PENDETEKSIAN BAHASA KASAR (*ABUSIVE LANGUAGE*) DAN UJARAN KEBENCIAN (*HATE SPEECH*) DARI KOMENTAR DI JEJARING SOSIAL

Oleh : Luh Putu Ary Sri Tjahyanti[1]

Abstrak

Bahasa kasar merupakan ekspresi yang berisi kata-kata kasar atau frase kasar atau kotor baik dalam konteks lelucon, pelecehan seks vulgar atau mengutuk seseorang. Namun bahasa kasar sering mengarah ke ujaran kebencian yang penyebarannya dilarang di ruang publik seperti jejaring sosial. Jejaring sosial yang digunakan dalam penelitian ini adalah Twitterkarena data tweets-nya dapat diambil melalui Twitter API dan Tweepy Library. Ujaran kebencian dapat ditentukan berdasarkan tingkatan, target, dan kategori. Artikel ini membahas klasifikasi teks multi-label menggunakan algoritma klasifikasi seperti*Naïve Bayes* (NB), *Support Vector Machine* (SVM), dan *Random Forest Decision Tree* (RFDT). Ekstraksi fitur yang digunakan untuk klasifikasi adalah fitur frekuensi istilah seperti kata *n-gram* dan huruf*n-gram*. Penelitian ini melakukan dua skenario dengan jenis label yang berbeda untuk menemukan akurasi tertinggi yang mungkin dapat dicapai oleh pengklasifikasi. Pada skenario pertama, NB dengan fitur kata *unigram + bigrams* memberikan hasil terbaik dengan 71,15% dari F_1– *Score*.Sedangkan untuk skenario kedua, terlihat bahwa NB dengan kata *unigram* memberikan hasil terbaik dengan 87,26% dari F_1– *Score*.

Kata Kunci: bahasa kasar, ujaran kebencian, Naïve Bayes (NB), Support Vector Machine, Random Forest Decision Tree.

Abstract

Abusive language is an expression that contains rude words or abusive or dirty phrases in the context of jokes, vulgar sex or condemning someone. However, abusive language often leads to hate speech which are forbidden in public spaces such as social networks. The social network used in this study is Twitter because the tweets data can be retrieved via Twitter API and Tweepy Library. Hate speech can be determined based on levels, targets and categories. This article discusses the classification of multi-label text using classification algorithms such as Naïve Bayes (NB), Support Vector Machine (SVM), and Random Forest Decision Tree (RFDT). Feature extraction used for classification is the frequency feature of terms such as the word n-gram and the char n-gram. This study conducted two scenarios with different label types to find the highest accuracy that might be achieved by the classifier. In the first scenario, NB with the word unigram + bigrams features gives the best results with 71.15% of F_1 - *Score*. As for the second scenario, it appears that NB with the word unigram gives the best results with 87.26% of F_1 - *Score*.

Keywords: abusive language, hate speech, Naïve Bayes, Support Vector Machine, Random Forest Decision Tree.

PENDAHULUAN

[1] *Luh Putu Ary Sri Tjahyanti adalah staf edukatif pada FKIP Universitas Panji Sakti Singaraja*

Setiap orang yang terlibat secara online, baik di forum kolom pesan, komentar, atau jejaring sosial, selalu ada risiko serius bahwa dirinya mungkin menjadi sasaran ejekan dan bahkan pelecehan dalam bahasa / kalimat kasar. Kata-kata dan kalimat kasar yang diungkapkan dapat berdampak besar pada kesopanan sebuah komunitas atau pengalaman pengguna. Untuk memerangi bahasa yang kasar, banyak perusahaan internet memiliki standar dan pedoman yang harus dipatuhi pengguna, bersama dengan sistem yang menggunakan ekspresi reguler dan daftar hitam bahasa kasar, untuk menangkap bahasa yang buruk dan yang kemudian tidak dapat di-posting atau terhapus secara otomatis. Seiring semakin banyak orang berkomunikasi secara online, kebutuhan akan pengklasifikasi bahasa kasar otomatis yang berkualitas tinggi menjadi sangat diperlukan.

Indonesia adalah salah satu negara dengan pengguna jejaring sosial terbanyak (Statiska, 2020). Orang Indonesia sering menggunakan jejaring sosial untuk berbagai keperluan, seperti mencari dan berbagi informasi, menjalin komunikasi, media iklan, atau untuk sekedar melepaskan perasaan hati (curhat). Sejumlah besar pengguna jejaring sosial sering mengarah pada komunikasi yang tidak terkendali dan banyak netizen yang berkomunikasi dengan bahasa yang kasar. Bahasa kasar adalah ekspresi yang mengandung kata-kata atau frase yang kasar / kotor, baik lisan maupun teks. Menurut, penyebab tidak terkontrolnya penggunaan kata-kata kasar di jejaring sosial adalah tidak adanya alat yang efektif untuk menyaring bahasa kasar di jejaring sosial, kurangnya empati di antara warga, dan kurangnya bimbingan orang tua. Bahasa kasar di jejaring sosial perlu disaring sehingga tidak ada anak-anak dan remaja yang belajar bahasa kasar dari jejaring sosial yang mereka gunakan. Namun, hampir tidak mungkin untuk memfilter bahasa kasar di jejaring sosial secara manual karena sejumlah besar orang yang menulis bahasa kasar.

