunakan hasil dari kelompok hierarki dengan jumlah kelompok yang sesuai.

2.2.9 Confusion Matrix

Teknik evaluasi data merupakan suatu teknik dalam mengukur validitas suatu data. Pada data mining untuk mengukur kinerja dari model yang dibuat dapat dilakukan dengan menggunakan confusion matrix (akurasi). Confusion matrix adalah salah satu metode yang digunakan untuk melakukan perbandingan akurasi pada konsep data mining. Presisi atau

confidence adalah proporsi kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya. Recall atau sensitivity adalah proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar. Pada jenis clustering binary yang hanya memiliki dua kelas seperti tabel 2.2 dibawah ini:

Tabel 2. 2 Tabel Confusion Matrix

Kelas	**Terclustering Positif**	**Terclustering Negatif**
Positif	True Positives (A)	False Negatives (B)
Negatif	False Positives (C)	True Negatives (D)

Perhitungan akurasi dengan tabel confusion matrix adalah sebagai berikut :

Akurasi = (A + D)/(A + B + C + D)……. **(3)**

Presisi di definisikan sebagai rasio Presisi didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih. Presisi dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan tersebut. Rumus presisi adalah:

Presisi = A/(C + A)……. **(4)**

Recall didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia. Recall dihitung dengan rumus:

Recall = A/(A + D)……. **(5)**

Akurasi, Presisi dan Recall dapat diberi nilai dalam bentuk angka dengan menggunakan perhitungan persentase (1-100%) atau dengan menggunakan bilangan antara 0 – 1 . Sistem rekomendasi akan dianggap baik jika nilai presisi dan recallnya tinggi. Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan confusion matrix. ROC adalah grafik dua dimensi dengan false positive sebagai garis horizontal dan true positive sebagai garis vertikal. AUC (the area under curve) dihitung untuk mengukur perbedaan performansi metode yang digunakan. ROC memiliki tingkat nilai diagnosa yaitu:

Tabel 2. 3 Skala Confusion Matrix

Nilai Akurasi %	**Keputusan**
90 - 100	Sangat baik
80 - 90	Baik
70 - 80	Cukup
60 -70	Sedang
< 60	Kurang

2.3 Kerangka Pemikiran

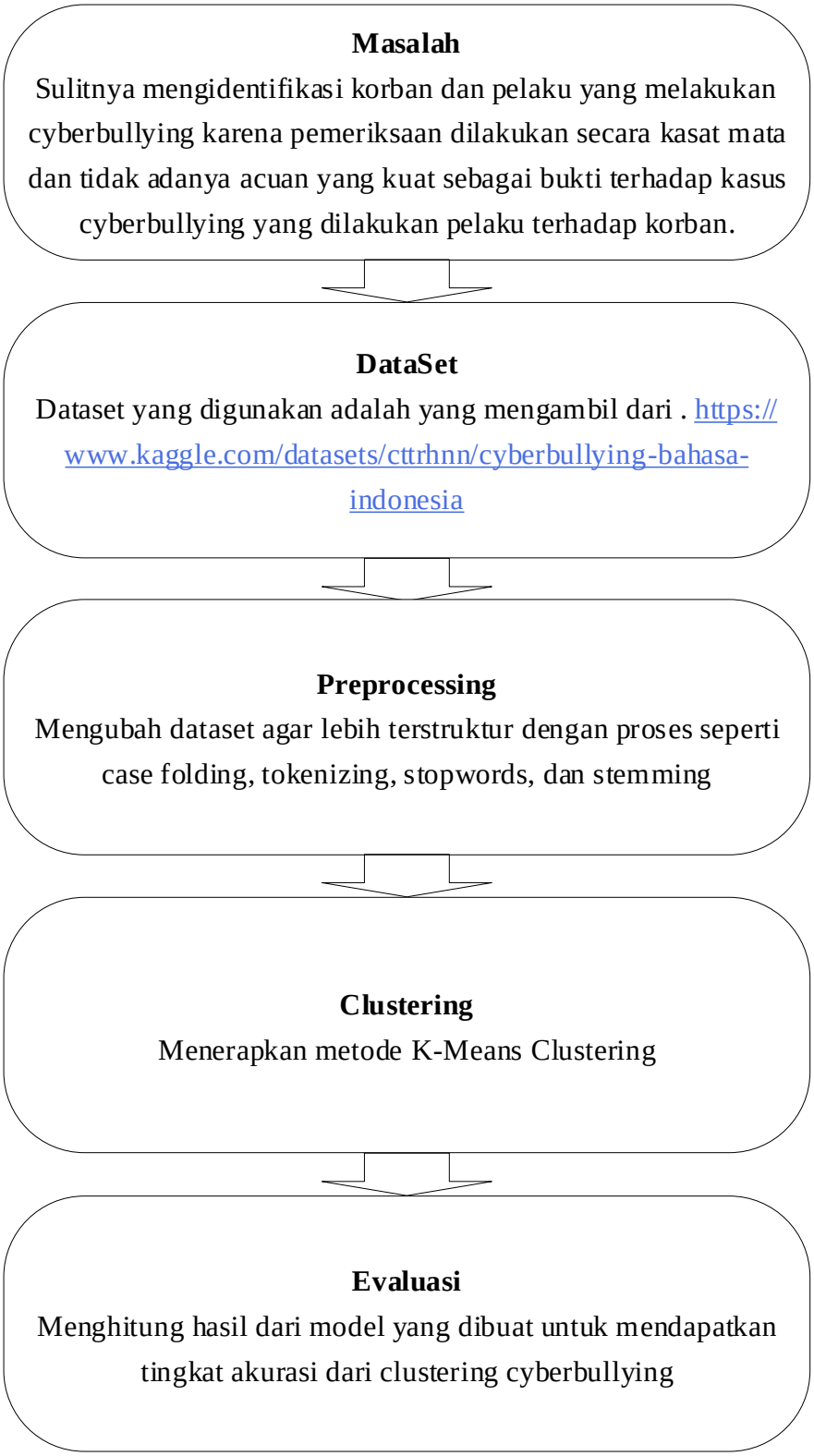
Kementrian Komunikasi Dan Informatika menyatakan bahwa penggunaan internet di indonesia mencapai 175,5 juta orang, 95 persen menggunakan internet untuk mengakses jejaring sosial, situs yang sering diakses adalah Instagram, Facebook dan Twiter. Sebanyak 120 juta orang di indonesia menggunakan perangkat mobile, seperti smartphone atau tablet untuk mengakses media sosial, dengan penetrasi 45 persen dan Instagram menjadi media sosial yang paling umum cyberbullying di internet. Cyberbullying yang dimaksud di sini adalah komentar negatif pada postingan, pesan personal yang tak baik, serta mengolok-olok. Sehingga dibutuhkannya keahlian di bidang forensika digital. Selain itu sering dijumpai permasalahan dalam cyberbullying yaitu sulitnya mengidentifikasi korban dan pelaku yang melakukan cyberbullying karena pemeriksaan dilakukan secara kasat mata dan tidak adanya acuan yang kuat sebagai bukti terhadap kasus cyberbullying yang dilakukan pelaku terhadap korban.

Dari indikator permasalahan diatas perlu diadakan penelitian yang dapat mengatasi masalah tersebut. Dari hasil analisis masalah utama dalam kasus ini adalah untuk mengklompokkan jenis cyberbullying (cyberbullying dan non-cyberbullying) secara otomatis dan tepat sasaran. Data yang akan diolah sendiri adalah berupa dataset yang diambil dari beberapa sumber terutama internet. Dan dataset tersebut sudah ada label didalamnya, label tersebut terdiri dari dua yaitu cyberbullying dan non-cyberbullying. Setelah data didapatkan kemudian akan dilakukan preprocessing (You, Guo, and Peng 2017) yang bertujuan untuk mengurangi berbagai noise data yang dapat menghambat proses pengelompokkan secara otomatis.

Kerangka pemikiran adalah gambaran umum dari proses penelitian yang akan dilakukan. Berdasarkan tujuan dari penelitian ini, peneliti mengusulkan algoritma k-means clustering yang dipadukan dengan pembobotan TF-IDF untuk bisa meningkatkan akurasi dari k-means clustering untuk pengelompokan cyberbullying yang ada diinstagram. Sebelum dilakukan clustering data akan displit atau threshold dari mulai 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, dan 1.0. Berikut ini adalah gambaran umum dari penelitian yang akan dilakukan :

1. Pengumpulan data : mengambil dari dataset yang ada diinternet. Terutama yang sudah ada labelnya.
2. Preprocessing : case folding, tokenizing, stopword, normalisasi, dan stemming.
3. Pengelompokkan : Penerapan metode K-Means Clustering

4. Evaluasi : menerapkan confusion matrix yang digunakan untuk mengevaluasi hasil dari k-means clustering.



Gambar 2. 4 Kerangka Pemikiran

BAB 3

Metodologi

Penelitian yang terkait spesifik membahas clustering cyberbullying. Penelitian ini akan menggunakan algoritma k-means clustering yang dipadukan dengan pembobotan TF-IDF diharapkan dapat meningkatkan akurasi clustering. Selain itu pemilihan fitur yang terbaik juga akan dilakukan untuk menentukan record yang banyak mempengaruhi label atau yang sering muncul pada seluruh dokumen. Selanjutnya langkah – langkah metode penelitian yang akan dilakukan akan diuraikan sebagai berikut

3.1 Pengumpulan data

Dalam penelitian ini dilakukan studi literatur pada beberapa sumber untuk mendapatkan informasi terkait penelitian yang akan dilakukan. Studi literatur dilakukan dengan menggunakan kata kunci (keyword) yang sesuai dengan penelitian yang akan dilakukan. Dalam penelitian ini kata kunci (keyword) dan sumber yang dipilih adalah sebagai berikut :

1. Keyword

   {k-means algorithm}

   {k-means clustering optimation}

   {text mining preprocessing}

   {threshold}

   {term frequency}

   {inverse Document Frequency}

   {tf-idf}

   {confusion matrix}

2. Sumber

   https://www.kaggle.com/datasets/cttrhnn/cyberbullying-bahasa-indonesia

   https://ieeexplore.ieee.org

   https://www.sciencedirect.com

   https://dl.acm.org/dl.cfm

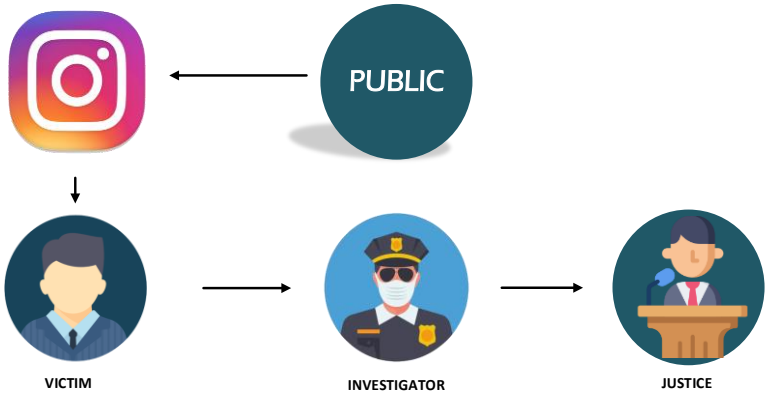
   https://www.google.com

   https://www.youtube.com

3.2 Simulasi Kasus

Simulasi kasus untuk Analisis Cyberbullying Pada Sosial Media Instagram menggunakan K-Means Clustering dapat dilihat sebagaimana pada gambar 3.1. Pada simulasi kasus ini,

Korban diasumsikan melakukan pelaporan kasus kepada penyidik dan selanjutnya ditindaklanjuti dalam bentuk investigasi siber.



