UNIVERSITAS INDONESIA

DETEKSI UJARAN KEBENCIAN DAN BAHASA KASAR PADA BLOG MIKRO BERBAHASA INDONESIA

SKRIPSI

NABILA KHANSA
1906293221

FAKULTAS ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER
DEPOK
JULI 2023

UNIVERSITAS INDONESIA

DETEKSI UJARAN KEBENCIAN DAN BAHASA KASAR PADA BLOG MIKRO BERBAHASA INDONESIA

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar

Sarjana Ilmu Komputer

NABILA KHANSA

1906293221

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI ILMU KOMPUTER

DEPOK

JULI 2023

HALAMAN PERNYATAAN ORISINALITAS

Skripsi ini adalah hasil karya saya sendiri,
dan semua sumber baik yang dikutip maupun dirujuk
telah saya nyatakan dengan benar.

Nama	:	**Nabila Khansa**
NPM	:	**1906293221**
Tanda Tangan	:	
Tanggal	:	**3 Juli 2023**

HALAMAN PENGESAHAN

Tugas Akhir ini diajukan oleh	:	
Nama	:	Nabila Khansa
NPM	:	1906293221
Program Studi	:	Sarjana Ilmu Komputer
Judul	:	Deteksi Ujaran Kebencian dan Bahasa Kasar pada Blog Mikro Berbahasa Indonesia

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Ilmu Komputer pada Program Studi Sarjana Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Indonesia

DEWAN PENGUJI

Pembimbing 1	:	Alfan Farizki Wicaksono, S.T., M.Sc., Ph.D.	(Nilai telah diberikan melalui SISIDANG pada 24-07-2023, 09:11:36) (Revisi telah disetujui melalui SISIDANG pada 24-07-2023, 09:24:17)
Penguji	:	Dipta Tanaya, S.Kom., M.Kom.	(Nilai telah diberikan melalui SISIDANG pada 04-07-2023, 07:37:30) (Revisi telah disetujui melalui SISIDANG pada 24-07-2023, 19:58:37)
Penguji	:	Evi Yulianti M.Comp.Sc, M.Kom., Ph.D.	(Nilai telah diberikan melalui SISIDANG pada 11-07-2023, 15:17:18) (Revisi telah disetujui melalui SISIDANG pada 24-07-2023, 10:54:46)

Ditetapkan di : Depok, Jawa Barat

Tanggal : 24 Juli 2023

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa karena atas berkat, rahmat, dan nikmat-Nya penulis dapat menyelesaikan penulisan tugas akhir dengan judul ”Deteksi Ujaran Kebencian dan Bahasa Kasar pada Mikroblog Berbahasa Indonesia”. Penulisan tugas akhir ini dilakukan dalam rangka memenuhi persyaratan untuk mendapatkan gelar Sarjana Ilmu Komputer Jurusan Ilmu Komputer Universitas Indonesia. Penulis mengucapkan terima kasih sebesar-besarnya kepada:

1. Pak Alfan Farizki Wicaksono selaku pembimbing yang telah memberi banyak saran serta bantuan dalam pengerjaan tugas akhir ini.
2. Bu Dipta Tanaya dan Bu Evi Yulianti selaku dosen Penguji.
3. Pak Suryana Setiawan selaku pembimbing akademis penulis selama menempuh pendidikan di Fakultas Ilmu Komputer Universitas Indonesia.
4. Milo sebagai sahabat yang selalu ada untuk menemani dan memberi semangat.
5. Orangtua penulis, Christia Nency dan Dwi Asmoro, yang selalu sabar menghadapi tingkah laku penulis.
6. Adik penulis, Pasha, yang selalu membelikan McD ketika penulis sibuk dan lupa makan.
7. Teman-teman kuliah penulis yaitu, Ilma Alpha Mannix, Jeremy Victor, Steven W.H., Fathan Muthahhari, Mardianto, Nathanael Pardosi, Sultan Daffa, dan Amanda C.A.
8. Teman-teman satu bimbingan penulis, Hendrico dan Raniah.
9. Teman dekat penulis, yaitu Yasmin Firasyan, Sad Khalishah, dan Aisyah Rania.
10. Teman-teman SMA penulis, yaitu Rembulan dan Shafira.
11. Arthur, Gabe, Nathan, Gendo, dan Nero sebagai teman bermain yang selalu menghibur.
12. Angel, Maggie, Val, Filbert, Sean, dan Min yang selalu bersedia mendengarkan keluh kesah penulis.
13. Teman-teman proyek Elevania yaitu, Kak Anas dan Reinald yang telah memberi banyak dukungan.
14. Astolfo yang selalu mengisi hati penulis dengan harapan dan kebahagiaan.

15. Pihak-pihak lainnya yang sudah membantu penulis dalam melakukan penelitian, menyusun tugas akhir, dan mendukung penulis selama perkuliahan secara langsung maupun tidak langsung.

Penulis menyadari bahwa laporan Skripsi ini masih jauh dari sempurna. Oleh karena itu, apabila terdapat kesalahan atau kekurangan dalam laporan ini, Penulis memohon agar kritik dan saran bisa disampaikan langsung melalui *e-mail* nabila.khansa@gmail.com.

Depok, 3 Juli 2023

Nabila Khansa

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Indonesia, saya yang bertanda tangan di bawah ini:

Nama	:	Nabila Khansa
NPM	:	1906293221
Program Studi	:	Ilmu Komputer
Fakultas	:	Ilmu Komputer
Jenis Karya	:	Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Indonesia **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty Free Right*)** atas karya ilmiah saya yang berjudul:

Deteksi Ujaran Kebencian dan Bahasa Kasar pada Blog Mikro Berbahasa Indonesia

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Indonesia berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Depok
Pada tanggal : 3 Juli 2023
Yang menyatakan

(Nabila Khansa)

ABSTRAK

Nama	:	Nabila Khansa
Program Studi	:	Ilmu Komputer
Judul	:	Deteksi Ujaran Kebencian dan Bahasa Kasar pada Blog Mikro Berbahasa Indonesia
Pembimbing	:	Alfan Farizki Wicaksono, Ph.D.

Ujaran kebencian dan bahasa kasar mempermudah penyebarah kekerasan di kehidupan nyata, sehingga muncul urgensi adanya pendeteksian secara otomatis. Untuk melanjutkan pekerjaan yang sudah dilakukan oleh Ibrohim dan Budi (2019), penelitian ini membahas dua isu terkait deteksi ujaran kebencian dan bahasa kasar pada mikroblog berbahasa Indonesia. Isu pertama adalah kajian terkait *effect size* fitur dan pengembangan model menggunakan fitur-fitur tersebut. Metode *Analysis of Variance f-test*, *Logistic Regression Analysis*, dan nilai *Shapley* digunakan untuk melakukan kajian *effect size* pada fitur-fitur yang dirancang secara manual. Kemudian, digunakan beberapa algoritma pemelajaran mesin untuk mengembangkan model prediksi berbasis fitur-fitur tersebut. Isu kedua adalah kajian bias dalam pengembangan model terkait keberadaan kata-kata bersifat netral pada data yang merupakan ujaran kebencian atau bahasa kasar. Kajian terkait bias dilakukan dengan menggunakan *dataset* uji bias. *Dataset* ini dikembangkan dengan menggantikan kata-kata yang dideteksi memiliki potensi adanya bias pada model yang dilatih menggunakan *dataset* hasil pekerjaan Ibrohim dan Budi (2019). Penelitian ini menunjukkan bahwa keberadaan kata-kata tertentu berpengaruh terhadap hasil deteksi ujaran kebencian dan bahasa kasar. Di antara kata-kata tersebut, terdeteksi beberapa kata-kata yang berpotensi bias, karena memiliki pengaruh terhadap pendeteksian padahal secara sendiri kata-kata yang dideteksi sebagai potensi bias tidak memiliki unsur kebencian atau bersifat kasar. Hasil evaluasi pengambilan sampel *bootstrap* menunjukkan *Logistic Regression* dan *XGBoost* sebagai model dengan akurasi terbaik dalam pendeteksian ujaran kebencian dan bahasa kasar. Namun, ketika model yang sudah dikembangkan digunakan untuk memprediksi *dataset* sintetis, didapatkan penurunan akurasi dalam pendeteksian ujaran kebencian. Hasil ini menandakan adanya bias pada model yang dikembangkan. Hasil tersebut didukung juga oleh hasil prediksi dengan akurasi rendah ketika model digunakan untuk melakukan pendeteksian ujaran kebencian pada *dataset* yang dikembangkan secara manual, tetapi ketika kata-kata bias digantikan dari data, akurasi model meningkat. Kontribusi yang diberikan oleh penelitian ini adalah pengembangan dataset uji bias secara otomatis dari *dataset* yang dikembangkan oleh Ibrohim dan Budi (2019) dan juga *dataset* uji bias yang dikembangkan secara manual.

Kata kunci:
klasifikasi, *effect size*, pemodelan, bias stereotipe, *dataset* sintetis

ABSTRACT

Name	:	Nabila Khansa
Study Program	:	Computer Science
Title	:	Detection of Hate Speech and Abusive Language on Indonesian Microblogs
Counsellor	:	Alfan Farizki Wicaksono, Ph.D.

Hate speech and abusive language facilitate the spread of violence in real life, hence the urgency of automatic detection. To continue the work done by Ibrohim dan Budi (2019), this research addresses two issues related to the detection of hate speech and abusive language on Indonesian-language microblogs. The first issue is a study on the effect size of features and the development of models using these features. Analysis of Variance f-test, Logistic Regression Analysis, and Shapley values are used to investigate the effect size of manually designed features. Several machine learning algorithms are then employed to develop prediction models based on these features. The second issue involves studying bias in model development concerning the presence of neutral words in data that constitute hate speech or abusive language. The study related to bias is conducted by using a bias test dataset. This dataset is developed by replacing words that are detected to have the potential for bias in models trained using the dataset resulting from the work of Ibrohim dan Budi (2019). This research demonstrates that certain words significantly influence the detection of hate speech and abusive language. Among these words, some are identified as potentially biased, as they affect detection despite not inherently containing hate or abusive elements. The results of bootstrap sampling evaluation indicate that Logistic Regression and XGBoost are the models with the highest accuracy in detecting hate speech and abusive language. However, when the developed models are used to predict synthetic datasets, a significant decrease in accuracy is observed in hate speech detection. This finding indicates the presence of bias in the developed models. This result is further supported by low-accuracy predictions when the models are used to detect hate speech in manually developed datasets. However, when biased words are replaced in the data, the model's accuracy significantly improves. The contributions of this research include the development of an automatically generated bias test dataset from the dataset created by Ibrohim dan Budi (2019), as well as a manually developed bias test dataset.