Mendeteksi bahasa yang kasar di jejaring sosial adalah masalah yang sulit dipecahkan. Salah satu karya pertama yang membahas bahasa kasar adalah Yin dkk. (2009) yang menggunakan teknik klasifikasi terawasi dalam hubungannya dengan *n-gram*, pola ekspresi reguler yang dikembangkan secara manual, fitur kontekstual yang memperhitungkan pelecehan kalimat sebelumnya. Karena sebagian besar pendekatan dasar menggunakan daftar hitam yang telah ditentukan, Sood dkk. (2012) mencatat bahwa beberapa kata daftar hitam mungkin tidak kasar dalam konteks yang

tepat. Dalam pekerjaan mereka, mereka menunjukkan peningkatan dalam deteksi kata-kata kotor dengan memanfaatkan daftar serta metrik jarak edit. Yang terakhir memungkinkan mereka untuk menangkap istilah yang tidak dinormalisasi seperti *@ss* atau *sh1t*. Kontribusi lain dari pekerjaan ini adalah mereka yang pertama menggunakan *crowdsourcing* untuk menjelaskan bahasa yang kasar. Dalam tugas mereka, mereka menggunakan pekerja Amazon Mechanical Turk untuk memberi label 6.500 komentar internet sebagai kasar atau tidak kasar.

Nobata dkk.(2016) mengatakan bahwa mendeteksi bahasa yang kasar di jejaring sosial tidak bisa hanya menggunakan pencocokan kata. Selain itu, ejaan dan tata bahasa dari netizen ketika berbicara bahasa kasar di jejaring sosial sangat informal. Terutama dalam data teks pendek, mengklasifikasikan data teks pendek untuk mendeteksi bahasa yang kasar lebih sulit untuk diselesaikan. Misalnya dalam data Twitter, ada banyak netizen memposting tweet menggunakan singkatan karena kata membatasi tweet. Hanafiah dkk. (2017) mengatakan bahwa beberapa kata non-formal yang sering digunakan oleh orang Indonesia adalah kata-kata yang menunjukkan perasaan, pengulangan huruf untuk menekankan makna, menggunakan kata-kata slang, dan mengubah vokal ke angka.

Penelitian tentang deteksi bahasa yang kasar di jejaring sosial telah dilakukan dalam beberapa tahun terakhir dengan berbagai pendekatan. Turaob dkk. (2017) mendiskusikan deteksi bahasa yang kasar dalam bahasa Thailand. Dataset yang mereka gunakan adalah data pos dan komentar Facebook. Mereka menggunakan beberapa pengklasifikasi (*classifier*) yaitu NB, *k-Nearest Neighbor* (kNN), SVM, RFDT, dan lain-lain. Dengan beberapa fitur pembobotan istilah seperti kata *n-gram* (kata *unigram* dan kata *bigrams*) dan Frekuensi Istilah - Frekuensi Dokumen Balik atau *Term Frequency – Inverse Document Frequency* (TFIDF).

Chen (2012) menjelaskan pentingnya penelitian tentang deteksi bahasa yang kasar di jejaring sosial. Menurut Chen, bahasa kasar di jejaring sosialperlu diklasifikasikan sehingga dapat membuat jejaring sosial yang lebih baik untuk anak-anak dan remaja. Penelitian Chen dilakukan menggunakan bagian kolom komentar Bahasa Inggris Youtube sebagai dataset dengan NB dan SVM sebagai classifier. Fitur yang digunakan dalam penelitian mereka adalah kata *n-gram*, pendekatan penilaian, dan fitur sintaksis leksikal.

Sebagian besar penelitian sebelumnya di bidang deteksi bahasa kasar sebenarnya telah tersebar di beberapa bidang yang tumpang tindih. Ini dapat menyebabkan beberapa kebingungan karena karya yang berbeda dapat menangani aspek spesifik dari bahasa yang kasar, mendefinisikan istilah secara berbeda, atau menerapkannya pada domain online tertentu saja (Twitter, forum online, dan lain-lain.). Untuk semakin mempersulit perbandingan antara pendekatan, hampir semua penelitian sebelumnya menggunakan set evaluasi yang berbeda. Salah satu kontribusi dari artikel ini adalah menyediakan dataset publik untuk memajukan bidang ini dengan lebih baik.