Gambar 3. 1 Desain Penelitian

3.3 Alat Penelitian

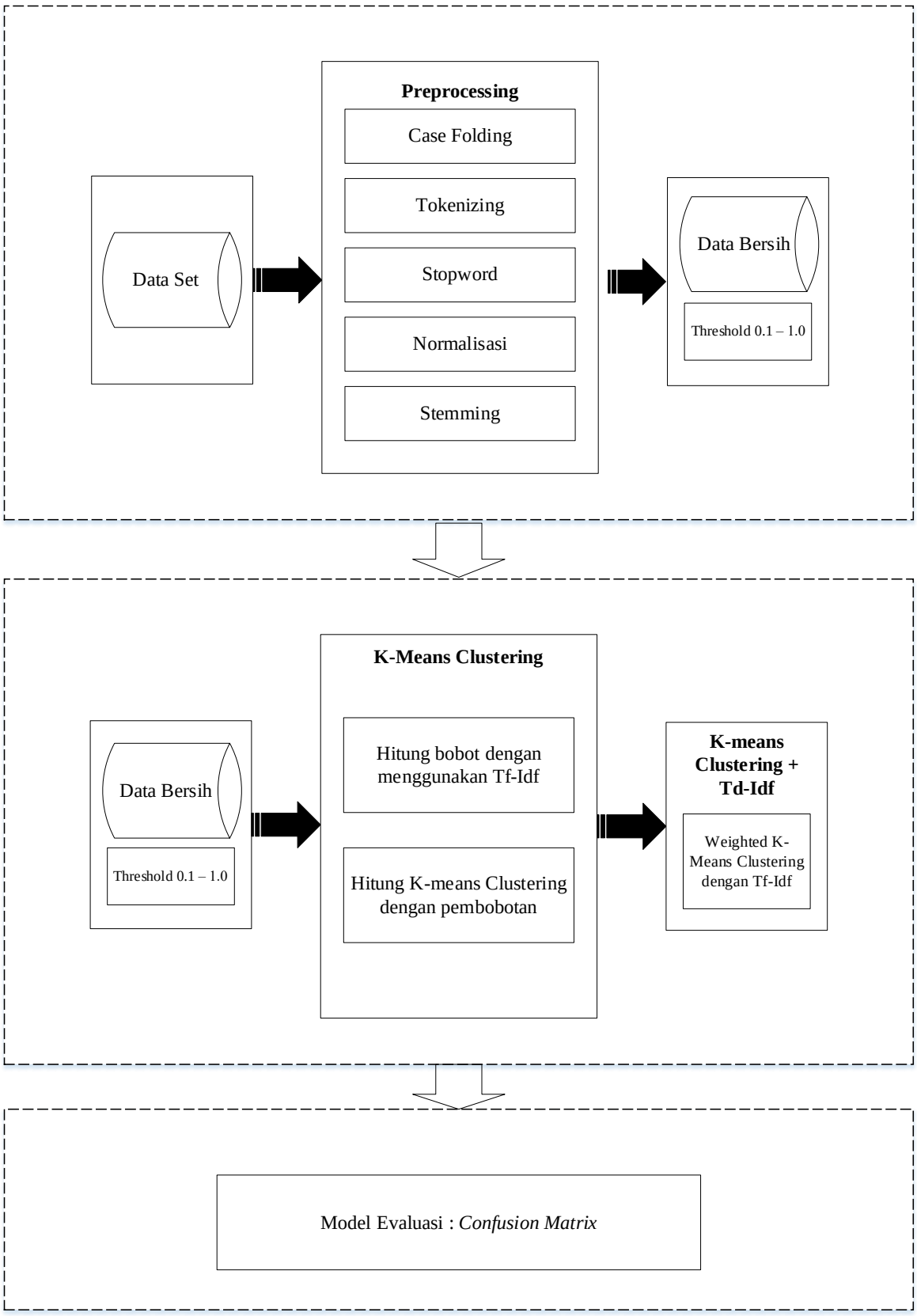
Dalam penelitian ini akan menggunakan alat atau tools antara lain yaitu Jupiter Notebook, Microsoft Excel dan Rapidminer Studio. Masing – masing alat ini akan digunakan sebagai berikut :

1. Jupiter Notebook digunakan untuk memproses data di phyton.
2. Microsoft Excel digunakan untuk proses pembuatan grafik dan sebagai alat bantu ketika data tidak dapat dioleh di Rapidminer atau Jupiter Notebook.
3. Rapidminer akan digunakan untuk proses penghitungan akurasi dan pengolahan data dengan k-means clustering dan pembobotan tf-idf.

3.4 Metodologi Penelitian

Berdasarkan rancangan kerangka pemikiran yang sudah penulis paparkan diatas maka penelitian ini akan mengusulkan algoritma k-means clustering yang dipadukan dengan pembobotan TF-IDF dengan split data atau menentukan nilai threshold terbaik untuk cyberbullying yang ada di Instagram. Metode yang diusulkan belum pernah dilakukan oleh peneliti sebelumnya dengan data yang spesifik yaitu cyberbullying yang ada di Instagram. Selain itu dalam penelitian ini juga berfokus pada tahap preprocessing yang akan dilakukan, karena pada tahap ini akan menghilagkan noise pada data yang akan diolah agar menjadi lebih terstruktur. Hasil dari tahap preprocessing akan berupa nilai numerik sehingga dapat dijadikan sebagai sumber data yang akan diolah lebih lanjut. Hasil dari penelitian ini berupa perbandingan confusion matrix dan split data atau threshold yang berbeda-beda terhadap metode k-means clustering dengan kesimpulan tingkat akurasi paling tinggi adalah

eksperimen yang menghasilkan model terbaik. Berikut ini adalah merupakan gambaran dari metode yang akan diusulkan :



Gambar 3. 2 Desain Penelitian

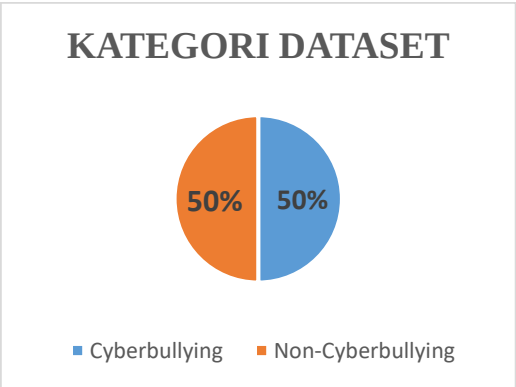
3.5 Data Penelitian

Data yang digunakan jika menggunakan teknik crawling maka akan menimbulkan penilaian yang kurang objektif pada manual labeling, sehingga menggunakan dataset yang ada dari sumber penelitian yang sudah pernah digunakan. Data penelitian ini menggunakan data sekunder yang diambil dari https://www.kaggle.com/datasets/cttrhnn/cyberbullying-bahasa-indonesia . Data yang sudah diperoleh kemudian di export dengan bentuk raw atau excel. Data dalam penelitian ini bersifat open public berisi 650 record. Terdapat 5 atribut dan 1 atribut class yang digunakan dalam penelitian ini yang ditampilkan pada tabel 3.1 dibawah ini :

Tabel 3. 1 Contoh data cyberbullying di Instagram

1	Nama Instagram
2	Komentar
3	Kategori
4	Tanggal Posting
5	Nama Akun Korban

Terdapat satu atribut class / label yang terdapat pada dataset yang didapat yaitu kategori. Dalam dataset cyberbullying tersebut mempunyai atribut yang bernilai kategorial maupun atribut yang numerical, dengan metode k-means clustering akan sangat handal untuk menangani tipe data seperti dataset tersebut untuk proses clustering. Pada dataset tersebut terdapat dua kelas yang seimbang dengan persentase pada kelas cyberbullying 50 persen dan non cyberbullying 50 persen dari data yang ada. Sehingga record data yang digunakan seimbang. Berikut data perbandingan pada dataset cyberbullying yang digunakan dalam penelitian ini seperti pada gambar 3.3 :



Gambar 3. 3 Perbandingan Kelas pada Dataset

Dari diagram perbandingan kelas / label diatas menunjukkan kelas cyberbullying dan non-cyberbullying seimbang. Contoh dataset cyberbullying di social media Instagram akan ditampilkan dalam tabel 3.2 berikut :

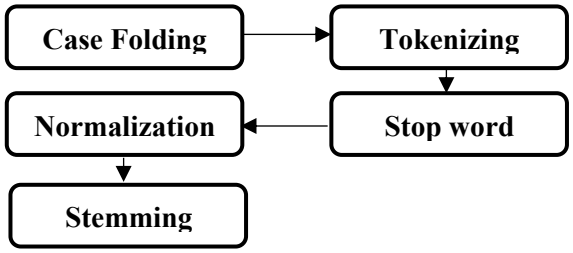
Tabel 3. 2 Contoh Dataset Cyberbullying di Sosial Media Instagram

NAMA INSTAGRAM	KOMENTAR	KATEGORI	TANGGAL POSTING	NAMA AKUN IG ARTIS
@dellianandа	"Kaka tidur yaa, udah pagi, gaboleh capek2"	*Non-bullying*	14 Oktober 2019	@isyanasarasvati
@fenninbl	"makan nasi padang aja begini badannya"	*Non-bullying*	14 Oktober 2019	@isyanasarasvati
@abdurahmanshq	"yang aku suka dari dia adalah selalu cukur jembut sebelum manggung"	*Bullying*	14 Oktober 2019	@isyanasarasvati
@najla.yoo	"Hai kak Isyana aku ngefans banget sama kak Isyana.aku paling suka lagu kak Isyana itu lagu tetap didalam jiwa"	*Non-bullying*	14 Oktober 2019	@isyanasarasvati
@dessy_____	"Manusia apa bidadari sih herann deh cantik terus 😊❤"	*Non-bullying*	14 Oktober 2019	@isyanasarasvati
@e.fril	"@ayu.kinantii isyan skrg berubah ya:(baju nya nakal"	*Bullying*	14 Oktober 2019	@isyanasarasvati
@bahasa.bayi.planet	"Gemesnya isyan kayak tango, berlapis lapis ciaaaa"	*Non-bullying*	16 September 2019	@isyanasarasvati
@khanayarudinita	"Makin jelek aja anaknya, padahal ibu ayahnya cakep2"	*Bullying*	22 Juni 2019	@tasyakamila
@reniaulia225	"Kok anaknya kayak udah tua gitu ya mukanya kk tasya"	*Bullying*	22 Juni 2019	@tasyakamila
@nurjanah.hani	"Muka anak nya ko tua banget yaa.. GK ngegemesin GK ada lucu2nya"	*Bullying*	22 Juni 2019	@tasyakamila
@rizalandriawanp	"Muka nya muka kolot wkwk bukan muka bayi2 lucu gt"	*Bullying*	6 Juni 2019	@randibachtiar
@yudharamm	"Udah gila sekarang ini orang wkwk, udah jomblo jadi psikolog ga laku, ya jadilah seperti itu"	*Bullying*	24 Desember 2020	@lutfiagizal
@mubarakbasslom	"Lo sintingnya udah tingkat dewa"	*Bullying*	24 Desember 2020	@lutfiagizal
@erma_09	"Sepertinya Lutfi setelah putus dgn salsa seperti orang stress"	*Bullying*	24 Desember 2020	@lutfiagizal
@rurrizki_maulana	"ANAK TOLOL INI MULAI AKTIF YA BUND"	*Bullying*	28 Oktober 2020	@lufiagizal
@lala.laxii	"bisa mati aja ga? gedeg gua liat lu. ANJIM"	*Bullying*	8 Desember 2020	@lutfiagizal

NAMA INSTAGRAM	KOMENTAR	KATEGORI	TANGGAL POSTING	NAMA AKUN IG ARTIS
@menisa6634	"Woy Lutfi asu , sok asik Lo ANJING !"	*Bullying*	8 Desember 2020	@lutfiagizal
@farsyahl_	"Apasih kontol berantem aja lah ayo bngst @lutfiagizal"	*Bullying*	6 Desember 2020	@lutfiagizal
@lennyliando	"Ibu kamu pelakor plus hamil di luar nikah ya? Pantesan anaknya kyk gini modelannya"	*Bullying*	16 Juni 2020	@listychanpokemon
@tatianayaaaa	"Anaknya udh pinter, maknya yang kelempengan.. berasa masih baby, tingginya azka aj melebihi mamanya.."	*Bullying*	25 Oktober 2020	@kalinaocktarany
@ferlista234	"Anak yg cerdas terbentuk dr pola asuh sekaligus pola didik ortu yg cerdas pula sih pada saranya. Sesimple itu"	*Non-bullying*	25 Oktober 2020	@kalinaocktarany
@sekiya.sat	"Setengah iri setengah jadi motivasi ini"	*Non-bullying*	25 Oktober 2020	@kalinaocktarany
@emi_muslikah	"Ancur hidup lu kalau sama dia… Sumpah.."	*Bullying*	15 November 2020	@kalinaocktarany
@tetikasmy7615	"serem banget muka lo mel kayak ayam potong merah banget"	*Bullying*	27 Desember 2020	@rosameldianti_
@xcnadilax	"Yaudah jd pengusaha kuliner aja gausah artis, makanannya emang bikin ngiler sih wkwkw"	*Non-bullying*	27 Desember 2020	@rosameldianti_
@ghy.14	"Kamu emang cocoknya jadi pengusaha kuliner udah fixs. Semoha lancat terusssss"	*Non-bullying*	27 Desember 2020	@rosameldianti_
@rubybee_16	"Tak sejelek IMAGE Owner ya..? Artis ups bukan Artis? Hallo secara tidak langsung semua orang itu tau IMAGE lu jelek gentong!"	*Bullying*	27 Desember 2020	@rosameldianti_
@syi_faalatas	"Apaan sih gak jelas hidupnya"	*Bullying*	28 Desember 2020	@rosameldianti_

3.6 Preprocessing Data

Tahap praproses ini dilakukan agar dalam pengelompokan dapat diproses dengan baik. Selain itu, preprocessing dalam penelitain ini juga bertujuan untuk menghilangkan noise pada data yang akan diproses yang dapat mempengaruhi hasil clustering. Berikut ini adalah beberapa metode yang dilakukan untuk tahap preprocessing :



Gambar 3. 4 Teknik Preprocessing Data

Pada umumnya ada 5 langkah preprocessing yang sering digunakan, akan tetapi ada juga langkah lemmatization. Namun dalam penelitian ini digunakan konsep stemming karena dataset yang digunakan berbentuk bahasa indoneisa dan librari yang dipakai adalah Sastrawi. Sedangkan jika penggunakan lemmatization sulitnya mencari library dalam bentuk Bahasa Indonesia.