Key words:
classification, effect size, modelling, stereotype bias, synthetic dataset

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR KODE PROGRAM

DAFTAR LAMPIRAN

BAB 1
PENDAHULUAN

Bab 1 mendiskusikan latar belakang dari permasalahan utama yang diangkat pada laporan tugas akhir ini, yaitu permasalahan deteksi ujaran kebencian dan penggunaan bahasa kasar serta kajian bias stereotipe yang ada pada koleksi data. Pertama, Subbab 1.1 membahas motivasi dan manfaat dari tugas deteksi ujaran kebencian dan bahasa kasar. Ringkasan pekerjaan terkait dan juga posisi penelitian ini juga dibahas di Subbab 1.1. Kemudian, Subbab 1.2 menyampaikan tiga buah rumusan masalah yang menjadi fokus pada tugas akhir ini dan Subbab 1.3 menampilkan daftar tujuan yang ingin dicapai dari tugas akhir ini. Selanjutnya, Subbab 1.4 menyampaikan batasan penelitian dalam hal penggunaan *dataset* dan domain pada koleksi data. Terakhir, Subbab 1.5 mengakhiri Bab 1 dengan ringkasan terkait sistematika penulisan yang digunakan pada laporan tugas akhir ini.

1.1 Latar Belakang

Ujaran kebencian (*hate speech*) adalah bentuk ujaran dengan tujuan merendahkan seseorang atau sekelompok orang yang dapat dilakukan secara langsung maupun tidak dan mengandung kebencian yang didasarkan oleh karakteristik khas dari orang atau kelompok yang menjadi target (HAM, 2015; Nockleby, 2000). Penyebaran ujaran kebencian umumnya didampingi oleh penggunaan bahasa kasar (*abusive language*), terutama di media sosial di mana pengguna mempunyai cukup kebebasan untuk menghasilkan konten secara anonim yang dapat dengan mudah tersebar (Davidson, Warmsley, Macy, & Weber, 2017). Sebagai tambahan, ujaran kebencian dan bahasa kasar sering kali digunakan dalam konteks politik, seperti untuk menyerang anggota parlemen tertentu (Agarwal et al., 2021) dan untuk mengangkat isu terkait etnis atau agama tertentu (Pettersson, 2019). Hal ini tentunya tidak diharapkan karena penggunaan ujaran kebencian dapat memancing terorisme domestik (Piazza, 2020) dan juga berdampak buruk pada emosi, empati, dan kelakuan seseorang yang terpapar (Bilewicz & Soral, 2020).

Bahasa kasar dapat diklasifikasikan berdasarkan (1) target kekasaran dan berdasarkan (2) tingkat kekasaran. Jika ditinjau berdasarkan target, bahasa kasar bisa diucapkan ter-

hadap target spesifik seperti individu, atau digeneralisasikan terhadap kelompok tertentu; dan dilihat dari tingkat kebencian, bahasa kasar bisa bersifat eksplisit, yaitu tidak ambigu bahwa bahasa tersebut kasar, atau bersifat implisit, yaitu tidak menunjukan secara langsung bahwa bahasa tersebut kasar (Waseem, Davidson, Warmsley, & Weber, 2017). Selain itu, bahasa kasar memiliki ketergantungan pada konteks dalam penggunaanny dan dapat tersembunyi dalam bentuk sindiran tidak langsung, sehingga dibutuhkan adanya pemahaman akan pengetahuan budaya yang bersangkutan (Lachenicht, 1980). Penelitian ini menggunakan *dataset* yang berasal dari domain politik. Beberapa terminologi ujaran kebencian yang terkait kontestasi politik tahun 2019, seperti ”cebong” dan ”kampret”, dapat ditemukan pada koleksi data.

Deteksi terhadap konten yang mengandung ujaran kebencian dan bahasa kasar kemudian menjadi sangat penting karena sifat penggunaannya yang memfasilitasi penyebaran kekerasan di dunia nyata (Piazza, 2020) dan dapat menyebabkan trauma pribadi, cyber-bullying, diskriminasi, dan kejahatan yang didasarkan diskriminasi (Lin, 2022; Park, Shin, & Fung, 2018). Pada tahun 2017, pihak kepolisian Indonesia telah menangani 3325 kasus ujaran kebencian, meningkat sebesar 44.99% dari 1829 kasus pada tahun sebelumnya (Medistiara, 2017). Meskipun sudah ada banyak penelitian deteksi ujaran kebencian untuk *dataset* berbahasa Inggris, ujaran kebencian memiliki implikasi budaya yang kuat, seperti apa yang dianggap menyinggung bisa berubah berdasarkan latar budaya (Schmidt & Wiegand, 2017). Sebagai contoh, kebanyakan kata kasar di bahasa Indonesia adalah nama-nama hewan, seperti ”bangsat”[1] dan ”kampret”[2], yang jika diterjemahkan ke bahasa lain belum tentu merupakan kata kasar di bahasa tersebut. Oleh karena itu, model yang dikembangkan untuk bahasa Inggris tidak serta merta dapat digunakan untuk deteksi ujaran kebencian dan bahasa kasar di dokumen berbahasa Indonesia; tetap diperlukan adalah penyesuaian dan penelitian lanjutan dalam domain bahasa Indonesia.

Beberapa pekerjaan sudah dilakukan untuk pengembangan model deteksi ujaran kebencian dan bahasa kasar berbahasa Indonesia, khususnya dalam hal pengembangan *dataset* (Alfina, Mulia, Fanany, & Ekanata, 2017; Ibrohim & Budi, 2018). Alfina et al. (2017) telah melakukan penelitian awal dan pengembangan *dataset* untuk topik deteksi ujaran kebencian. Kemudian, Ibrohim dan Budi (2018) melakukan penelitian untuk topik deteksi bahasa kasar pada media sosial bahasa Indonesia yang membahas tantan-

[1]Berdasarkan KBBI daring, bangsat adalah hewan sejenis kutu busuk atau kepinding.
[2]Berdasarkan KBBI daring, kampret adalah ”kelelawar kecil pemakan serangga”.

gan pendeteksian bahasa kasar dan pola penulisan bahasa kasar dalam bahasa Indonesia serta terkumpulnya *dataset* untuk eksperimen terkait. Pada penelitian yang dilakukan oleh Ibrohim dan Budi (2019), *dataset* dibangun untuk melakukan eksperimen klasifikasi multi-label untuk deteksi ujaran kebencian dan bahasa kasar.

Penelitian yang dilakukan oleh Ibrohim dan Budi (2019) juga membahas pembagian ujaran kebencian menjadi lima kategori berupa: *religion/creed* (agama/kepercayaan), *race/ethnicity* (ras/etnis), *physical/disability* (fisik/disabilitas), *gender/sexual orientation* (jenis kelamin/orientasi seksual), *other invective/slander* (makian lainnya). Selain itu terdapat klasifikasi target ujaran kebencian yang bisa merupakan individual atau group, dan tiga tingkat keparahan yaitu *weak* (lemah), *moderate* (sedang), dan *strong* (kuat). Kemudian, Prabowo, Ibrohim, dan Budi (2019) juga melakukan eksperimen serupa, namun dengan klasifikasi multi-label secara hierarkis. Selanjutnya, Ibrohim, Setiadi, dan Budi (2019) melakukan penelitian terkait identifikasi ujaran kebencian dan bahasa kasar menggunakan kombinasi dari fitur berbasis *word embeddings* dan juga *Part-of-Speech* atau kelas kata. Berikutnya, Ibrohim, Sazany, dan Budi (2019) melakukan identifikasi bahasa kasar menggunakan metode LSTM (*Long Short-Term Memory*).

Beberapa penelitian lain terkait topik deteksi ujaran kebencian adalah penelitian yang dilakukan oleh Sutejo dan Lestari (2018) yang membahas deteksi ujaran kebencian menggunakan pendekatan *deep learning* dengan metode LSTM pada dua tipe data, yaitu audio dan teks berbahasa Indonesia. Selain itu, Putri, Sriadhi, Sari, Rahmadani, dan Hutahaean (2020) mengumpulkan data dengan domain seputar politik dan agama untuk melakukan perbandingan model dalam klasifikasi ujaran kebencian. Terakhir, pada penelitian yang dilakukan oleh Fauzi dan Yuniarti (2018), deteksi ujaran kebencian dilakukan menggunakan metode *ensemble* dengan cara melatih beberapa model dan kemudian melakukan dua jenis *ensemble*, yaitu *hard voting* dan *soft voting* untuk mendapatkan keputusan akhir dari proses klasifikasi.

Sebagai ringkasan, pekerjaan terdahulu terkait deteksi ujaran kebencian pada dokumen berbahasa Indonesia fokus kepada pengembangan koleksi data dan pengembangan model klasifikasi. Setidaknya ada dua hal penting yang belum pernah dikaji secara mendalam pada penelitian-penelitian terdahulu. Pertama, belum terdapat pembahasan terkait *effect size*[3] dari fitur-fitur tekstual terhadap permasalahan deteksi ujaran kebencian dan bahasa kasar. Hasil kajian ini sangat berguna untuk mengedepankan aspek *explainability*

[3]Dalam penelitian ini, *effect size* bermakna besaran kontribusi dari suatu fitur terhadap hasil prediksi.

dari suatu model yang dapat menjadi dasar penilaian kekurangan model dan bias dalam data, untuk verifikasi hasil prediksi, untuk meningkatkan model, dan untuk mendapatkan wawasan terkait permasalahan (Samek, Wiegand, & Müller, 2017). Kedua, belum ada kajian tentang bias pada model yang disebabkan oleh data latih.