Selain memilih pengklasifikasi dan fitur yang akan digunakan dalam proses klasifikasi, ada hal penting lainnya yang harus dilakukan sebelum proses klasifikasi, yaitu menyeimbangkan nomor dataset ke masing-masing kelas label data. Ganganwar (2012) mengatakan bahwa dataset yang tidak seimbang dapat memberikan hasil negatif untuk klasifikasi. Ini karena jumlah dataset yang tidak seimbang antara kelas mayor dan minor cenderung membuat kelas mayor memiliki kinerja yang lebih baik daripada yang minor. Masalah ini dapat diselesaikan dengan teknik resampling data. Pengambilan sampel data adalah teknik untuk menyeimbangkan dataset dengan menghapus beberapa data kelas utama atau menduplikasi beberapa data kelas minor sehingga jumlah dataset pada setiap kelas menjadi lebih seimbang.

Dalam penelitian ini, dataset Twitter baru dibangun untuk deteksi bahasa Indonesia yang kasar. Secara umum, kontribusi penelitian ini adalah menganalisis jenis-jenis bahasa kasar Indonesia, membangun dataset Twitter baru untuk penelitian dalam deteksi bahasa Indonesia yang kasar, dan menganalisisnyadengan beberapa pengklasifikasi dengan fitur kata *n-gram* dan *huruf-gram*.

1. Bahasa Kasar (*Abusive Language*) dalam Bahasa Indonesia

Dalam bahasa Indonesia, kata-kata kasar / kotor biasanya berasal dari suatu kondisi, hewan, makhluk astral, benda, bagian tubuh, anggota keluarga, aktivitas, dan profesi (Triadi, R., 2017). Di bawah ini adalah penjelasan lebih lanjut tentang jenis referensi kata kasar / kotor dalam bahasa Indonesia(Ibrohim dan Indra Budi, 2018).

- **Kondisi**.Kata-kata yang mengungkapkan kondisi yang tidak menyenangkan dalam percakapan biasanya digunakan sebagai kata-kata kasar. Secara umum,

ada tiga hal yang dapat atau mungkin berhubungan dengan kondisi tidak menyenangkan ini, yaitu gangguan mental (misalnya: *gila*, *bego*, *goblok*, *idiot*, *sinting*, *bodoh*, *tolol*, *sontoloyo*, *geblek*, *sarap*), penyimpangan seksual (misalnya: *lesbi*, *homo*, *banci*, *waria*), kurangnya modernisasi (misalnya: *kampungan*, *udik*, *alay*), cacat fisik (misalnya: *buta*, *budek*, *bolot*, *bisu*), kondisi di mana seseorang tidak memiliki etika (misalnya: *brengsek*, *bejat*, *bajingan*) kondisi yang tidak disetujui oleh Tuhan atau agama (misalnya: *keparat*, *jahanam*, *terkutuk*, *kafir*, *najis*), dan kondisi yang terkait dengan keadaan yang tidak menguntungkan (misalnya: *celaka*, *mati*, *modar*, *sialan*, *pantek*, *mampus*).

- **Hewan**.Tidak semua hewan dapat digunakan sebagai kata-kata kasar. Hewan yang digunakan sebagai kata-kata ofensif biasanya merujuk pada karakteristik buruk tertentu, yang menjijikkan bagi beberapa orang (misalnya: *anjing*, *kampret*, *cebong*, *kodok*), menjijikkan dan dilarang dalam agama tertentu (misalnya: *babi*), menjengkelkan (misalnya: *bangsat*, *kucing*, *kunyuk*), parasit (mis: *lintah*), sehat (mis: *buaya*, *bandot*), dan berisik (mis: *beo*).
- **Makhluk astral**.Contoh makhluk astral yang biasanya digunakan sebagai kata-kata kasar adalah *setan*, *setan alas*, *iblis*, *tuyul*, dan *kunti*. Mereka semua adalah makhluk astral yang sering mengganggu kehidupan manusia.
- **Sebuah Objek**.Sama seperti binatang dan makhluk astral, benda-benda yang biasanya digunakan sebagai kata-kata kasar didasarkan pada karakteristik buruk mereka, seperti bau busuk (misalnya: *tai*, *tai kucing*, *bangkai*), kotor dan usang (mis: *gembel*, *gombal*), dan suara yang mengganggu (misalnya: *sompret*).
- **Bagian dari tubuh**. Bagian tubuh yang digunakan sebagai kata-kata kasar biasanya berkaitan erat dengan aktivitas seks seperti *kontol*, *memek*, *tempik*, dan *jembut*. Bagian tubuh lain yang sering digunakan dalam kutukan adalah mata (mata dalam bahasa Indonesia) dalam bentuk *matamu* yang berarti salah satu mengutuk yang lain karena tidak menggunakan mata mereka dengan benar dan membuat kesalahan karenanya. Ungkapan lain adalah *mata belang* dan *mata duitan* yang digunakan secara kiasan untuk mengutuk seorang pria cabul dan orang yang memilih uang atas segalanya, masing-masing.