3.5.1 Case Folding

Case Folding, proses mengubah semua huruf dalam data menjadi huruf kecil. Kemudian karakter selain huruf dan angka dihilangkan dan dianggap delimiter. Contoh hasil data yang sudah di case folding dapat dilihat perbandingan hasil dari tabel 3.2 dengan Tabel 3.3

Tabel 3. 3 Contoh hasil proses case folding

No	User	Komentar
1	@dellianandа	“kaka tidur yaa, udah pagi, gaboleh capek2”
2	@fenninbl	“makan nasi padang aja begini badannya”
3	@abdurahmanshq	“yang aku suka dari dia adalah selalu cukur jembut sebelum manggung”
4	@najla.yoo	“hai kak isyana aku ngefans banget sama kak isyana.aku paling suka lagu kak isyana itu lagu tetap didalam jiwa”
5	@dessy_____	“manusia apa bidadari sih herann deh cantik terus 😊❤”
6	@e.fril	“@ayu.kinantii isyan skrg berubah ya☹ baju nya nakal”
7	@bahasa.bayi.planet	“gemesnya isyan kayak tango, berlapis lapis ciaaaa”
8	@khanayarudinita	“makin jelek aja anaknya, padahal ibu ayahnya cakep2”
9	@reniaulia225	“kok anaknya kayak udah tua gitu ya mukanya kk tasya”
s/d…	…	…
650	@dikha.wirasathya	“inimah bukan main alat 28ahas lagi. olahraga jari dan kaki ini mah”

3.5.2 Tokenizing

Tokenizing, adalah tahapan pemotongan string berdasarkan tiap kata yang menyusunnya. Contoh hasil tokenizing dapat dilihat perbandingan pada tabel 3.3 dengan 3.4 dibawah ini :

Tabel 3. 4 Contoh hasil proses tokenizing

No	User	Komentar
1	@delliananda	‘kaka’, ‘tidur’, ‘yaa’, ‘udah’, ‘pagi’, ‘gaboleh’, ‘capek’
2	@fenninbl	‘makan’, ‘nasi’, ‘padang’, ‘aja’, ‘begini’, ‘badannya’
3	@abdurahmanshq	‘yang’, ‘aku’, ‘suka’, ‘dari’, ‘dia’, ‘adalah’, ‘selalu’, ‘cukur’, ‘jembut’, ‘sebelum’, ‘manggung’
4	@najla.yoo	‘hai’, ‘kak’, ‘isyana’, ‘aku’, ‘ngefans’, ‘banget’, ‘sama’, ‘kak’, ‘isyanaaku’, ‘paling’, ‘suka’, ‘lagu’, ‘kak’, ‘isyana’, ‘itu’, ‘lagu’, ‘tetap’, ‘didalam’, ‘jiwa’
5	@dessy_____	‘manusia’, ‘apa’, ‘bidadari’, ‘sih’, ‘herann’, ‘deh’, ‘cantik’, ‘terus’
6	@e.fril	‘kinantii’, ‘isyan’, ‘skrg’, ‘berubah’, ‘ya’, ‘baju’, ‘nya’, ‘nakal’
7	@bahasa.bayi.planet	‘gemesnya’, ‘isyan’, ‘kayak’, ‘tango’, ‘berlapis’, ‘lapis’, ‘ciaaaa’
8	@khanayarudinita	‘makin’, ‘jelek’, ‘aja’, ‘anaknya’, ‘padahal’, ‘ibu’, ‘ayahnya’, ‘cakep’
9	@reniaulia225	‘kok’, ‘anaknya’, ‘kayak’, ‘udah’, ‘tua’, ‘gitu’, ‘ya’, ‘mukanya’, ‘kk’, ‘tasya’
s/d	…	…
650	@dikha.wirasathya	‘inimah’, ‘bukan’, ‘main’, ‘alat’, ‘musik’, ‘lagi’, ‘olahraga’, ‘jari’, ‘dan’, ‘kaki’, ‘ini’, ‘mah’

3.5.3 Stopword

Stopword merupakan kosakata yang bukan merupakan kata yang uniq atau yang tidak mencerminkan ciri pada suatu dokumen. Hasil perubahan yang dilakukan stopword dapat dilihat pada tabel 3.4. Tabel 3.5 bersisi tiga kolom, kolom pertama berisi nomor, kemudian kolom kedua berisi username dan kolom ketiga berisi komentar yang diungkapkan di Instagram.

Tabel 3. 5 Contoh hasil proses stopword

No	User	Komentar
1	@delliananda	‘kaka’, ‘tidur’, ‘yaa’, ‘udah’, ‘pagi’, ‘gaboleh’, ‘capek’
2	@fenninbl	‘makan’, ‘nasi’, ‘padang’, ‘*aja*’, ‘begini’, ‘badannya’
3	@abdurahmanshq	‘*yang*’, ‘*aku*’, ‘suka’, ‘*dari*’, ‘*dia*’, ‘*adalah*’, ‘*selalu*’, ‘cukur’, ‘jembut’, ‘*sebelum*’, ‘manggung’
4	@najla.yoo	‘hai’, ‘kak’, ‘isyana’, ‘*aku*’, ‘ngefans’, ‘banget’, ‘*sama*’, ‘kak’, ‘isyanaaku’, ‘*paling*’, ‘suka’, ‘lagu’, ‘kak’, ‘isyana’, ‘*itu*’, ‘lagu’, ‘*tetap*’, ‘didalam’, ‘jiwa’
5	@dessy_____	‘manusia’, ‘*apa*’, ‘bidadari’, ‘*sih*’, ‘herann’, ‘*deh*’, ‘cantik’, ‘terus’
6	@e.fril	‘kinantii’, ‘isyan’, ‘skrg’, ‘berubah’, ‘*ya*’, ‘baju’, ‘*nya*’, ‘nakal’
7	@bahasa.bayi.planet	‘gemesnya’, ‘isyan’, ‘kayak’, ‘tango’, ‘berlapis’, ‘lapis’, ‘ciaaaa’

No	User	Komentar
8	@khanayarudinita	‘*makin*’, ‘jelek’, ‘*aja*’, ‘anaknya’, ‘*padahal*’, ‘*ibu*’, ‘ayahnya’, ‘cakep’
9	@reniaulia225	‘*kok*’, ‘anaknya’, ‘kayak’, ‘udah’, ‘tua’, ‘gitu’, ‘*ya*’, ‘mukanya’, ‘kk’, ‘tasya’
s/d	…	…
650	@dikha.wirasathya	‘inimah’, ‘*bukan*’, ‘main’, ‘alat’, ‘musik’, ‘*lagi*’, ‘olahraga’, ‘jari’, ‘*dan*’, ‘kaki’, '*ini*', 'mah'

3.5.4 Normalization

Normalisasi merupkan proses untuk menyamakan kata yang tidak sesuai kaidah dalam Bahasa Indonesia, seperti singkatan, bahasa gaul dan bahasa daerah yang dapat mempengarui hasil dari proses stemming. Dalam penelitian ini proses normalisasi dilakukan secara otomatis dengan memanfaatkan library dan manual dengan mencari sample – sample kata yang tidak sesuai dan akan diganti dengan kata yang sesuai. Dapat dilihat perbandingan tabel 3.5 dengan tabel 3.6 menunjukkan contoh hasil mapping kata – kata yang tidak sesuai dan akan dinormalisasi ke bentuk kata yang sesuai.

Tabel 3. 6 Contoh normalisasi kata

No	User	Komentar
1	@delliananda	'kaka', 'tidur', *'iya*', 'sudah', 'pagi', 'gaboleh', 'capek'
2	@fenninbl	'makan', 'nasi', 'padang', 'badannya'
3	@abdurahmanshq	'suka', 'cukur', 'jembut', 'manggung'
4	@najla.yoo	'hai', 'kak', 'isyana', 'ngefans', 'banget', 'kak', 'isyanaaku', 'suka', 'lagu', 'kak', 'isyana', 'lagu', 'didalam', 'jiwa'
5	@dessy_____	'manusia', 'bidadari', 'herann', 'deh', 'cantik'
6	@e.fril	'kinantii', 'isyan', '*sekarang*', 'berubah', 'baju', 'nakal'
7	@bahasa.bayi.planet	'gemesnya', 'isyan', '*seperti*', 'tango', 'berlapis', 'lapis', 'ciaaaa'
8	@khanayarudinita	'jelek', 'anaknya', 'ayahnya', 'cakep'
9	@reniaulia225	'anaknya', '*seperti*', 'sudah', 'tua', 'gitu', 'mukanya', 'kk', 'tasya'
s/d	…	…
650	@dikha.wirasathya	'inimah', 'main', 'alat', 'musik', 'olahraga', 'jari', 'kaki', 'mah'

3.5.5 Stemming

Stemming, yaitu proses untuk mendapatkan kata dasar dengan cara menghilangkan awalan, akhiran, sisipan dan kombinasi dari awalan dan akhiran atau confixes. Tahap stemming merupakan tahap mencari akar (root). Perbedannya Pada tabel 3.6 dengan tabel 3.7 menunjukkan hasil dari proses stemming yang dilakukan dengan library sastrawi.

Tabel 3. 7 Contoh hasil proses stemming

No	User	Komentar
1	@dellianandа	'kaka', 'tidur', 'iya', 'sudah', 'pagi', 'gaboleh', 'capek'
2	@fenninbl	'makan', 'nasi', 'padang', '*badan*'
3	@abdurahmanshq	['suka', 'cukur', 'jembut', 'manggung'
4	@najla.yoo	'hai', 'kak', 'isyana', 'ngefans', 'banget', 'kak', 'isyanaaku', 'suka', 'lagu', 'kak', 'isyana', 'lagu', '*dalam*', 'jiwa'
5	@dessy_____	'manusia', 'bidadari', 'herann', 'deh', 'cantik'
6	@e.fril	'kinantii', 'isyan', 'sekarang', '*berubah*', 'baju', 'nakal'
7	@bahasa.bayi.planet	'*gemes*', 'isyan', 'seperti', 'tango', 'berlapis', 'lapis', 'ciaaaa'
8	@khanayarudinita	'jelek', *'anak*', '*ayah*', 'cakep'
9	@reniaulia225	'*anak*', 'seperti', 'sudah', 'tua', 'gitu', '*muka*', 'kk', 'tasya'
s/d	…	….
650	@dikha.wirasathya	'inimah', 'main', 'alat', 'musik', 'olahraga', 'jari', 'kaki', 'mah'

3.7 Threshold Data

Threshold data adalah proses membagi dataset menjadi beberapa, penelitian ini akan membagi data yang digunakan dengan 10 fold cross validation untuk mengevaluasi kinerja model atau algoritma. Nilai k yang digunakan akan dibandingkan dari 0.1 sampai 1.0 artinya dari 10 persen data yang ada sampai dengan semua data yang ada. dataset ini akan diolah atau dilakukan preprocessing untuk menghilangkan noise data sebelum proses clustering dan berguna untuk meningkatkan akurasi hasil clustering.