Pada pembuatan model, bias adalah informasi awal milik suatu sistem yang menggambarkan kecerdasan sistem berdasarkan hasil *training*. Data yang dikumpulkan di dunia nyata tidak bersifat homogen karena dihasilkan oleh sub kelompok sosial yang berbeda dengan karakteristiknya masing-masing, dan demografi manusia yang melakukan data labelling juga dapat memengaruhi bias (Bansal, 2022). Contoh potensi bias pada *dataset* Ibrohim dan Budi (2019) adalah ”jokowi” ditemukan sebagai kata yang paling sering muncul dalam data berlabel ”*hate speech*”, padahal kata ”jokowi” secara sendiri seharusnya tidak mengimplikasi ujaran kebencian. Pentingnya kajian terkait bias adalah karena model yang dilatih menggunakan data yang memiliki bias berpotensi menyebabkan hasil prediksi yang tidak adil dan tidak akurat (Bansal, 2022). Sehingga apabila dideteksi keberadaan bias pada *dataset*, penelitian ini diharapkan dapat menjadi dasar bagi penelitian selanjutnya terkait ujaran kebencian dan bahasa kasar untuk melakukan penanganan potensi bias sebelum melakukan pelatihan model.

Penelitian ini melanjutkan pekerjaan yang dilakukan oleh Ibrohim dan Budi (2019) dan fokus kepada kajian *effect size* fitur-fitur tekstual dan kepada potensi bias yang ditimbulkan oleh data latih. Terkait isu bias, pekerjaan ini berkontribusi dalam menghasilkan *dataset* uji bias yang dikembangkan secara otomatis dengan basis *dataset* oleh Ibrohim dan Budi (2019) dan juga yang dibangun secara manual. Beberapa model pemelajarn mesin kemudian diuji menggunakan *dataset* sintetik tersebut untuk melihat apakah menimbulkan bias atau tidak.

1.2 Rumusan Masalah

Seperti yang sudah dijelaskan pada Subbab 1.1, pekerjaan ini menggunakan koleksi data yang sudah dikembangkan oleh Ibrohim dan Budi (2019) untuk mempelajari hal-hal lanjutan seperti *effect size* fitur-fitur dan juga kemungkinan terjadinya bias pada model prediksi. Berikut adalah tiga rumusan masalah yang dibahas pada laporan penelitian ini:

- Faktor-faktor apa saja yang menyebabkan sebuah pesan blog mikro diklasifikasikan sebagai *Hate Speech* dan *Abusive*?

- Sejauh mana *hate speech* dan *abusive language* dapat dideteksi?
- Apakah ada bias stereotipe di *dataset* yang dibangun oleh Ibrohim dan Budi (2019)?

Terkait Rumusan Masalah 2, maknanya adalah untuk melakukan eksperimen dan mengusulkan berbagai macam metode untuk memprediksi data, kemudian menyimpulkan metode yang terbaik bisa mencapai berapa akurasinya. Fitur-fitur yang digunakan untuk prediksi data didapat berdasarkan hasil penelitian dalam menjawab Rumusan Masalah 1. Terakhir, makna dari Rumusan Masalah 3 adalah untuk melakukan uji deteksi bias terhadap metode-metode prediksi data yang diusulkan dalam menjawab Rumusan Masalah 2.

1.3 Tujuan Penelitian

Berikut ini adalah tujuan dari penelitian yang dilakukan:

- Melakukan analisis secara mendalam untuk mengidentifikasi faktor-faktor klasifikasi *Hate Speech* dan *Abusive Language* pada blog mikro berbahasa Indonesia menggunakan *dataset* yang dibangun oleh Ibrohim dan Budi (2019).
- Membuat, mengkaji, dan melakukan perbandingan antara model pembelajaran mesin dengan rekayasa fitur.
- Melakukan analisis mengenai keberadaan dan pengaruh bias stereotip pada *dataset* yang dibangun oleh Ibrohim dan Budi (2019).

1.4 Batasan Masalah

Semua eskperimen dan analisis yang dilakukan pada pekerjaan ini secara terbatas hanya terkait dengan *dataset* yang dikembangkan oleh Ibrohim dan Budi (2019). Sebagai informasi, *dataset* tersebut berisi kumpulan pesan blog mikro dan mayoritas berasal dari domain politik, khususnya kondisi politik saat pemilu 2019 di Indonesia. Perlu dicatat bahwa *dataset* berbahasa Indonesia yang lain (dan dari domain yang lain) mungkin saja tersedia di Internet dan bisa memberikan hasil analisis dan eksperimen yang berbeda dari apa yang dilaporkan di skripsi ini.

1.5 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
  Bab 1 mendiskusikan latar belakang dari permasalahan utama yang diangkat pada laporan tugas akhir ini, yaitu permasalahan deteksi ujaran kebencian dan penggunaan bahasa kasar serta kajian bias stereotipe yang ada pada koleksi data.

- Bab 2 STUDI LITERATUR
  Bab 2 mendiskusikan hasil studi literatur yang telah dilakukan terkait permasalahan yang diangkat.

- Bab 3 METODOLOGI
  Bab 3 mendiskusikan langkah-langkah yang dilakukan dalam penelitian ini terkait penyelesaian masalah dimulai dari *exploratory data analysis*, *feature selection*, perancangan model, evaluasi model, dan metode deteksi bias pada model.

- Bab 4 PERANCANGAN DAN IMPLEMENTASI
  Bab 4 mendiskusikan perancangan dan implementasi yang dilakukan untuk melakukan eksperimen berdasarkan metodologi yang di bahas di Bab 3.

- Bab 5 EKSPERIMEN DAN ANALISIS
  Bab 5 mendiskusikan hasil dan analisis dari eksperimen yang dilakukan.

- Bab 6 PENUTUP
  Bab 6 mendiskusikan kesimpulan yang dapat ditarik berdasarkan hasil analisa dari eksperimen yang telah dilakukan dan saran untuk penelitian berikutnya.

BAB 2
STUDI LITERATUR

Bab 2 mendiskusikan hasil studi literatur yang telah dilakukan terkait permasalahan yang diangkat. Pertama, Subbab 2.1 membahas definisi dan penggunaan dari ujaran kebencian dan bahasa kasar, serta pentingnya pendeteksian otomatis untuk keduanya. Subbab 2.2 menyajikan informasi terkait proses pengecekan data menggunakan metode *exploratory data analysis*. Kemudian, konsep relevansi fitur dan metode pemilihan fitur dijelaskan di Subbab 2.3. Selanjutnya, Subbab 2.4 menyampaikan tentang pendekatan yang ada untuk tugas deteksi ujaran kebencian dan bahasa kesar. Subbab 2.5 membahas model pemelajaran mesin yang dipakai dalam penelitian ini. Subbab 2.6 menjelaskan metrik evaluasi yang dipakai untuk menilai hasil pemodelan yang dilakukan di penelitian ini. Terakhir, Subbab 2.7 membahas kajian bias stereotipe dan metode deteksi bias pada model.

2.1 Ujaran Kebencian dan Bahasa Kasar

Penelitian ini membahas pendeteksian ujaran kebencian dan bahasa kasar pada data berbentuk blog mikro berbahasa Indonesia. Sebagai pengetahuan dasar untuk memahami target deteksi, perlu diketahui definisi dari ujaran kebencian dan bahasa kasar, karakteristik yang membuat suatu ujaran memiliki unsur kebencian, dan karakteristik yang menunjukkan penggunaan bahasa kasar.

2.1.1 Definisi Ujaran Kebencian dan Bahasa Kasar

Ujaran kebencian adalah bentuk komunikasi yang dilakukan dengan tujuan merendahkan atau menyatakan prasangka buruk (stigma) terhadap seseorang atau sekelompok orang berdasarkan karakteristik khas mereka seperti ras, warna kulit, gender, orientasi seksual, kebangsaan, agama, dan sebagainya (Nockleby, 2000). Pada penelitian yang dilakukan oleh Waseem dan Hovy (2016) beberapa kriteria yang dipakai untuk melakukan deteksi ujaran kebencian adalah penggunaan bahasa kasar yang bersifat merendahkan gender atau ras tertentu, menargetkan kelompok minoritas, mempromosikan ujaran kebencian atau kekerasan, memutar balikan fakta, penggunaan hashtag kontroversial, mendukung xeno-

fobia atau seksisme, dan penggunaan username yang bersifat menyinggung (Alfina et al., 2017).

Bahasa kasar adalah suatu ucapan atau ekspresi yang mengandung kata kasar atau kotor. Dalam bahasa Indonesia, penggunaan bahasa kasar biasanya terdiri dari kata-kata yang menyinggung kondisi atau situasi yang memiliki implikasi negatif atau berupa nama-nama hewan. Contoh kondisi atau situasi dengan implikasi negatif bisa berupa kesehatan mental, tingkat kecerdasan, orientasi seksual, status finansial, dan sebagainya. Dengan tujuan mengurangi kontradiksi dalam literatur yang mendefinisikan bahasa kasar dan mempermudah tugas klasifikasi yang berhubungan dengan kata kasar, Waseem et al. (2017) mengelompokkan bahasa kasar berdasarkan dua faktor yaitu target dan tingkatnya. Berdasarkan faktor target, bahasa kasar dapat dikelompokkan ke dalam bahasa kasar yang memiliki target spesifik atau individu dan target general terhadap kelompok tertentu. Berdasarkan tingkatnya, bahasa kasar dapat dikelompokkan ke dalam bahasa kasar eksplisit dan implisit. Berikut adalah contoh-contoh kalimat bahasa kasar berdasarkan kategori-kategori yang telah disebutkan:

1. “Udah gila ya lu?”
2. “Terserah dah, debat sama orang tolol ga ada kelarnya.”
3. ”Emang si paling pinter, *rush mid* bawa SMG.”
4. “Harga naik aja udah berisik, dasar orang-orang miskin.”

Contoh 1 adalah bahasa kasar dengan target individu yang menyinggung kesehatan mental dan termasuk kategori eksplisit karena dengan jelas menggunakan kata kasar ”gila[1]”. Contoh 2 dan 3 keduanya merupakan bahasa kasar dengan target individu yang menyinggung kecerdasan target, tetapi contoh 2 termasuk kategori eksplisit karena langsung mengandung kata kasar ”tolol[2]” sedangkan contoh 3 menyembunyikannya dengan kata yang bermakna sebaliknya. Contoh 4 adalah bahasa kasar dengan target kelompok yang menyinggung sekelompok orang berdasarkan kondisi finansial mereka.