- **Anggota keluarga**. Orang Indonesia biasanya menambahkan akhiran *-mu* pada katan yang mengacu ke hubungan sebagai kutukan, seperti *ibumu*, *bapakmu*, *kakekmu*, dan *nenekmu*.
- **Aktivitas**. Kata-kata kasar pada kegiatan biasanya lebih mengarah ke seksual, seperti *ngentot*, *kentu*, dan *ngewe*.
- **Profesi**. Pekerjaan seseorang, terutama pekerjaan kelas rendah yang dilarang oleh agama, sering digunakan oleh orang Indonesia sebagai kata-kata kasar. Pekerjaan-pekerjaan itu termasuk *maling*, *sundel*, *copet*, *lonte*, *cecenguk*, *kacung*, *pelacur*, *pecun*, *jablay*, dan *perek*.

Mendeteksi bahasa kasar di jejaring sosial Indonesia menjadi lebih sulit karena banyak netizen di Indonesia menggunakan kata-kata kasar dalam bahasa asing dalam percakapan mereka dalam konteks lelucon (bahasa kasar tapi bukan ujaran kebencian) atau mengandung unsur SARA dan makian (ujaran kebencian).Contoh kata-kata kasar dibahasa asing yang sering digunakan oleh netizen Indonesia adalah *fuck*, *shit*, *bitch*, *motherfucker* (Inggris) dan *cyka blyat*(Bahasa Rusia). Penggunaan kata-kata kasar dalam bahasa asing tidak hanya dalam bentuk formal, tetapi juga informal, contohnya '*Fak yu!*' yang bentuk formalnya '*Fuck you!*'. Di bawah ini adalah penjelasan tentang pola penulisan kata kasar di jejaring sosial Indonesia.

- **Menggunakan bentukinformal dari bahasa kasar**.

  Sebuah kata kasar biasanya dibuat dari sebuah kata kasar yang diperpanjang. Bentuk informal dibuat dengan membuat kosakata baru yang pengucapannya mirip dengan kata kasar yang asli. Sebagai contoh, banyak netizen mengetik *meki* untuk mengatakan *memek* dan mengetik *kintil* untuk mengucapkan *kontol*.

- **Menggunakan bahasa asing dan lokal**.

  Seperti yang dijelaskan sebelumnya, banyak netizen Indonesia yang mengucapkan kata-kata kasar menggunakan bahasa asing (misalnya: *fuck*, *shit*, *bitch*, *motherfucker*, *cyka blyat*, dan lain-lain). Atau bahasa lokal (misalnya:*asu*, *kimak*, *kampang*, *jancuk*, dan lain-lain). Tidak hanya dalam bentuk formal, mereka juga biasanya mengetik dalam bentuk informal. Beberapa netizen biasanya mengetik bahasa kasar dalam satu bahasa, baik bahasa Indonesia murni, bahasa asing murni, atau bahasa lokal murni. Namun,

ada juga netizen yang mengetik bahasa kasar dengan bahasa campuran.

- **Menghapus bagian vokalnya**.

  Di banyak jejaring sosial terutama di Twitter yang membatasi jumlah huruf dalam sebuah postingan, sering ditemukan netizen yang mengetik kata kasar dengan menghapus vokal. Sebagai contoh, mereka mengetik*bgst* untuk mengucapkan *bangsat* dan mengetik*anjg* untuk mengatakan *anjing*.

- **Pengulangan huruf**.

  Dalam keadaan yang sangat marah, netizen terkadang mengetik kata-kata kasar dengan mengulangi beberapa huruf untuk menunjukkan kemarahan mereka. Contoh untuk pola ini adalah *baaaanggsaaaattt* (*bangsat*), *taaiiiii* (*tai*), *annjiiiiingg* (*anjing*), dan lain-lain.

- **Pergantian huruf**.

  Banyak netizen Indonesia mengganti beberapa huruf dalam kata-kata kasar saat menggunakan bahasa kasar karena beberapa alasan. Untuk menunjukkan kemarahan mereka, beberapa netizen misalnya mengganti *t* dengan *d*, misalnya: *bangsad* (*bangsat*), *jembud* (*jembut*), *bejad* (*bejat*), *ngentod* (*ngentot*), dan lain-lain. Dalam konteks lelucon, netizen biasanya mengubah *s* dengan *c*, misalnya : *bangcat* (*bangsat*), *acu* (*asu*), dan lain-lain. Beberapa pola lainnya berubah *k* dengan *q* (misalnya: *qontol* (*kontol*), *qimak* (*kimak*)), mengubah *j* dengan *dj* (misalnya: *djembut* (*jembut*), *djancuk* (*jancuk*)), mengganti *u* dengan *oe* (mis: *jemboet* (*jembut*), *djancoek* (*jancuk*)), dan mengganti vokal dengan angka (mis: *b4ngs4t* (*bangsat*), *b3g0* (*bego*), *j3mbut* (*jembut*)). Selain itu, kita sering menemukan netizen mengubah huruf dalam kata-kata kasar dengan huruf tertentu untuk berusaha mensensor kata-kata kasar, misalnya: *mem*k* (*memek*), *g*blok* (*goblok*), *b#ngsat* (*bangsat*), dan lain-lain.