3.8 Term Weighting

Setelah dilakukannya preprocessing dokumen (case folding, tokenization, filtering, dan stemming) dan threshold data, dimana Tahapan preprocessing akan menghasilkan kumpulan term atau kata yang sudah besih, selanjutkan dilakukan proses term weighting yang nantinya akan diberikan bobot atau nilai dimana bobot tersebut mengindikasikan pentingnya sebuah term terhadap dokumen. Penghitungan bobot tiap term dicari pada setiap dokumen bertujuan untuk dapat mengetahui ketersediaan dan kemiripan suatu term di dalam dokumen (Yugianus, Dachlan, and Hasanah 2013). Semakin banyak term tersebut muncul pada koleksi dokumen, semakin tinggi nilai atau bobot term tersebut. Setelah tahapan pemberian bobot selesai barulah dilanjutkan ke proses clustering. Dalam Term Weighting, metode yang digunakan dalam melakukan pembobotan adalah metode Tf-Idf.

Dalam perhitungan bobot menggunakan Tf-Idf, dihitung terlebih dahulu nilai tf perkata dengan bobot masing-masing kata adalah 1. sedangkan nilai idf diformulasikan pada Persamaan (6).

$$\text{Idf} = \log\left(\frac{Jumlah\ seluruh\ dokumen\ dalam\ koleksi}{Jumlah\ dokumen\ yang\ mengandung\ istilah}\right) \quad (6)$$

Dengan demikian maka rumus untuk penghitungan *Tf-Idf* merupakan penggabungan rumus *Tf* dengan rumus *Idf* dengan cara mengalikan nilai *Term Frequency (Tf)* dengan *Inverse Document Frequency (Idf).*

3.9 K-Means Clustering Yang Dipadukan Dengan Pembobotan TF-IDF

Setelah data diolah dengan beberapa tahap preprocessing kemudian data akan di proses dengan menggunakan algoritma weighted k-means clustering untuk mendapatkan hasil clustering cyberbullying sesuai kategorinya. Pada tahap ini setelah data dilakukan preprocessing kemudian akan dicari nilai tf-idf menggunakan persamaan.

3.10 Evaluasi

Dari model yang dihasilkan akan dilakukn evaluasi dengan menggunakan *confusion metrix* untuk melihat seberapa efektif dan akurat model yang dibuat. *Confusion matrix* akan memuat informasi tentang TP, FP, TN, dan FN yang akan berguna untuk melihat hasil clustering yang pada umumnya tidak dapat direpresentasikan hanya bengan satu penilaian saja. Semakin tinggi angka yang dihasilkan pada *Confusion matrix* semakin bagus model yang dihasilkan dan menandakan tingkay akurasi yang tinggi untuk proses clustering. Ada tiga poin yang dijadikan patokan nilai dari hasil model yang didapat, yaitu *accuracy, precision, dan recall.* Evaluasi Pengujian kecocokan objek *image* dengan metode SIFT ini menggunakan penilaian ketepatan (*accuration*) terkait kecocokan objek dari sisi ketelitian (*precision*) dan jumlah perolehan (*recall*). Berikut ini rumus persamaan dari *precision* (1), *recall* (2) dan *accuracy* (3).

$$Precision = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \text{X } 100\% \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

$$Recall\ = \frac{\text{TP}}{\text{TP+FN}} \text{X } 100\% \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

$$Acuracy\ = \frac{\text{TP}}{\text{TP+FP}} \text{X } 100\% \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

Penjelasan :

1. True Positive (TP): Memprediksi kategori yang sudah ada dan kategori sistem dari komentar yang sama ada kecocokan dan benar ada kecocokan.
2. True Negative (TN): Memprediksi kategori yang sudah ada dan kategori sistem dari komentar yg sama tidak ada kecocokan dan benar tidak ada kecocokan.

3. False Positive (FP): Memprediksi kategori yang sudah ada dan kategori sistem dari komentar yang sama ada kecocokan dan ternyata salah tidak ada kecocokan.
4. False Negative (FN): Memprediksi kategori yang sudah ada dan kategori sistem dari komentar yang sama tidak ada kecocokan dan ternyata salah ada kecocokan.

BAB 4
Hasil dan Pembahasan

4.1 Deskripsi Penelitian

Penelitian ini berkaitan dengan clustering cyberbullying pada aplikasi Instagram. Dataset yang akan digunakan adalah data komentar yang bersifat random yang diambil dari beberapa komentar pada beberapa akun Instagram yang lagi trand pada tahun 2020. Model clustering yang digunakan yaitu menggunakan algoritma k-means clustering. algoritma k-means clustering yang dipadukan dengan pembobotan TF-IDF. Tahapan penelitian yang akan dilakukan yaitu pengumpulan data, preprocessing, menghitung peluang kata perkategori, menghitung tf-idf, percobaan model clustering dan evaluasi confusion metrix.

4.2 Dataset

Data yang digunakan jika menggunakan teknik crawling maka akan menimbulkan penilaian yang kurang objektif pada manual labeling, sehingga menggunakan dataset yang ada dari sumber penelitian yang sudah pernah digunakan. Dataset cyberbullying pada sosial media Instagram ini sudah pernah dilakukan dalam penelitian sebelumnya dengan menggunakan klasifikasi. Dataset cyberbullying pada sosial media Instagram bersumber dari https://www.kaggle.com/datasets/cttrhnn/cyberbullying-bahasa-indonesia . Dataset yang akan digunakan ini diambil pada tahun 2020 dengan jumlah 650 records, 5 atribut dan 1 atribut class yang digunakan. Dataset tersebut akan displit atau threshold yang berbeda-beda terhadap metode k-means clustering dengan kesimpulan tingkat akurasi paling tinggi adalah eksperimen yang menghasilkan model terbaik. Tabel 4.1 merupakan metadata dari dataset yang digunakan dalam penelitian ini, sedangkan untuk dataset secara kesuluruhan akan disampaikan dalam lampiran dengan format excel.

Tabel 4. 1 Metadata komentar cyberbullying pada social media Instagram

NAMA INSTAGRAM	KOMENTAR	KATEGORI	TANGGAL POSTING	NAMA AKUN IG ARTIS
@dellianandа	"Kaka tidur yaa, udah pagi, gaboleh capek2"	*Non-bullying*	14 Oktober 2019	@isyanasarasvati
@fenninbl	"makan nasi padang aja begini badannya"	*Non-bullying*	14 Oktober 2019	@isyanasarasvati
@abdurahmanshq	"yang aku suka dari dia adalah selalu cukur jembut sebelum manggung"	*Bullying*	14 Oktober 2019	@isyanasarasvati
@najla.yoo	"Hai kak Isyana aku ngefans banget sama kak Isyana.aku paling suka lagu kak Isyana itu lagu tetap didalam jiwa"	*Non-bullying*	14 Oktober 2019	@isyanasarasvati
@dessy_____	"Manusia apa bidadari sih herann deh cantik terus 😊❤"	*Non-bullying*	14 Oktober 2019	@isyanasarasvati
@e.fril	"@ayu.kinantii isyan skrg berubah ya:(baju nya nakal"	*Bullying*	14 Oktober 2019	@isyanasarasvati
@bahasa.bayi.planet	"Gemesnya isyan kayak tango, berlapis lapis ciaaaa"	*Non-bullying*	16 September 2019	@isyanasarasvati
@khanayarudinita	"Makin jelek aja anaknya, padahal ibu ayahnya cakep2"	*Bullying*	22 Juni 2019	@tasyakamila
@reniaulia225	"Kok anaknya kayak udah tua gitu ya mukanya kk tasya"	*Bullying*	22 Juni 2019	@tasyakamila
@nurjanah.hani	"Muka anak nya ko tua banget yaa.. GK ngegemesin GK ada lucu2nya"	*Bullying*	22 Juni 2019	@tasyakamila
@rizalandriawanp	"Muka nya muka kolot wkwk bukan muka bayi2 lucu gt"	*Bullying*	6 Juni 2019	@randibachtiar
@yudharamm	"Udah gila sekarang ini orang wkwk, udah jomblo jadi psikolog ga laku, ya jadilah seperti itu"	*Bullying*	24 Desember 2020	@lutfiagizal
@mubarakbasslom	"Lo sintingnya udah tingkat dewa"	*Bullying*	24 Desember 2020	@lutfiagizal
@erma_09	"Sepertinya Lutfi setelah putus dgn salsa seperti orang stress"	*Bullying*	24 Desember 2020	@lutfiagizal
@rurrizki_maulana	"ANAK TOLOL INI MULAI AKTIF YA BUND"	*Bullying*	28 Oktober 2020	@lufiagizal
@lala.laxii	"bisa mati aja ga? gedeg gua liat lu. ANJIM"	*Bullying*	8 Desember 2020	@lutfiagizal

NAMA INSTAGRAM	KOMENTAR	KATEGORI	TANGGAL POSTING	NAMA AKUN IG ARTIS
@menisa6634	"Woy Lutfi asu , sok asik Lo ANJING !"	*Bullying*	8 Desember 2020	@lutfiagizal
@farsyahl_	"Apasih kontol berantem aja lah ayo bngst @lutfiagizal"	*Bullying*	6 Desember 2020	@lutfiagizal
@lennyliando	"Ibu kamu pelakor plus hamil di luar nikah ya? Pantesan anaknya kyk gini modelannya"	*Bullying*	16 Juni 2020	@listychanpokemon
@tatianayaaaa	"Anaknya udh pinter, maknya yang kelempengan.. berasa masih baby, tingginya azka aj melebihi mamanya.."	*Bullying*	25 Oktober 2020	@kalinaocktarany
@ferlista234	"Anak yg cerdas terbentuk dr pola asuh sekaligus pola didik ortu yg cerdas pula sih pada saranya. Sesimple itu"	*Non-bullying*	25 Oktober 2020	@kalinaocktarany
@sekiya.sat	"Setengah iri setengah jadi motivasi ini"	*Non-bullying*	25 Oktober 2020	@kalinaocktarany
@emi_muslikah	"Ancur hidup lu kalau sama dia… Sumpah.."	*Bullying*	15 November 2020	@kalinaocktarany
@tetikasmy7615	"serem banget muka lo mel kayak ayam potong merah banget"	*Bullying*	27 Desember 2020	@rosameldianti_
@xcnadilax	"Yaudah jd pengusaha kuliner aja gausah artis, makanannya emang bikin ngiler sih wkwkw"	*Non-bullying*	27 Desember 2020	@rosameldianti_
@ghy.14	"Kamu emang cocoknya jadi pengusaha kuliner udah fixs. Semoha lancat terusssss"	*Non-bullying*	27 Desember 2020	@rosameldianti_
@rubybee_16	"Tak sejelek IMAGE Owner ya..? Artis ups bukan Artis? Hallo secara tidak langsung semua orang itu tau IMAGE lu jelek gentong!"	*Bullying*	27 Desember 2020	@rosameldianti_
@syi_faalatas	"Apaan sih gak jelas hidupnya"	*Bullying*	28 Desember 2020	@rosameldianti_

4.3 Implementasi Preprocessing Data

Sebelum dataset dimasukkan ke dalam model yang akan diusulkan, terlebih dahulu dilakukan preprocessing data. Preprocessing adalah proses pembersihan dan mempersiapkan teks sebelum diolah dalam model. Terdapat beberapa tahapan yang dilakukan untuk preprocessing data, yaitu :

a. Case Folding

Case folding adalah proses yang berguna untuk mengubah dataset menjadi huruf kecil, contoh hasil dari case folding dapat dilihat pada tabel 4.2 dibawah ini :

Tabel 4. 2 Hasil case folding

User	Komentar
@dellianandа	“kaka tidur yaa, udah pagi, gaboleh capek2”
@fenninbl	“makan nasi padang aja begini badannya”
@abdurahmanshq	“yang aku suka dari dia adalah selalu cukur jembut sebelum manggung”
@najla.yoo	“hai kak isyana aku ngefans banget sama kak isyana.aku paling suka lagu kak isyana itu lagu tetap didalam jiwa”
@dessy_____	“manusia apa bidadari sih herann deh cantik terus 😊 ❤ ”
@e.fril	“@ayu.kinantii isyan skrg berubah ya☹ baju nya nakal”
@bahasa.bayi.planet	“gemesnya isyan kayak tango, berlapis lapis ciaaaa”
@khanayarudinita	“makin jelek aja anaknya, padahal ibu ayahnya cakep2”
@reniaulia225	“kok anaknya kayak udah tua gitu ya mukanya kk tasya”
@dikha.wirasathya	“inimah bukan main alat 37ahas lagi. olahraga jari dan kaki ini mah”

b. Tokenizing

Tokenizing adalah tahap untuk pemotongan string berdasarkan tiap kata yang tersusun, proses ini akan menghasilkan term. Tabel 4.3 menunjukkan hasil dari proses tokenizing.