[1]Berdasarkan KBBI daring, gila memiliki arti gangguan jiwa; sakit ingatan; sakit jiwa.
[2]Berdasarkan KBBI daring, tolol memiliki arti sangat bodoh; bebal

2.1.2 Ujaran Kebencian Implisit dan Eksplisit

Pada awal dilakukannya penelitian terkait ujaran kebencian, terdapat lebih banyak fokus terhadap ujaran kebencian yang bersifat eksplisit dengan pemanfaatan fitur-fitur leksikal. Nyatanya, ujaran kebencian tidak selalu diungkapkan secara eksplisit, tetapi juga dapat disampaikan menggunakan pilihan bahasa yang secara implisit mengandung kebencian (ElSherief et al., 2021; Lin, 2022). Berikut adalah contoh ujaran kebencian implisit dan eksplisit dalam bahasa Indonesia:

1. "Kita balas di sini, gimana kalau kita usir dan bantai umat Buddha."
2. "Indonesia belum merdeka, Indonesia masih dijajah. Takbir!"

Contoh 1 adalah ujaran kebencian eksplisit yang secara langsung menunjukkan kebencian terhadap kelompok agama tertentu, sedangkan contoh 2 adalah ujaran kebencian implisit yang secara tidak langsung mengasingkan kelompok tertentu di Indonesia. Pada contoh 1 terlihat terdapat penggunaan kata-kata tertentu yang secara jelas mengekspresikan kebencian, sedangkan pada contoh 2 kebencian sekedar berupa implikasi. Dalam kasus ujaran kebencian implisit, diperlukan pengetahuan kultur atau latar mengenai isi dari ujaran untuk dapat menemukan aspek kebenciannya. Seseorang mungkin tidak akan melihat aspek kebencian dalam contoh 2 apabila tidak memiliki pengetahuan awal atau mengetahui adanya kelompok yang memiliki opini bahwa Indonesia adalah negara Islam dan cenderung menstigmatisasi agama lain di Indonesia bahkan sampai terjadi kekerasan terhadap masyarakat dan tempat suci dari agama-agama terkait (Herbayu, 2013).

2.1.3 Urgensi Deteksi Ujaran Kebencian dan Bahasa Kasar

Adanya penyebaran ujaran kebencian dapat menyebabkan bencana sosial termasuk diskriminasi dan konflik antar kelompok masyarakat, hingga genosida manusia, seperti kasus genosida etnis Tutsi di Rwanda pada tahun 1994 (HAM, 2015; Stanton, 2009). Kepentingan adanya deteksi ujaran kebencian pada konten online muncul karena ujaran kebencian pada media sosial memudahkan tumbuhnya kekerasan di dunia nyata (Lin, 2022). Terdapat hubungan erat antara ujaran kebencian dan terjadinya kejahatan berdasarkan kebencian (*hate crime*) dan pendeteksian terhadap ujaran kebencian dapat memungkinkan adanya intervensi untuk mencoba menghentikan eskalasi dari ujaran ke aksi kejahatan (Waseem & Hovy, 2016).

2.1.4 Penelitian Terdahulu dan Posisi Penelitian

Penelitian terdahulu terkait ujaran kebencian dan bahasa kasar berbahasa Indonesia dilakukan dengan fokus terhadap pengembangan *dataset* dan pengembangan model klasifikasi. Penelitian awal terkait deteksi ujaran kebencian telah dilakukan oleh Alfina et al. (2017) yang telah mengembangkan *dataset* ujaran kebencian dan melakukan deteksi ujaran kebencian menggunakan model pemelajaran mesin konvensional. Sedangkan untuk deteksi bahasa kasar, penelitian awal berupa pengembangan *dataset* bahasa kasar dan deteksi bahasa kasar pada media sosial berbahasa Indonesia telah dilakukan pada penelitian oleh Ibrohim dan Budi (2018). Kemudian, penelitian oleh Ibrohim, Sazany, dan Budi (2019) juga melakukan identifikasi bahasa kasar dengan mengusulkan penggunaan metode *deep learning* berupa LSTM (*Long Short-Term Memory*).

Penelitian tugas akhir skripsi ini menggunakan *dataset* yang dikembangkan oleh Ibrohim dan Budi (2019) untuk klasifikasi multi-label ujaran kebencian dan bahasa kasar. Pada penelitian yang dilakukan oleh Ibrohim dan Budi, anotasi dilakukan untuk membagi data yang berlabel ujaran kebencian ke dalam lima kategori ujaran kebencian, lalu membagi berdasarkan target kebencian berupa individu atau kelompok, dan menilai tingkat keparahan dari ujaran kebencian. *Dataset* ini juga digunakan dalam penelitian yang dilakukan oleh Prabowo et al. (2019) untuk melakukan klasifikasi multi-label secara hirarkis terkait label spesifik dari data berlabel ujaran kebencian. Selanjutnya, penelitian oleh Ibrohim, Setiadi, dan Budi (2019) melakukan klasifikasi ujaran kebencian dan bahasa kasar menggunakan fitur semantik berupa kombinasi dari fitur berbasis *word embeddings* dan juga *Part-of-Speech* atau kelas kata.

Selain itu, penelitian oleh Sutejo dan Lestari (2018) melakukan deteksi ujaran kebencian pada data bertipe teks dan audio dengan menggunakan pendekatan *deep learning* berupa metode LSTM. Penelitian oleh Putri et al. (2020) melakukan pengembangan *dataset* ujaran kebencian dengan domain politik dan agama untuk mengembangkan dan membandingkan model pemelajaran mesin dalam memprediksi ujaran kebencian. Terakhir, Fauzi dan Yuniarti (2018) telah melakukan penelitian yang mengusulkan metode *ensemble* untuk melakukan deteksi ujaran kebencian. Tabel 2.1 menggambarkan posisi penelitian tugas akhir skripsi ini dibandingkan sembilan referensi yang telah disebut terkait deteksi ujaran kebencian dan bahasa kasar berbahasa Indonesia.

Tabel 2.1: Tabel Posisi Penelitian

No.	Penelitian	Tujuan Penelitian	Model
1	Alfina et al. (2017)	Mengembangkan dataset ujaran kebencian Deteksi ujaran kebencian	*Naive Bayes* *Support Vector Machine* *Bayesian Logistic Regression* *Random Forest Decision Tree*
2	Ibrohim dan Budi (2018)	Mengembangkan dataset bahasa kasar Deteksi bahasa kasar	*Naive Bayes* *Support Vector Machine* *Random Forest Decision Tree*
3	Ibrohim dan Budi (2019)	Mengembangkan dataset ujaran kebencian dan bahasa kasar Deteksi ujaran kebencian dan bahasa kasar secara *multi-label*	*Naive Bayes* *Support Vector Machine* *Random Forest Decision Tree*
4	Prabowo et al. (2019)	Deteksi ujaran kebencian dan bahasa kasar secara *multi-label* Klasifikasi *multi-label* secara hirarkis	*Naive Bayes* *Support Vector Machine* *Random Forest Decision Tree*
5	Ibrohim, Setiadi, dan Budi (2019)	Identifikasi ujaran kebencian dan bahasa kasar Klasifikasi menggunakan fitur semantik Menggunakan fitur *word2vec*, , *emoji*	*Logistic Regression* *Random Forest Decision Tree* *Support Vector Machine*
6	Ibrohim, Sazany, dan Budi (2019)	Identifikasi bahasa kasar Membandingkan performa NB dengan LSTM	*Naive Bayes* *Long Short-Term Memory*
7	Sutejo dan Lestari (2018)	Deteksi ujaran kebencian Menggunakan fitur tekstual dan fitur akustik	*Long Short-Term Memory*
8	Putri et al. (2020)	Mengembangkan dataset ujaran kebencian Deteksi ujaran kebencian Membandingkan algoritma klasifikasi	*Naive Bayes* *Decision Tree* *Multi-layer Perceptron* *AdaBoost Classifier*
9	Fauzi dan Yuniarti (2018)	Deteksi ujaran kebencian Menggunakan metode *ensemble*	*Naive Bayes* *K-Nearest Neighbours* *Maximum Entropy* *Random Forest* *Support Vector Machine*
10	Penelitian ini	Mengkaji *effect size* dan bias Deteksi ujaran kebencian secara biner Deteksi bahasa kasar secara biner	*Logistic Regression* *Extreme Gradient Boosting* *CatBoost Classifier* *Multi-layer Perceptron* *Support Vector Machine*

2.2 Exploratory Data Analysis

Exploratory Data Analysis atau EDA adalah proses yang dilakukan peneliti untuk mendapatkan informasi mengenai *dataset* dari berbagai sudut pandang. Informasi dapat ditampilkan dalam bentuk tabel atau visualisasi menggunakan grafik. Tahap ini memberikan peneliti gambaran umum mengenai struktur dari data, visualisasi dan kesimpulan dari fitur-fitur pada data, dan dasar untuk analisis lebih lanjut (Chatfield, 1986). Menurut Chatfield, EDA tidak bisa didefinisikan secara pasti dan memberi penjelasan secara garis besar

bahwa pengecekan kualitas data, perhitungan kesimpulan statistik, visualisasi grafik, dan penggunaan teknik data analisis seperti principal componen analysis (PCA) semua merupakan tugas yang termasuk dalam proses EDA. Sedangkan menurut Seltman (2012), EDA mengacu pada eksplorasi data dengan metode apapun tanpa menggunakan pemodelan atau inferensi statistik (Data, Komorowski, Marshall, Salciccioli, & Crutain, 2016). Pada penelitian ini, EDA digunakan untuk melakukan eksplorasi data teks secara ortografi dan leksikal. Kemudian, temuan yang didapat dari hasil proses EDA akan dijadikan dasar penentuan fitur-fitur yang akan dikaji *effect size*-nya dan digunakan dalam pelatihan model pemelajaran mesin.