2. Ujaran Kebencian (*Hate Speech*) dalam Bahasa Indonesia

Berdasarkan informasi dari *Focus Group Discussion* (FGD) dengan staf Direktorat Tindak Pidana Siber Badan Reserse Kriminal Keanggotaan Negara Republik Indonesia (Bareskrim Polri) yang dijelaskan dalam (Hernanto dan Jeihan, 2018), diketahui bahwa bicara kebencian atau *hate speech* memiliki target, kategori,

dan level tertentu. Setiap ujaran kebencian ditujukan pada target tertentu. Secara umum, sasaran ujaran kebencian dibagi menjadi dua macam, yaitu individu dan kelompok. Perkataan yang mengandung kebencian dengan target individu adalah ujaran kebencian yang ditujukan pada seseorang (seorang individu), sedangkan ujaran kebenciandengan target kelompok adalah ucapan benci yang ditujukan pada kelompok, asosiasi, atau komunitas tertentu. Kelompok, asosiasi, dan komunitas ini bisa dalam bentuk kelompok agama, ras, politik, klub penggemar, komunitas hobi, dan lain-lain. Ada beberapa kategori dalam ujaran kebencian:

1. **Agama / kepercayaan**, yang merupakan ujaran kebencian berdasarkan pada agama (Islam, Kristen, Katolik, dan lain-lain), organisasi / aliran keagamaan, atau kepercayaan tertentu;
2. **Ras / etnis**, yang merupakan ujaran kebencian berdasarkan ras manusia (kelompok manusia berdasarkan karakteristik fisik seperti bentuk wajah, tinggi, warna kulit, dan lain-lain) atau etnis (kelompok manusia berdasarkan kewarganegaraan umum atau tradisi budaya bersama di area geografis);
3. **Fisik / kecacatan**, yaitu kebencian yang didasarkan pada kekurangan / perbedaan fisik (misalnya: bentuk wajah, mata, dan bagian tubuh lainnya) atau kecacatan (misalnya: autisme, idiot, buta, tuli, dan lain-lain), baik hanya mengutuk seseorang (atau kelompok) dengan kata-kata yang berhubungan dengan fisik / kecacatan atau yang benar-benar dialami oleh mereka yang menjadi target dari ujaran kebencian;
4. **Orientasi gender / seksual**, yang merupakan ujaran kebencian berdasarkan jenis kelamin (pria dan wanita), mengutuk seseorang (atau kelompok) menggunakan kata-kata yang merendahkan gender (misalnya: gigolo, perempuan jalang, dan lain-lain.), atau orientasi seksual yang menyimpang (misalnya: homoseksual, lesbian, dan lain-lain);
5. **Makian / fitnah lainnya**, yaitu ujaran kebencian dalam bentuk sumpah / ejekan menggunakan kata / frasa kasar atau fitnah / penghasutan lainnya yang tidak terkait dengan empat kelompok di atas.

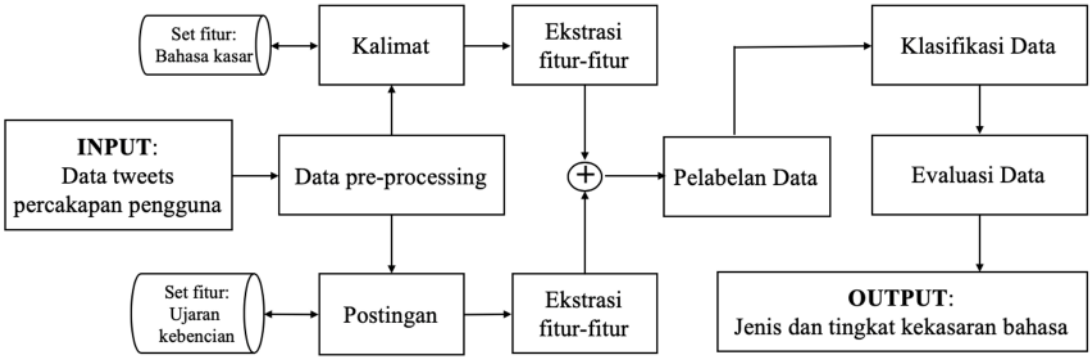
Dari kelima kategori tersebut, kategori makian / fitnah sudah menjurus pada tingkat ujaran kebencian yang kuat, dan tidak dapat dikategorikan di bawah kategori lainnya, dan begitu juga sebaliknya.