Tabel 4. 3 Hasil tokenizing

Komentar	Tokenizing
kaka tidur yaa, udah pagi, gaboleh capek2	‘kaka’, ‘tidur’, ‘yaa’, ‘udah’, ‘pagi’, ‘gaboleh’, ‘capek’
makan nasi padang aja begini badannya	‘makan’, ‘nasi’, ‘padang’, ‘aja’, ‘begini’, ‘badannya’
yang aku suka dari dia adalah selalu cukur jembut sebelum manggung	‘yang’, ‘aku’, ‘suka’, ‘dari’, ‘dia’, ‘adalah’, ‘selalu’, ‘cukur’, ‘jembut’, ‘sebelum’, ‘manggung’
hai kak isyana aku ngefans banget sama kak isyana.aku paling suka lagu kak isyana itu lagu tetap didalam jiwa	‘hai’, ‘kak’, ‘isyana’, ‘aku’, ‘ngefans’, ‘banget’, ‘sama’, ‘kak’, ‘isyanaaku’, ‘paling’, ‘suka’, ‘lagu’, ‘kak’, ‘isyana’, ‘itu’, ‘lagu’, ‘tetap’, ‘didalam’, ‘jiwa’
manusia apa bidadari sih herann deh cantik terus 😊 ❤	‘manusia’, ‘apa’, ‘bidadari’, ‘sih’, ‘herann’, ‘deh’, ‘cantik’, ‘terus’

Komentar	Tokenizing
@ayu.kinantii isyan skrg berubah ya☹ baju nya nakal	‘kinantii’, ‘isyan’, ‘skrg’, ‘berubah’, ‘ya’, ‘baju’, ‘nya’, ‘nakal’
gemesnya isyan kayak tango, berlapis lapis ciaaaa	‘gemesnya’, ‘isyan’, ‘kayak’, ‘tango’, ‘berlapis’, ‘lapis’, ‘ciaaaa’
makin jelek aja anaknya, padahal ibu ayahnya cakep2	‘makin’, ‘jelek’, ‘aja’, ‘anaknya’, ‘padahal’, ‘ibu’, ‘ayahnya’, ‘cakep’
kok anaknya kayak udah tua gitu ya mukanya kk tasya	‘kok’, ‘anaknya’, ‘kayak’, ‘udah’, ‘tua’, ‘gitu’, ‘ya’, ‘mukanya’, ‘kk’, ‘tasya’
inimah bukan main alat 38ahas lagi. olahraga jari dan kaki ini mah	‘inimah’, ‘bukan’, ‘main’, ‘alat’, ‘musik’, ‘lagi’, ‘olahraga’, ‘jari’, ‘dan’, ‘kaki’, ‘ini’, ‘mah’

c. Stopword

Stopword adalah proses untuk menghilangkan kata – kata yang tidak memiliki nilai atau kontribusi dalam pengklasifikasian teks. Kata – kata yang akan dihilangkan sudah tersimpan di stop list. Dalam penelitian ini juga ditambahkan kata – kata stop list Bahasa jawa yang dianggap perlu dihilangkan. Pada tabel 4.4 adalah proses menambahkan stop list yang akan dihilangkan dari dataset untuk meningkatkan hasil pengelompokkan. Hasil dari stopword dapat dilihat pada tabel 4.5 dibawah ini.

Tabel 4. 4 Stop List Tambahan

Stop List Tambahan untuk Bahasa Jawa dan Singkatan
"yg", "dg", "rt", "dgn", "ny", "d", 'klo', 'kalo', 'amp', 'biar', 'bikin', 'bilang', 'gak', 'ga', 'krn', 'nya', 'nih', 'sih', 'si', 'tau', 'tdk', 'tuh', 'utk', 'ya', 'jd', 'jgn', 'sdh', 'aja', 'n', 't', 'wa', 'wr', 'ass', 'nyg', 'hehe', 'pen', 'u', 'nan', 'loh', 'rt', '&', 'yah', 'sing', 'ana' ,'wis', 'ora','liya', 'ing', 'yen', 'karo', 'aja', 'aku','ala','amarga','amargi','ta','tah','opo','koe','sampean','ngapunten','kulo','badhe','tanglet','menowo','nyuwun','tulung','byk', 'tlg', 'hp', 'sy', 'yth', 'pak', 'pa', 'dr', 'lg', 'trims', 'ga' ,'tksh', 'jgn'

Tabel 4. 5 Hasil Stopword

Komentar	Normalisasi
kaka tidur yaa, udah pagi, gaboleh capek2	‘kaka’, ‘tidur’, ‘yaa’, ‘udah’, ‘pagi’, ‘gaboleh’, ‘capek’
makan nasi padang aja begini badannya	‘makan’, ‘nasi’, ‘padang’, ‘*aja*’, ‘begini’, ‘badannya’
yang aku suka dari dia adalah selalu cukur jembut sebelum manggung	‘*yang*’, ‘*aku*’, ‘suka’, ‘*dari*’, ‘*dia*’, ‘*adalah*’, ‘*selalu*’, ‘cukur’, ‘jembut’, ‘*sebelum*’, ‘manggung’
hai kak isyana aku ngefans banget sama kak isyana.aku paling suka lagu kak isyana itu lagu tetap didalam jiwa	‘hai’, ‘kak’, ‘isyana’, ‘*aku*’, ‘ngefans’, ‘banget’, ‘*sama*’, ‘kak’, ‘isyanaaku’, ‘*paling*’, ‘suka’, ‘lagu’, ‘kak’, ‘isyana’, ‘*itu*’, ‘lagu’, ‘*tetap*’, ‘didalam’, ‘jiwa’
manusia apa bidadari sih herann deh cantik terus 😊 ❤	‘manusia’, ‘*apa*’, ‘bidadari’, ‘*sih*’, ‘herann’, ‘*deh*’, ‘cantik’, ‘terus’
@ayu.kinantii isyan skrg berubah ya☹ baju nya nakal	‘kinantii’, ‘isyan’, ‘skrg’, ‘berubah’, ‘*ya*’, ‘baju’, ‘*nya*’, ‘nakal’

Komentar	Normalisasi
gemesnya isyan kayak tango, berlapis lapis ciaaaa	‘gemesnya’, ‘isyan’, ‘kayak’, ‘tango’, ‘berlapis’, ‘lapis’, ‘ciaaaa’
makin jelek aja anaknya, padahal ibu ayahnya cakep2	‘*makin*’, ‘jelek’, ‘*aja*’, ‘anaknya’, ‘*padahal*’, ‘*ibu*’, ‘ayahnya’, ‘cakep’
kok anaknya kayak udah tua gitu ya mukanya kk tasya	‘*kok*’, ‘anaknya’, ‘kayak’, ‘udah’, ‘tua’, ‘gitu’, ‘*ya*’, ‘mukanya’, ‘kk’, ‘tasya’
inimah bukan main alat 39ahas lagi. olahraga jari dan kaki ini mah	‘inimah’, ‘*bukan*’, ‘main’, ‘alat’, ‘musik’, ‘*lagi*’, ‘olahraga’, ‘jari’, ‘*dan*’, ‘kaki’, '*ini*', 'mah'

d. Normalisasi

Normalisasi adalah proses mengubah struktur kata menjadi seragam atau untuk menyeragamkan term yang memiliki makna sama namun penulisannya berbeda, bisa diakibatkan karena penyingkatan kata, ataupun “bahasa gaul” dengan ini maka dinyatakan, bahwa dari tahun ketahun penyingkatan kata ataupun “Bahasa gaul” akan semakin berkembang. Contoh normalisasi dapat dilihat pada tabel 4.6 dan hasilnya dapat dilihat pada tabel 4.7 dibawah ini:

Tabel 4. 6 Normalisasi Kata

Sebelum	Sesudah
knp	kenapa
mb	mbak
brengsek	berengsek
bangettttt	banget
ky	kaya
yg	yang
knpaa	kenapa
ak	aku
ortu	orang tua
anj	anjing
jablayyy	lonte
k3ntil	kontol
nga	engga
spt	seperti
lont	lonte
ngewe	kawin
nenen	susu
gilakk	gila
asu	anjing

Sebelum	Sesudah
mba	mbak
enaq	enak
Netizennnnn	orang
mnding	mending

Tabel 4. 7 Hasil Normalisasi

Komentar	Normalisasi
kaka tidur yaa, udah pagi, gaboleh capek2	'kaka', 'tidur', *'iya*', 'sudah', 'pagi', 'gaboleh', 'capek'
makan nasi padang aja begini badannya	'makan', 'nasi', 'padang', 'badannya'
yang aku suka dari dia adalah selalu cukur jembut sebelum manggung	'suka', 'cukur', 'jembut', 'manggung'
hai kak isyana aku ngefans banget sama kak isyana.aku paling suka lagu kak isyana itu lagu tetap didalam jiwa	'hai', 'kak', 'isyana', 'ngefans', 'banget', 'kak', 'isyanaaku', 'suka', 'lagu', 'kak', 'isyana', 'lagu', 'didalam', 'jiwa'
manusia apa bidadari sih herann deh cantik terus 😊❤	'manusia', 'bidadari', 'herann', 'deh', 'cantik'
@ayu.kinantii isyan skrg berubah ya☹ baju nya nakal	'kinantii', 'isyan', '*sekarang*', 'berubah', 'baju', 'nakal'
gemesnya isyan kayak tango, berlapis lapis ciaaaa	'gemesnya', 'isyan', '*seperti*', 'tango', 'berlapis', 'lapis', 'ciaaaa'
makin jelek aja anaknya, padahal ibu ayahnya cakep2	'jelek', 'anaknya', 'ayahnya', 'cakep'
kok anaknya kayak udah tua gitu ya mukanya kk tasya	'anaknya', '*seperti*', 'sudah', 'tua', 'gitu', 'mukanya', 'kk', 'tasya'
inimah bukan main alat 40ahas lagi. olahraga jari dan kaki ini mah	'inimah', 'main', 'alat', 'musik', 'olahraga', 'jari', 'kaki', 'mah'

e. Stemming

Pada tahap preprocessing ini akan dilakukan dengan menggunakan library sastrawi yang berguna untuk mengembalikan kata kebentuk dasarnya. Sastrawi dikembangkan berdasarkan aturan Bahasa Indonesia yang kata – katanya menggunakan imbuhan, awalan (prefix), sisipan (infix), akhiran (suffix), dan kombinasi awalan serta akhiran (confixes). Aturan pemenggalan kata dapat dilihat pada tabel 4.8 dibawah ini :

Tabel 4. 8 Aturan Pemenggalan Kata

No	Format Kata	Pemenggalan
1	berV..	Ber-V.. | be-rV..
2	berCAP..	Ber-CAP.. dimana C!- 'r' & P!-'er'
3	berCAerV..	Ber-CaerV.. dimana C!='r'
4	Belajar	Bel-ajar
5	beC$_1$erC$_2$..	beC$_1$erC$_2$.. dimana C$_1$!={'r'|'l'}
6	terV..	Ter-V.. | te-rV..
7	terCerV..	Ter-CerV.. dimana C!='r'
8	terCP…	Ter-CP.. dimana C!='r' dan P!='er'
9	teC$_1$erC$_2$ ..	Te-C$_1$erC$_2$… dimana C$_1$!='r'
10	Me{l|r|w|y}V..	Me-{l|r|w|y}V…
11	Mem{b|f|v}…	Mem-{b|f|v}…
12	Mempe{r|l}…	Mem-pe..
13	Mem{rV|V}…	Me-m{rV|V}… | me-p {rV|V}…
14	Men{c|d|j|z}…	Men-{c|d|j|z}…
15	menV…	Me-nV.. |me-tV..
16	Meng{g|h|q}…	Meng-{g|h|q}…
17	mengV…	Meng-V… |meng-kV…
18	menyV…	Meny-sV…
19	mempV…	mempV… dimana V!='e'
20	Pe{w|y}V…	Pe-{w|y}V…
21	perV…	Per-V…|pe-rV…
22	perCAP..	Per-CAP.. dimana C!='r' dan P!='er'
23	perCAerV…	Per-CAerV…dimana C!='r'
24	Pem{b|f|V}..	Pem-{b|f|V}..
25	Pem{rV|V}…	Pe-m{rV|V}… | pe-p{rV|V}…
26	Pen{c|d|j|z}…	Pen-{c|d|j|z}…
27	penV…	Pe-nV…|pe-lV..
28	Peng{g|h|q}…	Peng-{g|h|q}…

Untuk mempercepat proses stemming pada dataframe dalam penelitian ini digunakan pula library swifter dengan menjalankan task secara parallel. Dengan library ini kecepatan pemprosesan bisa lebih cepat sampai dua kali lipat daripada tanpa menggunakan swifter. Pada tabel 4.9 menunjukkan hasil proses stemming dengan menggunakan library sastrawi dan swifter.