2.3 Feature Selection

Seperti yang sudah disampaikan di Bab 1, salah satu tujuan pada penelitian ini adalah untuk mengembangkan model berbasis *data-driven* yang mampu mendeteksi apakah sebuah pesan di blog mikro termasuk ujaran kebencian atau bukan, dan termasuk bahasa kasar atau bukan. *Feature selection* adalah sebuah tahap penting pada pengembangan model klasifikasi berbasis pemelajaran mesin. Pada tahap ini, dilakukan kajian *effect size* berdasarkan fitur-fitur yang telah didapat dari tahap sebelumnya. Caranya adalah dengan menggunakan metode-metode *feature selection* untuk melakukan pemilihan fitur-fitur relevan. Fitur-fitur yang tidak memberikan informasi bermanfaat dianggap tidak relevan dan fitur-fitur yang tidak memberi informasi tambahan dari fitur yang sudah terpilih dianggap fitur berulang. Tujuan dari *feature selection* adalah untuk meningkatkan performa algoritma data mining, meningkatkan akurasi dari hasil prediksi, dan membuat data lebih mudah dipahami (Kumar & Minz, 2014).

2.3.1 Analysis of Variance F-test

Analysis of variance (ANOVA) adalah metode statistik untuk melakukan uji hipotesis yang digunakan ketika terdapat variabel numerik dan variabel kategorik, seperti pada penelitian ini di mana variabel input fitur memiliki nilai numerik dan target klasifikasi merupakan variabel kategorik. Penggunaan metode ANOVA biasanya memiliki tujuan untuk mendapatkan informasi terkait pengaruh dari berbagai faktor terhadap hasil atau reaksi yang didapat. Untuk tujuan tersebut, metode ANOVA menggunakan sebuah kelas uji statistik yang disebut F-test (St, Wold, et al., 1989). Perhitungan F-test dilakukan

sebagai berikut:

$$F = \frac{\mathsf{MSB}}{\mathsf{MSW}}. \tag{2.1}$$

Pada Persamaan 2.1, MSB (*mean square between*) diperoleh dengan membagi *sum of squares between* (SSQb) dengan *degree of freedom between* (dfB) dan MSW (*mean square within*) diperoleh dengan membagi *sum of squares within* (SSQw) dengan *degree of freedom within* (dfQ) seperti pada persamaan berikut:

$$\mathsf{SSQb} = \sum_{j=1}^{p} n_j (x_j - x)^2, \text{ dengan } \mathsf{dfB} = p - 1, \tag{2.2}$$

$$\mathsf{SSQw} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{i,j} - x_j)^2, \text{ dengan } \mathsf{dfW} = \sum_{j=1}^{p} n_j - 1, \tag{2.3}$$

dengan variabel p adalah banyak kelas; n_j adalah jumlah pengamatan dalam kelas j; x_j adalah rata-rata kelas j; dan x adalah rata-rata umum. Formula ini dibangun sehingga hasil perhitungan nilai F adalah suatu bilangan positif di mana jika F memiliki nilai yang signifikan lebih dari $1,0$ menunjukkan adanya perbedaan antar kelas (St et al., 1989). Penelitian yang dilakukan oleh Steiger (2004) menunjukkan bahwa F score dalam metode ANOVA memiliki informasi mengenai *effect size* dan estimasi presisi. Pada penelitian ini ANOVA F-test digunakan untuk menentukan *effect size* dari masing-masing fitur yang diusulkan untuk pengembangan model klasifikasi.

2.3.2 *Logistic Regression Analysis*

Logistic Regression Analysis (LRA) adalah sebuah penerapan teknik analisis regresi berganda untuk mempelajari kasus-kasus di mana target klasifikasi merupakan variabel kategorik dan dikotomis[3] (Dayton, 1992). Contohnya pada penelitian ini, prediksi dibuat untuk menilai apakah suatu dokumen blog mikro termasuk ujaran kebencian atau bukan, dan termasuk bahasa kasar atau bukan. Metode LRA bekerja berdasarkan probabilitas terkait dengan nilai Y (variabel target klasifikasi). Hubungan variabel prediktor dengan

[3]Berdasarkan KBBI daring, dikotomi adalah pembagian atas dua kelompok yang saling bertentangan.

kemungkinan variabel target bernilai 1 dimodelkan dalam persamaan:

$$\log_e \left[\frac{\pi}{1-\pi} \right] = \alpha + \beta_1 X_1 + ... + \beta_p X_p = \alpha + \sum_{j=1}^{p} \beta_j X_j \,, \tag{2.4}$$

$$\text{dengan} \;\; \pi = P(Y = 1 | X_1, ..., X_p) \,. \tag{2.5}$$

Secara teori, $\pi = P(Y = 1)$ pada Persamaan 2.5 didefinisikan sebagai proposi populasi dari kasus di mana $Y = 1$, dan $1 - \pi = 1 - P(Y = 1)$ untuk kasus di mana $Y = 0$. Tanpa informasi lain, perhitungan dilakukan menggunakan porporsi sampel di mana $Y = 1$ pada setiap kasus, tetapi untuk konteks regresi diasumsikan ada sejumlah variabel prediktor $X_1 + ... + X_p$ yang terhubung ke Y sebagai data tambahan untuk memprediksi Y (Dayton, 1992). Probabilitas $Y = 1$ dan $Y = 0$ masing-masing dapat direpresentasikan dengan persamaan berikut:

$$P(Y = 1 | X_1, ..., X_p) = \frac{\exp(\alpha + \sum_{j=1}^{p} \beta_j X_j)}{1 + \exp(\alpha + \sum_{j=1}^{p} \beta_j X_j)} = \frac{1}{1 + \exp(-\alpha - \sum_{j=1}^{p} \beta_j X_j)} \tag{2.6}$$

$$P(Y = 0 | X_1, ..., X_p) = 1 - P(Y = 1 | X_1, ..., X_p) = \frac{1}{1 + \exp(\alpha + \sum_{j=1}^{p} \beta_j X_j)} \tag{2.7}$$

Dari model di atas, menggunakan *maximum likelihood estimation* (MLE), nilai koefisien (*coef*) diestimasi dengan mencari nilai dari $\alpha, \beta_1, ..., \beta_p$ yang memberikan nilai maksimum pada fungsi *Likelihood* berikut:

$$L = \prod_{i=1}^{n} P(Y_i | X_{i,1}, ..., X_{i,p}) \,, \tag{2.8}$$

dengan n adalah banyak baris di data.

Sebagai contoh, Tabel 2.2 menampilkan sebuah hasil dari proses estimasi parameter pada proses analisis *logistic regression*, yaitu dengan:

$$\log_e \left[\frac{\pi}{1-\pi} \right] = -0,3968 + 1,3470 \times \text{jokowi} \,, \tag{2.9}$$

yang berarti nilai koefisien *Intercept* adalah $\alpha = -0,3968$ dan nilai koefisien yang terkait fitur jokowi adalah nilai $\beta = +,3470$. Mengikuti konteks penelitian ini, makna dari Persamaan 2.9 adalah ketika fitur biner "jokowi" bernilai 1, maka *odds ratio* dokumen blog

mikro dideteksi sebagai *hate speech* menjadi meningkat sebesar $exp(1.3470)$ atau 3,85 kali lipat daripada ketika ”jokowi” bernilai 0.

Tabel 2.2: Contoh LRA pada fitur keberadaan kata ”jokowi” untuk label ”*Hate Speech*”

	coef	std err	z	$P > \|z\|$
Intercept	-0,3968	0,037	-10,805	0,000
jokowi	1,3470	0,593	5,655	0,000

2.3.3 Shapley Additive Explanations (SHAP)

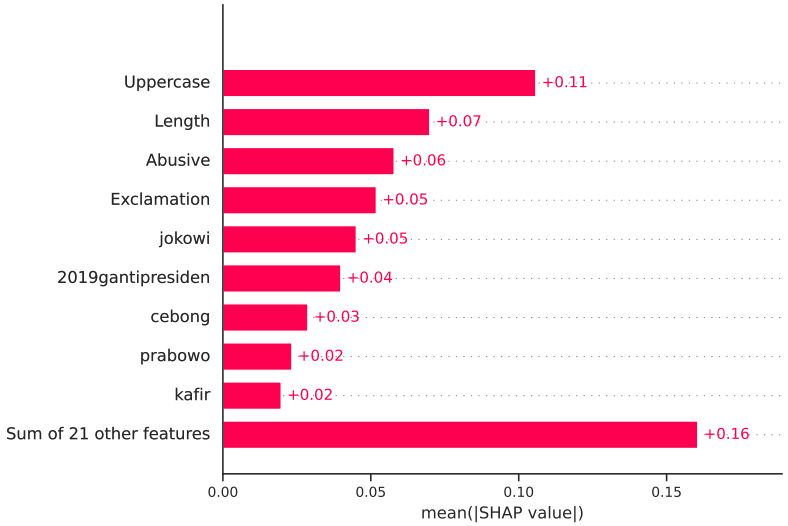
SHAP (*SHapley Additive exPlanations*) adalah teknik teori permainan yang merupakan sebuah pendekatan untuk menjelaskan output suatu model pemelajaran mesin dengan cara memberi nilai relevansi untuk prediksi tertentu pada setiap fitur, dengan kata lain setiap prediksi dijelaskan dengan cara menghitung kontribusi dari masing-masing fitur dari dataset ke hasil prediksi. SHAP menghubungkan alokasi kredit yang optimal dengan penjelasan lokal menggunakan nilai *Shapley* dari teori permainan dan ekstensi terkait lainnya. Sehingga fitur-fitur berpartisipasi sebagai ‘pemain’ dalam koalisi, dan nilai *Shapley* digunakan untuk menentukan cara membagi ‘pembayaran’, dalam kasus ini hasil prediksi, di antara para ‘pemain’ secara adil. Penggunaan nilai *Shapley* untuk menjelaskan hasil prediksi model digambarkan sebagai metode atribusi fitur tambahan (Lundberg & Lee, 2017; Marcílio & Eler, 2020; Molnar, 2020).