Selain memiliki target dan kategori, ujaran kebencian juga memiliki tingkatan tertentu. Di dalam (Hernanto dan Jeihan, 2018) juga dijelaskan tingkatan dari ujaran kebencian dapat dibagi menjadi tiga tingkatan, yaitu lemah, sedang, dan kuat. Penjelasan untuk setiap tingkat ujaran kebencian adalah sebagai berikut:

1. **Ujaran kebencian yang lemah**, yaitu kebencian dalam bentuk sumpah / fitnah yang ditujukan pada individu tanpa menyertakan hasutan / provokasi untuk membawa konflik terbuka. Di Indonesia, ujaran kebencian dalam bentuk ini dikategorikan sebagai ujaran kebencian yang lemah karena itu adalah masalah pribadi. Ini berarti, jika target ujaran kebencian tidak melapor ke pihak yang berwenang (merasa orang biasa dan memaafkan yang menyebarkan ujaran kebencian kepadanya) maka ujaran kebencian itu tidak terlalu diprioritaskan untuk diselesaikan oleh pihak berwenang.
2. **Ujaran kebencian moderat**, yaitu wacana kebencian dalam bentuk sumpah / penistaan / stereotip / pelabelan ditujukan pada kelompok tanpa menyertakan keterlibatan / provokasi untuk menghadirkan konflik terbuka. Meskipun dapat mengundang konflik antar kelompok, jenis ujaran kebencian ini termasuk dalam moderat karena konflik yang akan terjadi diperkirakan terbatas pada konflik di jejaring sosial.
3. **Ujaran kebencian yang kuat**, yaitu ucapan kebencian dalam bentuk sumpah / fitnah / penistaan / stereotip / pelabelan yang ditujukan pada individu atau kelompok termasuk hasutan / provokasi untuk membawa konflik terbuka. Ujaran kebencian semacam ini termasuk dalam ujaran kebencian yang kuat, karena itu ujaran kebencian yang perlu diprioritaskan untuk segera diselesaikan karena dapat mengundang konflik yang tersebar luas dan dapat menyebabkan konflik / kehancuran fisik di dunia nyata.

METODE PENELITIAN

Selanjutnya disini disajikan dataset pendeteksian bahasa kasar dan ujaran kebencian yang digunakan dalam penelitian ini. Dalam penelitian ini, dataset Twitter baru dibangun untuk deteksi bahasa yang kasar dalam tweet Indonesia. Secara umum, diagram alur penelitian untuk percobaan ini dalam artikel ini dapat dilihat pada Gambar 1.



Gambar 1. Diagram alur penelitian

Langkah pertama dari penelitian ini adalah pengambilan input data tweets di Twitter dengan pemrosesan awal (*pre-preocessing*) menggunakan Twitter API dan Tweepy Library. Dalam percobaan pemrosesan awal ini untuk deteksi bahasa Indonesia yang kasar, dilakukan pemeriksaan kalimat tweet menggunakan database set fitur kata-kata kasar, dan juga dilakukan pemeriksaan unsur ujaran kebencian terhadap kata-kata dalam postingan menggunakan database set fitur ujaran kebencian. Ini karena bentuk-bentuk pelecehan informal yang sering ditemukan di jejaring sosial Indonesia. Disini dikumpulkan sekitar 50.000 data tweet. Dari data yang dikumpulkan, datadifilter data dengan menghapus tweet yang digandakan dan tweet yang menggunakan bahasa asing atau lokal.

Proses penyaringan data memberikan 3.000 data yang siap diberi label. Label disini dibedakan dalam 3 label data yaitu *bahasa kasar*, *kasar tapi bukan ujaran kebencian*, dan *ujaran kebencian*. Dataset yang dihasilkan dari data ekstraksi fitur-fitur dilabelkan dan diverifikasi oleh 10 orang relawan penanda untuk memastikan pelabelan data sesuai harapan. Penanda dipastikan dapat membedakan 3 jenis data tersebut dengan tepat (disetujui 100%), dan apabila ada data yang di luar itu maka akan dibuang. Dari proses pelabelan didapatkan 2.735 data.

Fitur-fitur yang sudah diekstraksi dibedakan dalam bentuk fitur kata *n-gram* kata, dan huruf *n-gram*. Fitur-fitur ini disiapkan untuk klasifikasi data. Ada beberapa pengklasifikasi seperti *Naive Bayes* (NB) (Lewis D.D., 1998), *Support Vector Machine* (SVM) (Srivastava dan Bhambhu, 2010) dan *Random Forest Decision Tree* (RFDT) (Ali dkk., 2012). Proses klasifikasi ini dilakukan dalam bahasa Python di Linux dengan menggunakan library yang khusus untuk klasifikasi yaitu Scikit-Learn.

Untuk memastikan hasil klasifikasi bagus atau tidak dan untuk mendapatkan kombinasi fitur yang terbaik digunakan teknik 10-fold cross-validation. Teknik ini dapat membagi dataset menjadi 10 bagian dimana salah satu diantaranya menjadi data testing dan 9 sisanya menjadi data training. Proses testing dan training ini dilakukan 10 kali dan setiap data menjadi data training dan testing secara simultan. Untuk menghindari hasil yang overfitting dimana data untuk training itu "sangat baik" sehingga jika dilakukan testing pada data yang berbeda akan mengurangi keakuratannya, maka digunakan library SMOTE (Chawla dkk., 2002) untuk data yang tidak seimbang.