Tabel 4. 9 Hasil Proses Stemming

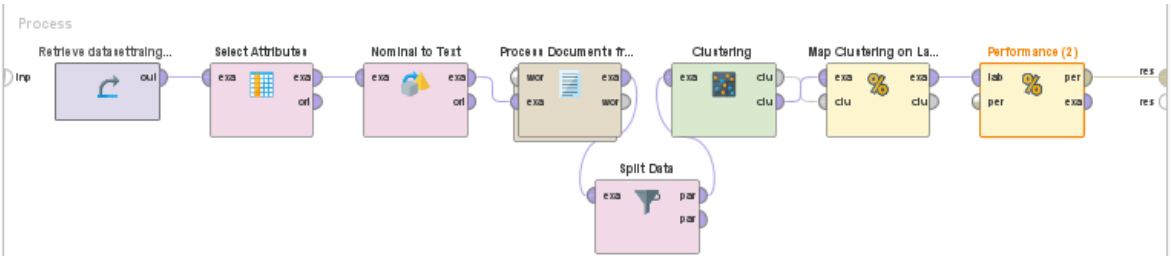
Komentar	**Normalisasi**
kaka tidur yaa, udah pagi, gaboleh capek2	'kaka', 'tidur', 'iya', 'sudah', 'pagi', 'gaboleh', 'capek'
makan nasi padang aja begini badannya	'makan', 'nasi', 'padang', '*badan*'
yang aku suka dari dia adalah selalu cukur jembut sebelum manggung	['suka', 'cukur', 'jembut', 'manggung'
hai kak isyana aku ngefans banget sama kak isyana.aku paling suka lagu kak isyana itu lagu tetap didalam jiwa	'hai', 'kak', 'isyana', 'ngefans', 'banget', 'kak', 'isyanaaku', 'suka', 'lagu', 'kak', 'isyana', 'lagu', '*dalam*', 'jiwa'
manusia apa bidadari sih herann deh cantik terus 😊 ❤	'manusia', 'bidadari', 'herann', 'deh', 'cantik'
@ayu.kinantii isyan skrg berubah ya☹ baju nya nakal	'kinantii', 'isyan', 'sekarang', '*berubah*', 'baju', 'nakal'

Komentar	Normalisasi
gemesnya isyan kayak tango, berlapis lapis ciaaaa	'gemes', 'isyan', 'seperti', 'tango', 'berlapis', 'lapis', 'ciaaaa'
makin jelek aja anaknya, padahal ibu ayahnya cakep2	'jelek', 'anak', 'ayah', 'cakep'
kok anaknya kayak udah tua gitu ya mukanya kk tasya	'anak', 'seperti', 'sudah', 'tua', 'gitu', 'muka', 'kk', 'tasya'
inimah bukan main alat 42ahas lagi. olahraga jari dan kaki ini mah	'inimah', 'main', 'alat', 'musik', 'olahraga', 'jari', 'kaki', 'mah'

Dari proses preprocessing di atas dapat dilihat bahwa, masih banyaknya kata atau trem yang mengandung Bahasa daerah dan Bahasa gaul. Maka diperlukannya update Bahasa daerah dan bahsa gaul didalam proses normalisasi dan stopword. selanjutnyaPreprocessing dilakukan untuk menyiapkan dataset agar lebih terstuktur. Hasil dari preprocessing akan berupa nilai numeric sehingga dapat dijakan sumberdata yang dapat diolah lebih lanjut.

4.4 Implementasi Data Menggunakan RapidMiner

RapidMiner merupakan perangakat lunak yang bersifat terbuka (open source). RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. RapidMiner memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi. RapidMiner merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan munggunakan bahasa java sehingga dapat bekerja di semua sistem operasi. Operator proses pada rapidminer yang digunakan dalam penelitian ini dapat di lihat pada gambar 4.1 di bawah ini.



Gambar 4. 1 Proses RapidMiner

4.5 Split Data atau Threshold Data

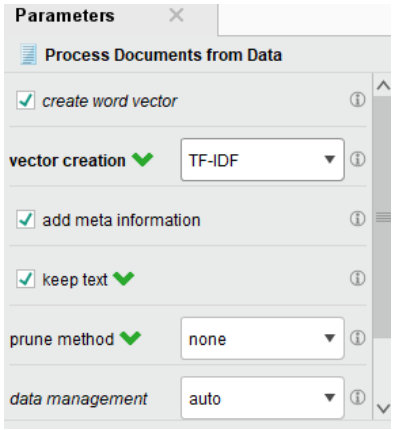
Setelah preprocessing selesai, penelitian ini akan membagi data yang digunakan dengan 10 fold cross validation untuk mengevaluasi kinerja model atau algoritma. Nilai k yang digunakan akan dibandingkan dari 0.1 sampai 1.0 artinya dari 10 persen data yang ada sampai dengan semua data yang ada. dataset ini akan diolah atau dilakukan preprocessing

untuk menghilangkan noise data sebelum proses clustering dan berguna untuk meningkatkan akurasi hasil clustering.

4.6 Pembobotan TF-IDF

Proses pengubahan data teks menjadi data vektor dilakukan dengan membaca kata satu persatu dan menghitung nilai tf-idf. Nilai tf-idf adalah kemunculan kata (term frequency) dalam kalimat dikalikan log jumlah dokumen/record dibagi jumlah dokumen/record yang mengandung kata yang dimaksud.

Pada penelitian ini pembobotan tf-idf dengan menggunakan rapidminer dengan cara tinggal cari operator Process Dokumen From Data lalu dibagian parameter ganti dengan TF-IDF. Lebih jelasnya dapat dilihat pada gambar 4.2 dan Setelah dilakukan tahapan preprocessing sebanyak 650 data lalu dengan mengubah term menjadi data vektor melalui perkalian Tf*Idf, maka didapatkanlah 1,377 term atau kata. Berikut term yang didapatkan, setelah hasil preprocessing dapat dilihat pada tabel 4.10 dibawah ini.



Gambar 4. 2 Pembobotan Tf-Idf

Tabel 4. 10 Term Dari Preprocessing

No	Term	No	Term
1	amin	9	adek
2	abai	10	adik
3	abal	11	aduh
4	activity	12	aduhh
5	adab	13	aduuh
6	adam	14	aesthetic
7	adaptasi	…	…
8	adek	1377	yutuber

Tahapan selanjutnya adalah proses prekalian antara Term Frequency (Tf) dengan Inverse Document Frequency (Idf). Berikut hasil proses Tf*Idf dapat dilihat pada Tabel 4.11.

Tabel 4. 11 Hasil Tf-Idf

Doc	amin	abai	abal	activity	adab	s/d	yutuber
1	0	0	0	0	0	…	0
2	0	0	0	0	0	…	0
3	0	0	0	0	0	…	0
4	0	0	0	0	0.370	…	0
5	0	0	0	0	0.305	…	0
6	0	0	0	0	0	…	0
7	0	0	0	0	0.232	…	0
8	0	0	0	0	0.221	…	0
9	0	0	0	0	0.154	…	0
10	0	0	0	0	0	…	0
s/d	…	…	…	…	…	…	0
650	0	0	0	0	0	…	0

4.7 K-Means Clustering yang Dipadukan Dengan Pembobotan TF-IDF

Setelah proses preprocessing selesai dan mengubah term menjadi data vektor melalui perkalian Tf*Idf, maka selanjutnya dilakukan proses clusterisasi dengan menggunakan K-Means. Hasil dari Clustering sistem dengan menggunakan K-Means, dapat dilihat pada gambar 4.3 dan gambar 4.4.

id	..	cluster	text
1	.	cluster_0	kaka tidur pagi gaboleh capek
2	.	cluster_0	makan nasi padang badan
3	.	cluster_0	suka cukur jembut manggung
4	.	cluster_0	isyana ngefans banget isyanaaku suka lagu isyana lagu jiwa
5	.	cluster_0	manusia bidadari herann cantik
6	.	cluster_0	kinanti isyan ubah baju nakal
7	.	cluster_0	gemesnya isyan tango lapis ciaaaa
8	.	cluster_0	jelek anak ayah cakep
9	.	cluster_0	anak gitu muka tasya
10	.	cluster_0	muka anak banget ngegemesin lucu

Gambar 4. 3 Hasil Cluster 0

Berdasarkan informasi hasil clustering sebagaimana pada gambar 4.3 didapat hasil clustering yang dilakukan. Hasilnya adalah cluster 0 memiliki data sebanyak 57 record dan data tersebut berada pada komentar yang mengandung unsur non-cyberbullying.

id	..	cluster ↓	text
17	.	cluster_1	lutfi anjing asik anjing
53	.	cluster_1	bagus pantat panci aing atuh anjing
122	.	cluster_1	mata katarak anjing ahahah
177	.	cluster_1	menderitaaa lora anjing
180	.	cluster_1	gaga kayak anjing
193	.	cluster_1	bagus anjing wkwkw
206	.	cluster_1	manusia anjing cari uang gitu kerja nyebokin sebentar gaji
208	.	cluster_1	anjing anak nang
211	.	cluster_1	anjing baku hantam

Gambar 4. 4 Hasil Cluster 1

Sementara berdasarkan informasi pada gambar 4.4, didapat hasil bahwa cluster 1 memiliki data sebanyak 9 record dan data tersebut berada pada komentar yang memiliki unsur cyberbullying.

4.8 Evaluasi

Skenario pengujian metode dengan menggunakan k-means clustering dengan menggunakan pembobotan tf-idf. Semua metode diuji dengan menggunakan 10-fold cross validation dengan cara menyeplit data atau Threshold Data dengan Nilai k yang digunakan akan dibandingkan dari 0.1 sampai 1.0 artinya dari 10 persen data yang ada sampai dengan semua data yang ada untuk mengetahui kinerja model atau algoritma yang digunakan.