2.3.3.1 Nilai *Shapley*

Nilai *Shapley* adalah kontribusi marjinal rata-rata dari nilai fitur di semua kemungkinan koalisi. Lloyd *Shapley* mengusulkan bahwa ada kemungkinan nilai dari memainkan suatu game dapat dievaluasi secara numerik dan mengembangkan fungsi yang dikenal sebagai nilai *Shapley*. Saat ini, nilai *Shapley* memiliki domain yang telah diperluas dan banyak fungsi terkait yang ditemukan dari melonggarkan beberapa asumsi yang ada (Shapley & Roth, 1988).

2.3.3.2 SHAP Feature Importance

Salah satu fungsi penggunaan SHAP adalah untuk menilai kepentingan suatu fitur dengan menggunakan nilai *Shapley*. Kepentingan suatu fitur dinilai dari seberapa besar



Gambar 2.1: Contoh hasil visualisasi *feature importance* menggunakan SHAP

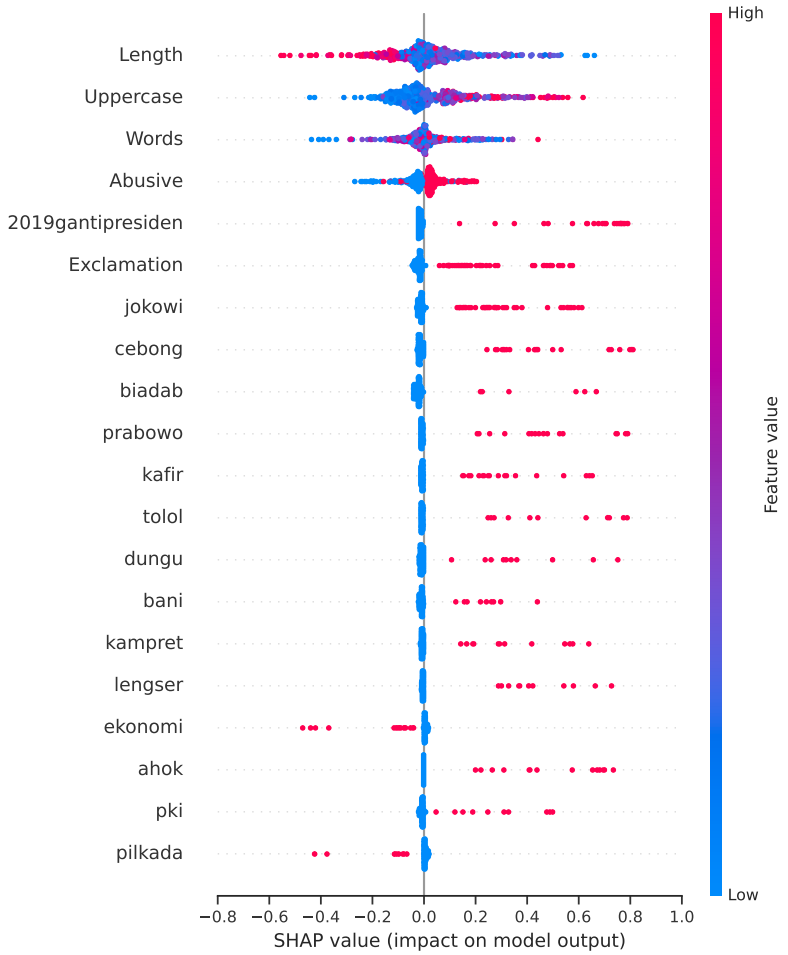
nilai *Shapley* absolut dari fitur tersebut. Nilai yang digunakan adalah rata-rata dari nilai *Shapley* absolut ($\text{mean}(|SHAPvalue|)$) untuk setiap fitur, karena tujuannya adalah untuk menemukan kepentingan fitur secara keseluruhan Molnar (2020). Dapat dilihat pada contoh Gambar 2.1 fitur *Uppercase* adalah fitur paling penting karena fitur tersebut mengubah rata-rata absolut dari probabilitas ujaran kebencian sebesar 11 poin persentase (0,11). Semakin besar nilai $\text{mean}(|SHAPvalue|)$ dari suatu fitur, semakin besar pengaruh fitur tersebut terhadap keputusan akhir model atau hasil prediksi.

2.3.3.3 SHAP Summary Plot

Pada penelitian ini SHAP juga digunakan untuk visualisasi *summary plot* dengan menggabungkan *feature importance* dan *feature effects*. Pada sumbu y ditampilkan beberapa fitur yang paling berpengaruh dan pada sumbu x ditampilkan nilai *Shapley*. Warna yang ditampilkan merupakan representasi *feature value* suatu fitur untuk masing-masing dokumen (Molnar, 2020). Pada contoh Gambar 2.2 fitur *Length* memiliki *feature importance* tertinggi dengan *feature value* yang tinggi dalam memutuskan suatu dokumen blog mikro tidak terdeteksi sebagai ujaran kebencian.

2.4 Pendekatan untuk Deteksi Ujaran Kebencian dan Bahasa Kasar

Tugas klasifikasi memiliki tujuan untuk memasukkan data ke dalam kelas-kelas yang diketahui. Dalam penelitian ini permasalahan deteksi ujaran kebencian dan bahasa kasar



Gambar 2.2: Contoh hasil visualisasi *feature importance* menggunakan SHAP

masing-masing adalah tugas klasifikasi biner, di mana dilakukan prediksi untuk mendapatkan hasil apakah suatu dokumen termasuk ke dalam kelas satu atau lainnya di antara dua pilihan. Sebagai contoh, pada penelitian ini data diprediksi apakah termasuk ke dalam kelas '*Hate Speech*' atau '*Non Hate Speech*', selain itu juga diprediksi apakah termasuk ke dalam kelas '*Abusive*' atau '*Not Abusive*'. Untuk melakukan tugas klasifikasi untuk mendeteksi apakah suatu dokumen memiliki unsur ujaran kebencian atau bahasa kasar, diketahui dua metode deteksi, yaitu metode deteksi *rule-based* dan pendekatan *data-driven* menggunakan pemelajaran mesin.

2.4.1 Rule-Based Detection

Rule-based detection adalah metode deteksi yang memerlukan keberadaan ahli sebagai penentu aturan-aturan (*rule*) untuk menjadi dasar klasifikasi. Aturan yang sudah dibuat oleh pakar atau ahli kemudian dimasukkan ke dalam komputer menjadi bentuk *if-then-*

else yang sangat banyak. Tidak seperti model pemelajaran mesin, model *rule-based* selalu dapat dijelaskan dan diinterpretasikan karena tidak bersifat *black box*, tetapi banyak usaha dan keahlian yang dibutuhkan dalam pembangunan model (Gémes, Kovács, Reichel, & Recski, 2021). Metode *rule-based detection* tidak akan dipakai dalam penelitian ini karena membutuhkan biaya untuk ahli yang menetapkan aturan-aturan, prosesnya memakan banyak waktu, dan diperlukannya *maintenance* terus menerus dengan munculnya kata kasar baru dan berubahnya sifat ujaran kebencian.

2.4.2 Pendekatan *Data-Driven* (Pemelajaran Mesin)

Data didefinisikan oleh *Oxford Learner's Dictionary* sebagai kumpulan fakta atau informasi yang digunakan untuk membuat keputusan atau sebagai sumber pemelajaran. Pemelajaran mesin adalah bidang dengan titik fokus dalam pengembangan algoritma yang meningkat secara performa seiring berjalannya waktu berdasarkan pengalaman, seperti bagaimana manusia belajar (Ketkar & Ketkar, 2017). Model pemelajaran mesin menentukan prediksi atau keputusan berdasarkan hasil pemelajaran dari data yang dimiliki, bukan berdasarkan aturan-aturan yang ditentukan dan tidak ada pedoman pasti mengenai apa yang menyebabkan suatu keputusan dibuat.

2.4.2.1 Pemelajaran Mesin Konvensional

Pemelajaran mesin konvensional adalah pengembangan model yang memerlukan *feature engineering* secara manual. Pemelajaran mesin konvensional memiliki kemampuan terbatas untuk memproses data secara langsung dalam bentuk aslinya, tidak seperti *deep learning* yang bisa melakukan ekstraksi fitur secara otomatis. Sehingga, dalam proses pengembangan suatu model pemelajaran mesin konvensional, diperlukan pemahaman dan keahlian untuk mengolah dan melakukan ekstraksi fitur dari data, seperti penggunaan teknik *feature selection* dan pengusulan fitur (Chauhan & Singh, 2018).

2.4.2.2 Deep Learning

Pemelajaran mesin memiliki sub-bidang yang disebut *deep learning*, sebuah pendekatan pemelajaran mesin yang mempunyai kemampuan untuk ekstraksi fitur secara otomatis tanpa perlu tahap *feature engineering*. Menurut Setiowati, Franita, Ardiyanto, et al. (2017) dan Chauhan dan Singh (2018), secara umum *deep learning* memiliki performa

lebih baik dibandingkan pemelajaran mesin konvensional untuk permasalahan seperti pemahaman bahasa manusia dan *computer vision*, tetapi ini tidak berlaku secara mutlak; pemelajaran mesin konvensional bisa lebih baik jika *feature engineering* berhasil dilakukan dengan tepat. *Deep neural network* menggunakan model non-linear dengan sejumlah *hidden layer* untuk menghasilkan sistem yang dapat memelajari hubungan kompleks antara *input* dan *output*-nya (Chauhan & Singh, 2018).

2.5 Model Pemelajaran Mesin

Subbab 2.5 menjelaskan cara kerja masing-masing model pemelajaran mesin yang digunakan dalam penelitian ini. Setiap model dipakai untuk dua tugas klasifikasi biner, yaitu klasifikasi dengan target "*Hate Speech*"/"*Not Hate Speech*" dan "*Abusive*"/"*Not Abusive*."