Proses berikutnya yaitu evaluasi data untuk mengukur tingkat kekasaran bahasa menggunakan pengukuran evaluasi F_1 – *Score* (yang disebut juga *F – Measure*). Output dari evaluasi ini dijabarkan dalam metrik evaluasi untuk jenis metode NB, SVM, dan RFDT.

HASIL PENELITIAN

Dalam artikel ini, peneliti melakukan dua skenario percobaan. Dalam skenario pertama, peneliti mengklasifikasikan tweet menjadi tiga label yang *bahasa kasar*, *bahasa kasar tapi bukan ujaran kebencian*, dan *ujaran kebencian.* Untuk skenario kedua, peneliti hanya mengelompokkan dataset menjadi dua label yang *bukanbahasa kasar* dan *bahasa kasar* untuk mengetahui apakah hasil klasifikasi akan lebih baik jika hanya mengklasifikasikan dua label. Dalam skenario kedua ini, tweet yang berlabel *bahasa kasar tapi bukan ujaran kebencian*dan *ujaran kebencian*akan dilabeli sebagai *bahasa kasar*.

Untuk kedua skenario, peneliti menggunakan fitur kata*n-gram* dan huruf *n-gram* dengan NB, SVM dan RFDT sebagai classifier. Kata *n-gram* yang peneliti gunakan adalah kata *unigram*, kata *bigrams*, kata *trigram*, dan kombinasi semuanya, sedangkan huruf*n-gram* yang peneliti gunakan adalah huruf*trigram*, huruf*quadgram*, dan juga kombinasi huruf*trigram* dan huruf*quadgram*. Tabel 1 dan Tabel 2 menunjukkan F_1 - *Score* untuk setiap skenario dalam%.

Tabel 1. F_1 – *Score* untuk tiga label class

Jenis ekstraksi fitur	NB	SVM	RFDT
kata unigram	70,14	68,67	68,34
kata bigrams	52,28	54,71	28,69
kata trigrams	29,46	24,45	48,17
kata unigram + bigrams	71,15	67,14	62,28
kata unigram + bigrams + trigrams	70,75	68,33	62,61
huruf trigrams	67,91	65,64	60,72
huruf quadgrams	70,24	67,39	63,35
huruf trigrams + quadgrams	70,34	65,28	62,40

Tabel 2. F_1 – *Score* untuk dualabel class

Jenis ekstraksi fitur	NB	SVM	RFDT
kata unigram	87,26	82,53	82,83
kata bigrams	56,14	80,69	41,34
kata trigrams	23,46	78,36	74,35
kata unigram + bigrams	86,38	82,62	82,71
kata unigram + bigrams + trigrams	82,38	82,31	80,56
huruf trigrams	84,87	78,28	81,33
huruf quadgrams	83,49	80,35	81,67
huruf trigrams + quadgrams	87,03	79,18	80,79

Berdasarkan Tabel 1, kita dapat melihat bahwa untuk skenario pertama NB dengan fitur kata *unigram* + *bigrams* memberikan hasil terbaik dengan 71,15% dari F_1 - *Score*, diikuti oleh NB dengan fitur kata *unigram* + *bigrams* + *trigram* (70,75%) dan NB dengan fitur huruf *quadgrams* (70,24%). Sedangkan dari Tabel 2 untuk skenario kedua, kita dapat melihat bahwa NB dengan kata *unigram* memberikan hasil terbaik dengan 87,26% dari F_1 - *Score*, diikuti oleh NB dengan fitur huruf *trigram* + *quadgram* (87,03%) dan NB dengan kata *unigram* + *bigrams* (86,38 %). Dari kedua skenario, dapat dilihat bahwa NB lebih baik daripada SVM dan RFDT untuk mengklasifikasikan tweet dalam percobaan ini. Begitu pula kata *unigram* dan kombinasi kata *n-gram* memberikan hasil yang lebih baik daripada fitur lain untuk setiap classifier yang peneliti gunakan. Oleh karena itu, untuk penelitian berikutnya pada bidang deteksi bahasa kasar di jejaring sosial Indonesia, peneliti menyarankan agar menggunakan NB dengan kata *unigram* dan kombinasi kata *n-gram* untuk pengklasifikasi informasi dasar dan ekstraksi fitur.

Dari kedua skenario, juga dapat dilihat bahwa mengklasifikasikan tweet menjadi tiga label lebih sulit daripada hanya mengklasifikasikan apakah tweet itu

bukan bahasa kasar atau *bahasa kasar*. Di sini, semua pengklasifikasi dengan semua fitur yang peneliti gunakan sulit untuk membedakan apakah tweet itu *bahasa kasar tapibukan ujaran kebencian* atau *ujaran kebencian*. Dari analisis dataset peneliti, ada pola tertentu yang membedakan apakah tweet tersebut merupakan *bahasa kasar tapi bukan ujaran kebencian* atau *ujaran kebencian*. Banyak netizen biasanya mengetik *ujaran kebencian* menggunakan huruf besar atau tanda seru (!), sedangkansaat mereka mengatakan *bahasa kasar tapi bukan ujaran kebencian*biasanya menambahkan kata yang berarti menertawakan sesuatu seperti:*wkwk*, *haha*, *hihi*, *hehe*, dan lain-lain.