4.8.1 Analisis Pengujian K-Means Clustering dengan Pembobotan Tf-Idf

Pada pengujian dengan menggunakan metode K-Means Clustering dengan Pembobotan Tf-Idf. Pengujian dengan pembobotan Tf-Idf dilakukan dengan menentukan jumlah Record yang akan digunakan. Jumlah atribut ditentukan mulai dari (R) = 0.1 sampai 1.0 artinya artinya dari 10 persen Record yang ada sampai dengan semua Record yang ada untuk mengetahui kinerja model atau algoritma yang digunakan. Penentuan Record ini dilakukan untuk mencari nilai akurasi terbaik berada pada jumlah seleksi Record berapa. Berikut ini adalah hasil dari confusion matrix yang diperoleh untuk jumlah R = 0.1 seperti yang ditampilkan di tabel 4.12 dibawah ini :

Tabel 4. 12 Hasil Confusion Matrix Algoritma K-Means Clustering + Tf-Idf untuk R = 0.1

		Aktual (Real)		**Total Prediksi**
		C1	C2	
Prediksi (sistem)	C1	32	25	**57**
	C2	1	8	**9**
Total Aktual		**33**	**33**	**N=66**

Berdasarkan tabel 4.12 diatas dapat dihitung nilai akurasi, presisi dan recall yang dapat dihasilkan dilakukan pada percobaan pertama dengan nilai R = 0.1. Persamaan untuk menghitung nilai akurasi dari clustering multi class adalah sebagai berikut :

$$akurasi = \sum_{i=1}^{i} \left(\frac{TP_i + TN_i}{Tp_i + TN_i + FP_i + FN_i} \right)$$

$$\text{Akurasi} = \frac{32 + 8}{32 + 25 + 1 + 8}$$

$$\text{Akurasi} = \frac{40}{66} x\ 100\% = 60{,}61\%$$

Sedangkan untuk menghitung nilai recall adalah dengan nilai FP dari sistem clustering multi class adalah sebagai berikut :

$$presisi = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} (FP_i + TP_i)}\ x\ 100\%$$

Tabel 4. 13 Nilai FP Masing – masing Kelas

Jumlah FP	**Kelas**
25	Non-Cyberbullying
1	Cyberbullying

Selanjutnya setelah menemukan nilai dari masing – masing FP untuk menghitung nilai recall masing – masing kelas dan jumlah semua hasil rata – rata tersebut dibagi dengan jumlah kelas seperti dibawah ini :

$$P\ (Non - Cyberbullying) = \frac{32}{32 + 25} x100\% = 56{,}14\%$$

$$P\ (Cyberbullying) = \frac{8}{8 + 1} x100\% = 88{,}89\%$$

$$all\ recall = \frac{\sum P(kelas)}{jumlah\ kelas}\ x\ 100\%$$

$$all\ recall = \frac{0{,}56 + 0{,}89}{2} x100\% = 72{,}5\%$$

dan selanjutnya persamaan untuk menghitung nilai precision dari sistem clustering multi class adalah sebagai berikut :

Tabel 4. 14 Nilai FN Masing – masing Kelas

Jumlah FN	Kelas
1	Non-Cyberbullying
25	Cyberbullying

$$P\ (Non - Cyberbullying) = \frac{32}{32 + 1} x100\% = 96{,}97\%$$

$$P\ (Cyberbullying) = \frac{8}{8 + 25} x100\% = 24{,}24\%$$

$$all\ precision = \frac{\sum P(kelas)}{jumlah\ kelas}\ x\ 100\%$$

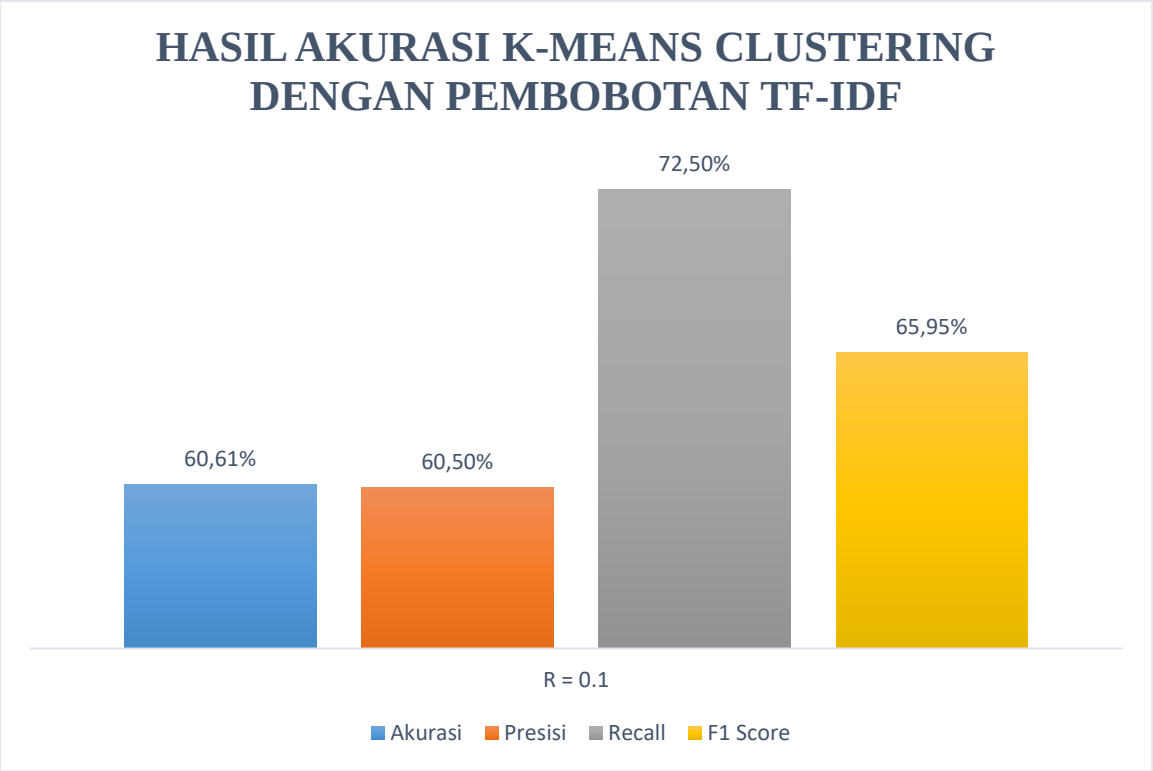
$$All\ Presisi = \frac{0{,}97 + 0{,}24}{2} x100\% = 60{,}5\%$$

F1 score digunakan untuk memperbaiki performa clustering agar tidak menyesatkan kinerja dari suatu klasifikasi. F1 score merupakan matrix yang baik digunakan ketika ada data yang tidak seimbang (imbalance), sehingga f1 score akan menjadi nilai rata – rata yang harmonis dari recall dan precision. Berikut adalah cara untuk menghitung nilai f1 score :

$$F1\ Score\ = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}\ x\ 100\%$$

$$F1\ Score = 2\ x\ \frac{0{,}605\ x\ 0{,}725}{0{,}605 +\ 0{,}725} x100\% = 65{,}95\%$$

Berdasarkan hasil percobaan dengan menggunakan (R) = 0.1 sampai 1.0 artinya artinya dari 10 persen Record yang ada sampai dengan semua Record yang ada untuk mengetahui kinerja model atau algoritma yang digunakan. Hasil terbaik ditemukan pada pengujian R = 0.5 dengan nilai akurasi sebesar 67.38 %. Gambar 4.3 menunjukkan hasil dari dari clustering dengan nilai R = 0.1. Hasil percobaan setiap jumlah atribut R yang ditentukan dapat dilihat pada tabel 4.15 dibawah ini.

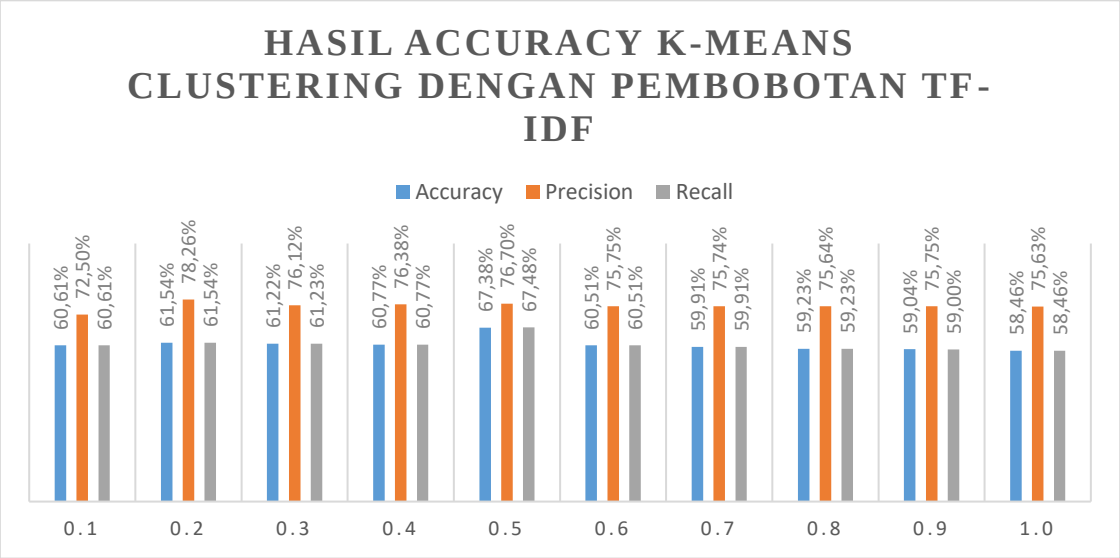


Gambar 4. 5 Hasil akurasi dengan R = 0.1

Tabel 4. 15 Hasil Percobaan R = 0.1 sampai R = 1.0

Percobaan Ke	Threshold (R)	Nilai Accuracy (%)	Precision (%)	Recall (%)
1	0.1	60.61%	72.5%	60.61%
2	0.2	61.54%	78.26%	61.54%
3	0.3	61.22%	76.12%	61.23%
4	0.4	60.77%	76.38%	60.77%
5	**0.5**	**67.38%**	**76.70%**	**67.48%**
6	0.6	60.51%	75.75%	60.51%
7	0.7	59.91%	75.74%	59.91%
8	0.8	59.23%	75,64%	59,23%
9	0.9	59.04%	75,75%	59,00%
10	1.0	58.46%	75,63%	58,46%

Dari hasil banyaknya percobaan yang telah dilakukan dapat dilihat bahwa pada pengujian dengan nilai threshold 0,5 memiliki tingkat accuracy, rata-rata precision dan recall paling tinggi yaitu 67,38% untuk accuracy nya 76,70% untuk precision nya sedangkan untuk recall nya sebesar 67,48% dari pengujian data yang lainnya. ini berarti bahwa penempatan data pada setiap cluster nya di pengujian dengan nilai threshold 0,5 kebanyakan sudah tepat dan dapat diartikan dalan tabel skala Confusion Matrix dikatakan cukup baik.



Gambar 4. 6 Akurasi K-Means Clustering dengan Pembobotan Tf-Idf

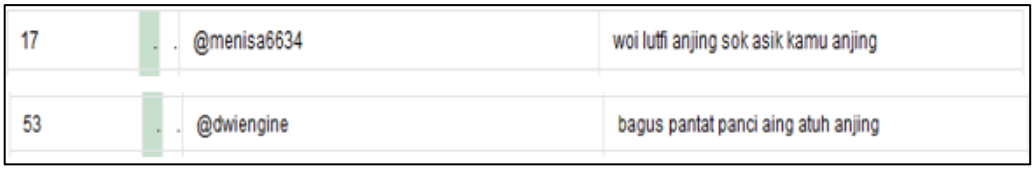
Berdasarkan hasil yang telah disampaikan dengan menggunakan sebanyak 650 record, maka terlihat bahwa algoritma K-Means mampu berjalan sesuai tujuan yang diharapkan. Algoritma K-Means merupakan salah satu algortima yang dianjurkan dan banyak digunakan untuk analisa pada media sosial Instagram (Habibi and Cahyo 2019). Penelitian ini menggunakan nilai K=2 dari jumlah cluster yang telah ditentukan tersebut. Pemroses data dilakukan menggunakan software rapidminer. Hasil penelitian mendapatlkan informasi bahwa pada pengujian dengan nilai threshold 0.5 menghasilkan cluster 0 yang terbentuk beranggotakan 67 record. Sementara cluster satu yang terbentuk beranggotakan 258 record. Informasi lain yang diperoleh yaitu terkait nilai accuracy, rata-rata precision, dan recall. Tingkat paling tinggi yaitu sebesar 67,38% untuk accuracy, 76, kemudian 70% untuk precision, dan 67,48% untuk recall dan dapat diartikan dalan tabel skala Confusion Matrix dikatakan cukup baik. Hasil tersebut didapatkan dari pengujian data yang lainnya. Sedangkan dalam penelitian sebelumnya (Luqyana, Cholissodin, and Perdana 2018) hasil pengujian cyberbullying pada komentar Instagram dengan memanfaatkan Metode Klasifikasi Support Vector Machine dengan 400 record data menghasilkan tingkat akurasi terbaik sebesar 90%, precision sebesar 94,44% dan recall sebesar 85%. Sehingga dapat

dikatakan bahwa penelitian dengan metode k-means clustering menghasilkan tingkat akurasi yang lebih rendah jika dibadingkan dengan penelitian sebelumnya yaitu menggunakan metode Support Vector Machine. Akurasi dengan metode k-means lebih rendah dari pada algoritma SVM (Support Vector Machine) dikarenakan beberapa hal yaitu karena sedikitnya kombinasi dari data yang diperoleh, jumlah dataset yang digunakan dan ragam karakteristik data yang digunakan.