2.5.1 Logistic Regression

Logistic Regression adalah sebuah model yang dapat digunakan untuk klasifikasi biner (Komarek, 2004). *Logistic Regression* memiliki cara kerja yang serupa dengan metode regresi lainnya, tetapi digunakan untuk melakukan prediksi pada *dataset* dengan target klasifikasi yang bersifat biner dan dikotomis. Contohnya seperti "benar" atau "salah", "1" atau "0", dan pada penelitian ini target klasifikasi adalah "Hate Speech" atau "Not Hate Speech" dan "Abusive" atau "Not Abusive". Selain itu *Logistic Regression* melakukan *fitting* data ke dalam fungsi logistik yang memiliki bentuk menyerupai huruf S dengan nilai *y-axis* mulai dari 0 ke 1, bukan ke dalam suatu garis lurus. Metode pencarian koefisien dalam pembangunan model klasifikasi *logistic regression* telah dibahas di Subbab 2.3 pada halaman 12 bagian *Logistic Regression Analysis*.

2.5.2 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boostling (XGBoost) adalah suatu sistem *tree boosting* yang dapat diskalakan. Penelitian terkait XGBoost pertama kali dilakukan oleh Chen dan Guestrin (2016) dengan memanfaatkan *gradient-boosted decision trees* untuk memaksimalkan kecepatan dan performa. XGBoost dapat memanfaatkan secara maksimal kapasitas perangkat keras sehingga memiliki kelebihan dalam peningkatan algoritma, *model tuning*, dan juga dapat diterapkan dalam lingkungan komputasi. XGBoost merupakan imple-

mentasi dari model *optimized gradient boosted tree* dan juga merupakan sebuah model *regularized* yang menghitung kompleksitas *tree* (Dhaliwal, Nahid, & Abbas, 2018). Sehingga fungsi objektif memiliki dua komponen, *Training Loss* (TL) yang mengukur seberapa prediktif model, dan *Regularization* (R) yang membantu menjaga kompleksitas model untuk tetap berada dalam batas:

$$Obj(\theta) = TL(\theta) + R(\theta). \tag{2.10}$$

XGBoost mencoba melakukan optimisasi *tree* yang dipelajari dengan menambah sebuah *tree* pada setiap langkah. Pada masing-masing langkah, *tree* yang ditambah adalah *tree* yang memenuhi optimisasi fungsi objektif. Cara XGBoost mengembangkan *tree* adalah menggunakan *greedy tree growing algorithm* dimulai dari *tree* dengan kedalaman 0. Untuk setiap *leaf*, akan dicoba dilakukan *split*. Dalam melakukan *splitting*, dicari nilai *split* yang memaksimalkan nilai *gain* dan *split* berhenti dilakukan ketika *split* terbaik memiliki nilai *gain* yang negatif. Nilai gain adalah jumlah nilai *left child* ditambah *right child* dan dikurangi dengan agregasi dari keduanya apabila tidak dilakukan *split*.

2.5.3 CatBoost

CatBoost adalah contoh lain dari implementasi *gradient boosting framework*. CatBoost menggunakan prediktor dasar berupa *binary decision trees*, sebuah model yang dibuat dengan membagi ruang fitur (*feature space*) secara rekursif menjadi berbagai daerah terpisah (*disconnected regions*) berdasarkan nilai dari beberapa atribut *splitting* (Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, 2018). Fungsi *hx* yang merepresentasikan *decision tree h* terhadap input *x* adalah:

$$h(x) = \sum_{j=1}^{J} b_j l_{\{x \in R_j\}}, \tag{2.11}$$

dengan R_j adalah daerah terpisah (*disjoint region*) yang sebanding dengan *leaves* dari *tree*; $l_{\{x \in R_j\}} = a$ adalah atribut-atribut *splitting*, biasanya merupakan variabel biner yang mengidentifikasi bahwa suatu fitur x^k melebihi suatu *threshold* t; dan $t = 0.5$ untuk kasus di mana x^k bernilai biner.

2.5.4 Multi-layer Perceptron

Multi-layer Perceptron (MLP) adalah suatu bentuk *artificial neural network* (ANN) yang digunakan untuk berbagai permasalahan seperti *pattern recognition* dan *interpolation* (Noriega, 2005). ANN adalah struktur yang dimodelkan berdasarkan cara otak beroperasi, sehingga memiliki kapabilitas untuk melakukan estimasi fungsi model dan menangani fungsi secara linear dan non-linear (Taud & Mas, 2018). Arsitektur model MLP memiliki setidaknya tiga layer, yaitu *input layer*, *hidden layer*, dan *output layer*, dengan jumlah *hidden layer* dan jumlah *node* pada masing-masing *layer* yang dapat bervariasi. Semakin banyak *node* yang digunakan, model akan lebih sensitif terhadap permasalahan, tetapi kemungkinan *overfitting* akan meningkat (Murtagh, 1991).

2.5.5 Support Vector Machine

Support Vector Machine (SVM) adalah sebuah algoritma pemelajaran mesin yang bekerja dengan menemukan *hyperplane* yang dapat memaksimalkan jarak antara dua kelas. Dalam SVM, fitur dipetakan menggunakan koordinat berdasarkan hubungan suatu fitur dengan fitur lain, membentuk *support vectors* (Pisner & Schnyer, 2020). Sebuah mesin *support vector* dapat berbentuk linear atau non-linear. SVM linear merepresentasikan hasil prediksi menggunakan model matematis $y = wx' + \gamma$ dengan W adalah vektor bobot atau koefisien; x' adalah vektor fitur; dan γ adalah *intercept*. Persamaan tersebut dimanipulasi agar didapatkan pembagian daerah secara linear (Suthaharan & Suthaharan, 2016). Ketika data tidak bisa dibagi secara linear menggunakan garis atau *hyperplane*, fungsi kernel digunakan untuk memetakan data ke dalam ruang dimensi lebih tinggi di mana pembagian daerah secara linear dapat dilakukan (Gidudu, Hulley, & Marwala, 2007).

2.6 Metode Evaluasi

Pada penelitian ini dilakukan evaluasi model untuk membandingkan performa satu model dengan model lainnya dari masing-masing tugas klasifikasi. Evaluasi model dilakukan dalam bentuk uji statistik untuk mendapatkan performa model secara umum, terlepas dari seberapa bagus pembagian data *train* dan *test*.

2.6.1 Bootstrap

Bootstrap adalah suatu metode uji statistik yang dapat digunakan pada sampel data berbentuk apapun (Singh & Xie, 2008). Cara kerja *bootstrap* adalah dengan melakukan pengambilan sampel ulang (dengan pengganti) pada *dataset* yang ada untuk membentuk beberapa *dataset* sampel *bootstrap*. Masing-masing *dataset* sampel *bootstrap* akan digunakan untuk pengembangan model, kemudian kesimpulan performa model diambil dari rata-rata performa model dari pengembangan menggunakan masing-masing *dataset* sampel *bootstrap*. Pengambilan sampel ulang (dengan pengganti) dilakukan pada *training dataset* dan performa pemodelan diambil berdasarkan hasil prediksi model yang dikembangkan untuk memprediksi *testing dataset* asli.

2.6.2 Uji Statistik

Metode uji statistik yang digunakan dalam penelitian ini adalah *paired t-test* untuk melakukan perbandingan *mean difference* pada sepasang target observasi (Hsu & Lachenbruch, 2014). Dalam penelitian ini perbandingan dilakukan dengan tujuan membandingkan performa sepasang model untuk menilai apakah model A lebih baik daripada model B. Dalam masing-masing pengamatan, diasumsikan setiap pasangan tidak bergantung pada pasangan lainnya; pengamatan pasangan dinyatakan sebagai x_i dan y_i; dan d_i adalah perbedaan dari x_i dan y_i. Perhitungan *t-test* dapat dinyatakan sebagai berikut:

$$t = \frac{\bar{d}}{(s_d/\sqrt{n})}, \tag{2.12}$$

dengan $\bar{d}$ adalah rata-rata di; s_d adalah standar deviasi perbedaan; dan n adalah jumlah pasangan.

2.7 Bias Stereotipe

Pada pembuatan model, bias adalah informasi awal milik suatu sistem yang menggambarkan kecerdasan sistem berdasarkan hasil *training*. Data yang dikumpulkan di dunia nyata tidak bersifat homogen karena dihasilkan oleh sub kelompok sosial yang berbeda dengan karakteristiknya masing-masing, dan demografi manusia yang melakukan data labelling juga dapat memengaruhi bias. Model yang dilatih menggunakan data yang memiliki bias berpotensi menyebabkan hasil prediksi yang tidak adil dan tidak akurat (Bansal,

2022). Berdasarkan penelitian yang dilakukan oleh Shah, Schwartz, dan Hovy (2019), ditemukan empat potensi sumber munculnya bias, yaitu:

1. **Bias Label**

   Bias label muncul ketika variabel target pada *dataset* sumber memiliki distribusi yang menyimpang secara substansial dari distribusi ideal. Pada kasus ini, variabel target atau label dianggap salah. Hal ini dapat terjadi karena anotasi dilakukan oleh kelompok anotator yang tidak representatif atau memegang prasangka (Joseph, Friedland, Hobbs, Tsur, & Lazer, 2017). Kemungkinan lain juga bisa dikarenakan anotator tidak memiliki keahlian domain yang cukup (Plank, Hovy, & Søgaard, 2014).

2. **Bias Seleksi**

   Bias seleksi muncul akibat pengamatan yang tidak representatif, di mana model dilatih menggunakan data latih yang tidak merepresentasikan data yang ingin diprediksi. Contohnya adalah ketika model dilatih menggunakan data blog mikro dengan domain politik digunakan untuk mendeteksi ujaran kebencian pada data blog mikro sehari-hari. Misalkan penggunaan kata-kata nama hewan pada *dataset* dengan domain politik pasti berupa umpatan atau digunakan dalam konteks kasar, model akan memiliki performa buruk pada data blog mikro sehari-hari ketika nama hewan dipakai dengan makna aslinya.

3. ***Overamplification***

   Overamplification dapat terjadi di proses pemelajaran ketika model memelajari faktor-faktor kecil yang diamplifikasi menjadi faktor besar dalam pembuatan keputusan. Bias yang disebabkan oleh *overamplification* terjadi tanpa pengaruh anotator, pengumpul data, atau penganalisis data, tetapi bisa mengamplifikasi bias yang sudah ada.

4. **Bias Semantik**

   Di bidang *natural language processing* (NLP), *embeddings* digunakan sebagai sumber informasi semantik. Penggunaan *embeddings* cenderung mengelompokkan kata-kata tertentu bersamaan dengan kata-kata feminin, menyebabkan bias yang tidak diingingkan.