SIMPULAN

Dalam artikel ini, peneliti membahas pendeteksian bahasa kasar dan ujaran kebencian dalam bahasa Indonesia di jejaring sosial yang umumnya berasal dari kondisi yang tidak menyenangkan dan dilarang secara etika dan agama. Di sini, peneliti membuat dataset baru untuk pendeteksian bahasa kasar dan ujaran kebencian.Hasil percobaan menunjukkan bahwa NB lebih baik daripada SVM dan RFDT untuk mengklasifikasikan bahasa kasar menggunakan dataset peneliti di semua skenario. Untuk ekstraksi fitur, kata *unigram* dan kombinasi kata *n-gram* memberikan hasil yang lebih baik daripada fitur lainnya, baik menggunakan NB, SVM atau RFDT. Hasil percobaan juga menunjukkan bahwa mengklasifikasikan tweet menjadi tiga label (*bahasa tidak kasar*, *bahasa kasar tapi bukan ujaran kebencian*, dan *ujaran kebencian*) lebih sulit daripada hanya mengklasifikasikan apakah tweet itu *bukan bahasa kasar* atau *bahasa kasar*. Dari hasil penelitian terlihat pengklasifikasi NB dengan kata *unigram* dan kombinasi kata *n-gram*dapat digunakan sebagai informasi dasar. Peneliti menyarankan penggunaan kamus huruf besar, tanda baca, dan tertawa untuk ekstraksi fitur untuk meningkatkan hasil klasifikasi dalam membedakan apakah tweet adalah bahasa yang kasar tetapi bukan ujaran kebencian.

DAFTAR PUSTAKA

Ali, J., Khan, R., Ahmad, N., dan Maqsood, I.. *Random Forests And Decision Trees*. IJCSI International Journal of Computer Science Issues (IJCSI), vol. 9, issues 5, no. 3, hal. 272–278.2012.

Chawla, N.V., Bowyer, K.W., Hall, L.O., dan Kegelmeyer, W.P.. *SMOTE: Synthetic*

Minority Over-Sampling Technique. Journal of ArtificialIntelligence Research, vol. 16, no. 1, hal. 321–357. 2002.

Chen, Y.. *Detecting offensive language in social media to protect adolescent online safety*. Expert Systems with Applications,vol. 36, hal. 71–80. 2012.

Ganganwar,V..*An overview of classification algorithms for imbalanced datasets*.InternationalJournalofEmergingTechnologyandAdvancedEngineering, vol. 2, no. 4, hal. 42–47, 2012.

Hernanto, B.,dan Jeihan. *Personal communication*.2018.

Hossin, M., Sulaiman, M.N.. A Review On Evaluation Metrics For Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process (IJDKP), vol. 5, no. 2, hal. 1-11. 2015.

Ibrohim, M.O., dan Indra Budi. *A Dataset and Preliminaries Study for Abusive Language Detectionin Indonesian Social Media*. Procedia Computer Science, vol. 135, hal. 222–229. 2018.

Lewis, D.D.. *Naive (bayes) at forty: The independence assumption in information retrieval*. European Conference on Machine Learning(EMCL), hal. 4–15. 1998.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., dan Chang, Y.. *Abusive language detection in online user content*. International WorldWide Web Conference Committee (IW3C2),hal. 145–153.2016.

Sood, S. O., Antin, J., dan Churchill, E. F.. *Using crowdsourcing to improve profanity detection*. AAAI Spring Symposium: Wisdom of the Crowd. 2012.

Srivastava, D.K. dan Bhambhu, L.. *Data Classification Using Support Vector Machine*. Journal of Theoretical and Applied Information Technology, vol. 12, no. 1, hal. 1–7. 2010.

Statiska. Number of social network users in selected countries in 2018 and 2023. https://www.statista.com/statistics/278341/number-of-social-network-users-in-selected-countries/. Diakses pada tanggal 1 Mei 2020.

Triadi, R.. *Penggunaan makian bahasa indonesia pada jejaring sosial (kajian sosiolinguisik)*. Jurnal Sasindo Unpam, vol. 5, no. 2, hal. 1–26.2017.

Turaob, S. dan Mitrpanont, J.. *Automatic discovery of abusive thai language*. International Conference on Asia-Pacific Digital Libraries,hal. 267–278.2017.

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., dan Edwards, L.. *Detection of harassment on web 2.0*. Proceedings of the Content Analysis in the WEB, vol. 2, hal. 1–7. 2009.