Terdapat kendala yang dihadapi dalam penelitian ini ketika proses analisis text berbahasa Indonesia. Hal ini dikarenakan corpus bahasa Indonesia yang lengkap masih sulit diperoleh. Selain itu seringkali komentar-komentar atau cuitan-cuitan dalam Instagram banyak menggunakan bahasa daerah dan hahasa gaul. Kendala ini menyebabkan hasil yang diperoleh dari proses data mining yang dilakukan masih belum mencapai tingkat akurasi yang diinginkan. Dikarenakan masih sulitnya mendapatkan corpus yang lengkap maka disarankan untuk selalu mengupdate corpus tersebut supaya menyesuaikan dengan Bahasa daerah dan Bahasa gaul yang lagi viral pada masanya.

4.8.2 Deteksi Kasus pada Simulasi

Informasi hasil clustering yang didapat ditunjukkan pada gambar 4.3 dan gambar 4.4. Terdapat 3 kolom yang disajikan pada kedua gambar tersebut. Kolom pertama menampilkan id, kolom kedua menampilkan cluster dan kolom ketiga menampilkan teks atau komentar instagram. Sebagai contoh pada cluster 1 memiliki data sebanyak 9 record. Data tersebut berada pada komentar yang memiliki unsur cyberbullying. Hasil dari penelitian yang telah dilakukan yaitu komentar dengan id 17 dan 53 terindikasi sebagai komentar cyberbullying. Dapat dilihat pada gambar



Gambar 4. 7 Contoh id terindikasi sebagai komentar bullying yang sudah diklarifikasi dengan data.

Output yang dihasilkan pada cluster 1 tersebut yang berindikasi masuk kedalam kategori bullying, perlu adanya tahapan berikutnya dimana pada hasil tersebut hanya menampilkan id nya. Dari id tersebut selanjutnya diklarifikasi dengan dataset yang ada sehingga menemukan username dan komentar pelaku yang melakukan Tindakan cyberbullying. Hasil tersebut dapat digunakan sebagai bukti digital awal untuk keperluan persidangan.

BAB 5
Kesimpulan dan Saran

5.1 Kesimpulan

Pada pengelompokkan komentar pada media sosial instagram menggunakan metode k-means clustering untuk identifikasi awal cyberbullying. Dataset yang digunakan adalah data komentar yang bersifat random yang diambil dari beberapa komentar pada beberapa akun Instagram yang sedang trand pada tahun 2020 dilakukan dengan menggunakan data sejumlah 659 records, dengan masing – masing kategori memiliki jumlah data yang sama yaitu sebesar 325 data. Model clustering yang dibuat menggunakan dua kombinasi metode. Model yang digunakan adalah algoritma k-means clustering yang dipadukan dengan pembobotan TF-IDF. Selain itu juga dilakukan penelitian terhadap jumlah record yang digunakan. Dari penelitian yang telah dilakukan, dapat diambil kesimpulan sebagai berikut :

1. Dapat dikatakan bahwa Metode algoritma k-means clustering yang dipadukan dengan pembobotan TF-IDF mampu mengelompokkan komentar ke dalam dua kelompok yaitu non-cyberbullying dan cyberbullying. Atribut terbaik dari hasil percobaan dengan nilai *threshold* 0.5 Pada beberapa kelompok data yang diuji, memiliki tingkat accuracy, rata-rata precision dan recall paling tinggi yaitu 67,38% untuk accuracy nya 76,70% untuk precision nya sedangkan untuk recall nya sebesar 67,48%.
2. Dalam mengidentifikasi tindakan awal *cyberbullying* sebagai contoh pada cluster 1 memiliki data sebanyak 9 record. Data tersebut berada pada komentar yang memiliki unsur cyberbullying. Hasil dari penelitian yang telah dilakukan yaitu komentar dengan id 17 terindikasi sebagai komentar cyberbullying. Hasil tersebut dapat digunakan sebagai bukti digital awal untuk keperluan persidangan.

5.2 Saran

Pada penelitian ini nilai akurasi yang didapatkan sudah cukup bagus. Namun untuk meningkatkan tingkat akurasi dari klasifikasi aduan masyarakat ini ada beberapa saran yang dapat dilakukan sebagai berikut :

1. Melakukan normalisasi data dengan maksimal untuk menghilangkan noise data yang ada ada dataset karena sangat penting untuk memaksimalkan hasil dari clustering.

2. Menambah kamus data untuk normalisasi kata. Hal ini disarankan karena masih banyak kata yang tidak standar atau mengandung Bahasa daerah yang dapat mempengaruhi hasil dari pembobotan tiap kata.
3. Menggunakan metode untuk menentukan nilai threshold (R) terbaik, sehingga tidak perlu melakukan banyak percobaan untuk mengetahui berapa banyak atribut terbaik yang akan digunakan untuk menghasilkan akurasi tertinggi.

Daftar Pustaka

Adiya, M Hasmil, And Yenny Desnelita. 2019. "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada Rsud Pekanbaru." 01: 17–24.

Al-Rahmi, Waleed Mugahed, Noraffandy Yahaya, Mahdi M Alamri, And Nada Ali. 2018. "A Model Of Factors Affecting Cyber Bullying Behaviors Among University Students." *Ieee Access* Pp(November): 1.

Andriansyah, Miftah Et Al. 2018. "Cyberbullying Comment Classification On Indonesian Selebgram Using Support Vector Machine Method." *Proceedings Of The 2nd International Conference On Informatics And Computing, Icic 2017* 2018-Janua: 1–5.

Bai, Ping, And Junqing Li. 2009. "The Improved Naive Bayesian Web Text Classification Algorithm." *Proceedings - 1st International Symposium On Computer Network And Multimedia Technology, Cnmt 2009*.

Casey, Eoghan. 2011. Paper Knowledge . Toward A Media History Of Documents *Digital Evidence And Computer Crime*.

Chukwuere, Precious Chibuike, Joshua Ebere Chukwuere, And Dickson Adom. 2021. "The Psychosocial Effects Of Social Media Cyberbullying On Students In Selected African Countries." *Acta Informatica Malaysia* 5(2): 62–70.

Fahrimal, Yuhdi. 2018. "Netiquette: Etika Jejaring Sosial Generasi Milenial Dalam Media Sosial." *Jurnal Penelitian Pers Dan Komunikasi Pembangunan* 22(1): 69–78.

Fazry, Laila, And Nurliana Cipta Apsari. 2021. "Pengaruh Media Sosial Terhadap Perilaku Cyberbullying Di Kalangan Remaja." *Jurnal Pengabdian Dan Penelitian Kepada Masyarakat* 2(1): 28–36. Https://Ejournal.Bsi.Ac.Id/Ejurnal/Index.Php/Cakrawala/Article/Viewfile/3680/2624.

Gorro, Kim D. Et Al. 2018. "Classification Of Cyberbullying In Facebook Using Selenium And Svm." *2018 3rd International Conference On Computer And Communication Systems, Icccs 2018*: 233–38.

Habibi, Muhammad, And Puji Winar Cahyo. 2019. "Clustering User Characteristics Based On The Influence Of Hashtags On The Instagram Platform." *Ijccs (Indonesian Journal Of Computing And Cybernetics Systems)* 13(4): 399.

Hamzah, Amir. 2012. "Klasifikasi Teks Dengan Naïve Bayes Classifier (Nbc) Untuk Pengelompokan Teks Berita Dan Abstract Akademis." *Prosiding Seminar Nasional*

Aplikasi Sains & Teknologi (Snast) Periode Iii (2011): 269–77.

Hang, Ong Chee, And Halina Mohamed Dahlan. 2019. "Cyberbullying Lexicon For Social Media." *International Conference On Research And Innovation In Information Systems, Icriis* December-2.

Imam Riadi, Sunardi, Panggah Widiandana. 2021. "Investigasi Cyberbullying Pada Whatsapp Menggunakan Digital Forensics." *Rekayasa Sistem Dan Teknologi Informasi* 1(10): 730–35.

Ishara Amali, H. M.A., And Shantha Jayalal. 2020. "Classification Of Cyberbullying Sinhala Language Comments On Social Media." *Mercon 2020 - 6th International Multidisciplinary Moratuwa Engineering Research Conference, Proceedings*: 266–71.

Kementerian Komunikasi Dan Informatika. 2021. "Warganet Meningkat, Indonesia Perlu Tingkatkan Nilai Budaya Di Internet." Https://Aptika.Kominfo.Go.Id/2021/09/Warganet-Meningkat-Indonesia-Perlu-Tingkatkan-Nilai-Budaya-Di-Internet/ (May 2, 2022).

Lei Xun Et Al. 2014. "Information Security In Big Data: Privacy And Data Mining." *Ieee Access* 2: 1149–76.

Luqyana, Wanda Athira, Imam Cholissodin, And Rizal Setya Perdana. 2018. "Analisis Sentimen Cyberbullying Pada Komentar Instagram Dengan Metode Klasifikasi Support Vector Machine." *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-Ptiik) Universitas Brawijaya* 2(11): 4704–13.

Mohan, Ashok Kumar, And D. Venkataraman. 2017. "Forensic Future Of Social Media Analysis Using Web Ontology." *2017 4th International Conference On Advanced Computing And Communication Systems, Icaccs 2017*.

Naf'an, Muhammad Zidny Et Al. 2019. "Sentiment Analysis Of Cyberbullying On Instagram User Comments." *Journal Of Data Science And Its Applications* 2(1): 88–98.

Nurrahmi, Hani. 2018. "Indonesian Twitter Cyberbullying Detection Using Text Classification And User Credibility." : 543–48.

Pawar, Rohit, And Rajeev R. Raje. 2019. "Multilingual Cyberbullying Detection System." *Ieee International Conference On Electro Information Technology* 2019-May: 040–044.

Riadi, Imam, Sunardi Sunardi, And Panggah Widiandana. 2020. "Mobile Forensics For Cyberbullying Detection Using Term Frequency - Inverse Document Frequency (Tf-

Idf).” *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika* 5(2): 68.

Suryanto, Joko. 2017. “Analisa Perbandingan Pengelompokkan Curah Hujan 15 Harian Provinsi Diy Menggunakan Fuzzy Clustering Dan K-Means Clustering.” Xvi: 229–42.

Tapia, Freddy, And Cristina Aguinaga. 2018. “Detección De Patrones De Comportamiento A Través De Redes Sociales Como Twitter , Utilizando Técnicas De Minería De Datos Como Método Para Detectar El Acoso Cibernético Detection Of Behavior Patterns Through Social Networks Like Twitter , Using Data Minin.” *2018 7th International Conference On Software Process Improvement (Cimps)*: 111–18.

Tapia, Freddy, Cristina Aguinaga, And Roger Luje. 2019. “Detection Of Behavior Patterns Through Social Networks Like Twitter, Using Data Mining Techniques As A Method To Detect Cyberbullying.” *Applications In Software Engineering - Proceedings Of The 7th International Conference On Software Process Improvement, Cimps 2018*: 111–18.

Usama Fayyad, Gregory Piatetsky-Shapiro, And Padhraic Smyth. 1996. “From Data Mining To Databases Usama.” *Lecture Notes In Computer Science (Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics)* 17: 3.

Widiandana, Panggah, And Imam Riadi. 2019. “Analisis Investigasi Forensik Cyberbullying Pada Whatsapp Messenger Menggunakan Metode National Institute Of Standards And Technology (Nist).” : 488–93.

Yoannes Romando, Reny Sulistyowati, And Iwan Setiawan Wibisono. 2019. “Identifikasi Komentar Negatif Berbahasa Indonesia Pada Instagram Dengan K-Means.” *Multimatrix* Ii(1): 6–8.

You, Wanhai, Yawei Guo, And Cheng Peng. 2017. “Twitter’s Daily Happiness Sentiment And The Predictability Of Stock Returns.” *Finance Research Letters* 23: 58–64. Http://Dx.Doi.Org/10.1016/J.Frl.2017.07.018.

Yugianus, Pausta, Harry Soekotjo Dachlan, And Rini Nur Hasanah. 2013. “Pengembangan Sistem Penelusuran Katalog Perpustakaan Dengan Metode Rocchio Relevance Feedback.” *Jurnal Eeccis* 7(1): 47–52.

Zhang, Lungan, Liangxiao Jiang, Chaoqun Li, And Ganggang Kong. 2016. “Two Feature Weighting Approaches For Naive Bayes Text Classifiers.” *Knowledge-Based Systems* 100: 137–44. Http://Dx.Doi.Org/10.1016/J.Knosys.2016.02.017.

Zhao, Rui, And Kezhi Mao. 2017. “Cyberbullying Detection Based On Semantic-Enhanced Marginalized Denoising Auto-Encoder.” *Ieee Transactions On Affective Computing* 8(3): 328–39.