Keberadaan bias yang tidak diingingkan dapat ditemukan dalam data yang mengandung terminologi identitas (Dixon, Li, Sorensen, Thain, & Vasserman, 2018; Mozafari, Farahbakhsh, & Crespi, 2020). Untuk menangani hal ini Dixon et al. melakukan mitigasi dengan cara memperluas data latih dan data *test* menggunakan teknik generalisasi untuk istilah identitas. Pada penelitian yang dilakukan oleh Bansal, salah satu pendekatan penanganan bias yang dilakukan adalah dengan menemukan vektor kata yang menunjukkan bias, vektor kata tersebut kemudian dinetralisir.

Penelitian-penelitian yang sudah dilakukan sebelumnya terkait deteksi ujaran kebencian dan bahasa kasar menggunakan *dataste* berbahasa Indonesia umumnya memiliki fokus kepada pengembangan koleksi data dan pengembangan model klasifikasi. Isu bias terkait deteksi ujaran kebencian dan bahasa kasar belum pernah dibahas untuk *dataset* berbahasa Indonesia. Oleh karena itu, penelitian ini akan melakukan kajian terkait bias pada pendeteksian ujaran kebencian dan bahasa kasar menggunakan *dataste* berbahasa Indonesia.

BAB 3
METODOLOGI

Bab 3 mendiskusikan langkah-langkah yang dilakukan dalam penelitian ini terkait penyelesaian masalah. Pertama, Subbab 3.1 membahas dataset yang digunakan dalam penelitian ini. Sebagai dasar dari langkah-langkah berikutnya, proses dan hasil EDA dibahas di Subbab 3.2. Kemudian, Subbab 3.3 membahas proses feature selection menggunakan metode pemilihan fitur ANOVA f-test yang hasilnya ditelusuri lebih lanjut menggunakan LRA. Visualisasi effect size menggunakan SHAP dan pemetaan penggunaan fitur secara bertambah terhadap akurasi model juga dibahas di Subbab 3.3. Subbab 3.4 membahas proses pengembangan model dan Subbab 3.5 membahas metode evaluasi yang digunakan untuk melakukan uji statistik dan perbandingan dari model-model yang telah dikembangkan. Selanjutnya, Subbab 3.6 membahas proses pengembangan dataset deteksi bias berupa *dataset* sintetis hasil manipulasi data *test* dan *dataset* yang dikembangkan secara manual. Terakhir, Subbab 3.7 membahas penggunaan dataset deteksi bias untuk diprediksi menggunakan model-model yang sudah dikembangkan di tahap sebelumnya.

3.1 *Dataset*

Dalam penelitian ini, pengembangan dan evaluasi model dilakukan berdasarkan *dataset* yang dikembangkan oleh Ibrohim dan Budi (2019). *Dataset* mengandung sebanyak 13169 pesan blog mikro dengan domain politik berdasarkan waktu dikembangkannya, yaitu pada tahun 2019 di masa-masa menjelang pemilihan umum presiden Indonesia. Data memiliki dua label utama, "*Hate Speech*" dan "*Abusive*". Selain kedua label tersebut, Ibrohim dan Budi juga membagi ujaran kebencian ke dalam tiga subkategori berupa kategori konten kebencian, target kebencian, dan tingkat keparahan dari ujaran kebencian. Lima kategori konten berupa: *religion/creed* (agama/kepercayaan), *race/ethnicity* (ras/etnis), *physical/disability* (fisik/disabilitas), *gender/sexual orientation* (jenis kelamin/orientasi seksual), *other invective/slander* (makian lainnya). Kategori target ujaran kebencian memiliki pilihan berupa: *individual* atau *group*. Terakhir kategori tingkat keparahan berupa: *weak* (lemah), *moderate* (sedang), dan *strong* (kuat). Penelitian ini hanya fokus

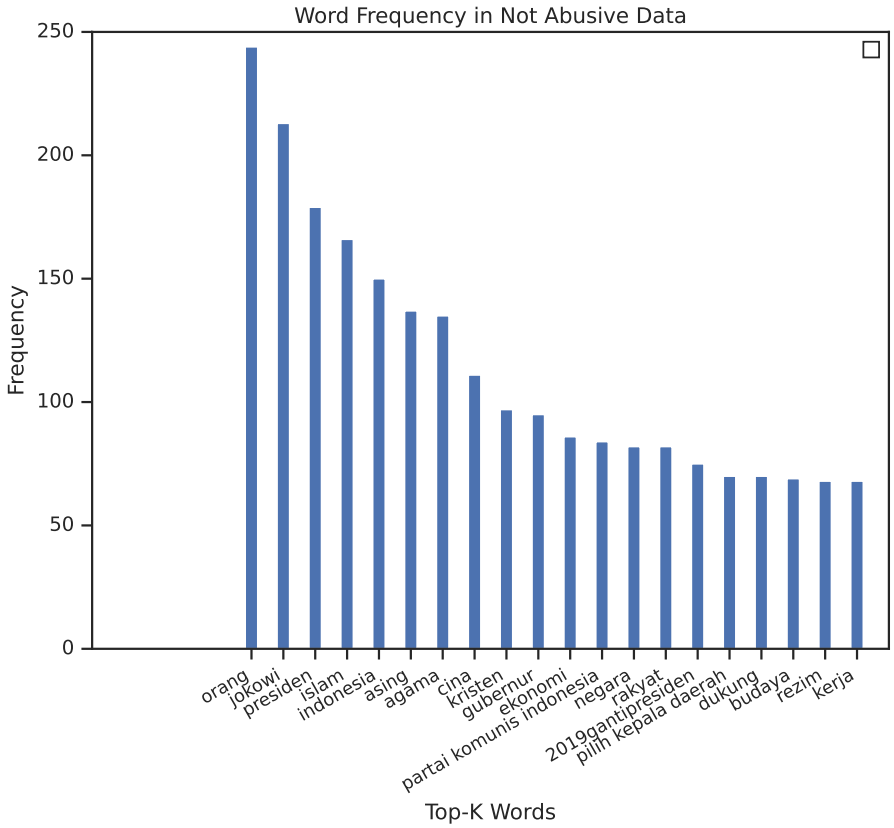
terhadap dua label utama: ”*Hate Speech*” dan ”*Abusive*”, yang masing-masing akan dideteksi secara biner.

3.2 *Exploratory Data Analysis*

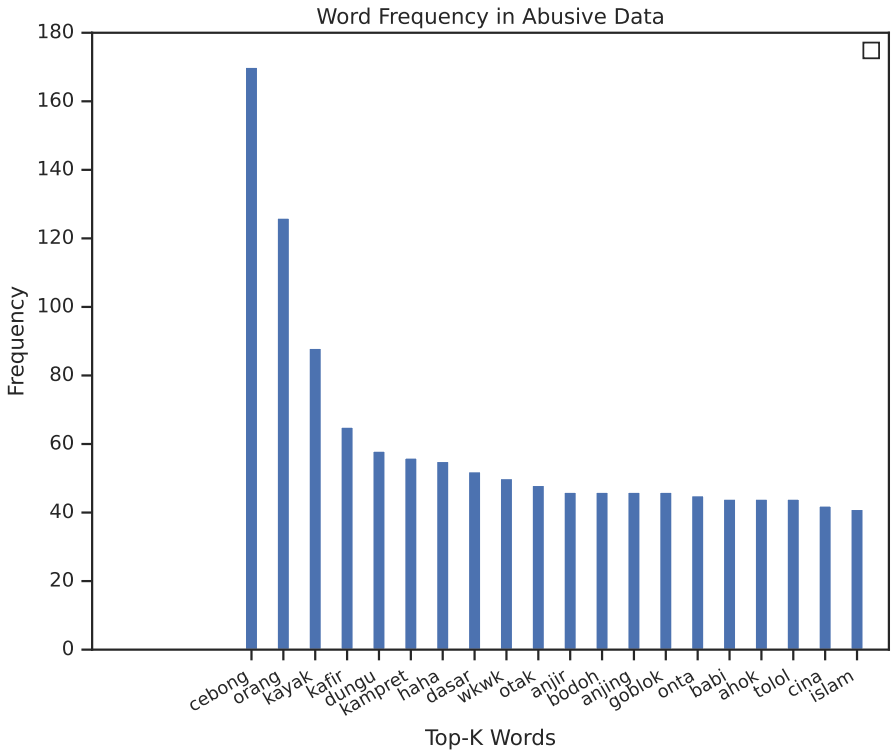
Pada awal tahap pra-proses, dilakukan *stratified sampling* untuk mendapatkan data sampel sebesar 3292 baris data yang akan digunakan untuk proses *exploratory data analysis* (EDA). Pada proses EDA, dilakukan analisis terhadap frekuensi kemunculan kata dan struktur leksikal dari pesan-pesan blog mikro berdasarkan kedua label terkait ujaran kebencian dan bahasa kasar. Pertama, untuk melakukan analisis terkait frekuensi kemunculan kata-kata yang bukan merupakan *stopwords*, dilakukan visualisasi menggunakan diagram balok untuk 20 kata dengan frekuensi kemunculan tertinggi berdasarkan label ”*Not Abusive*”, ”*Not Hate Speech*”, ”*Abusive*”, dan ”*Hate Speech*”. Secara lebih lengkap, dicatat 100 kata-kata dengan frekuensi tertinggi untuk masing-masing label dalam bentuk tabel yang dapat dilihat di Lampiran 1.

Gambar 3.1 menunjukkan urutan kata-kata dengan frekuensi tertinggi pada data berlabel ”*Not Abusive*”, didapatkan beberapa kata yang paling sering muncul adalah kata ”orang”, ”jokowi”, dan ”presiden”. Gambar 3.3 menunjukkan urutan kata-kata dengan frekuensi tertinggi pada data berlabel ”*Not Hate Speech*”, didapatkan beberapa kata yang paling sering muncul adalah ”orang”, ”presiden”, dan ”asing”. Dari kedua gambar tersebut, bisa dilihat bahwa kata ”orang” dan ”presiden” sering muncul untuk data bukan merupakan ujaran kebencian atau bahasa kasar.

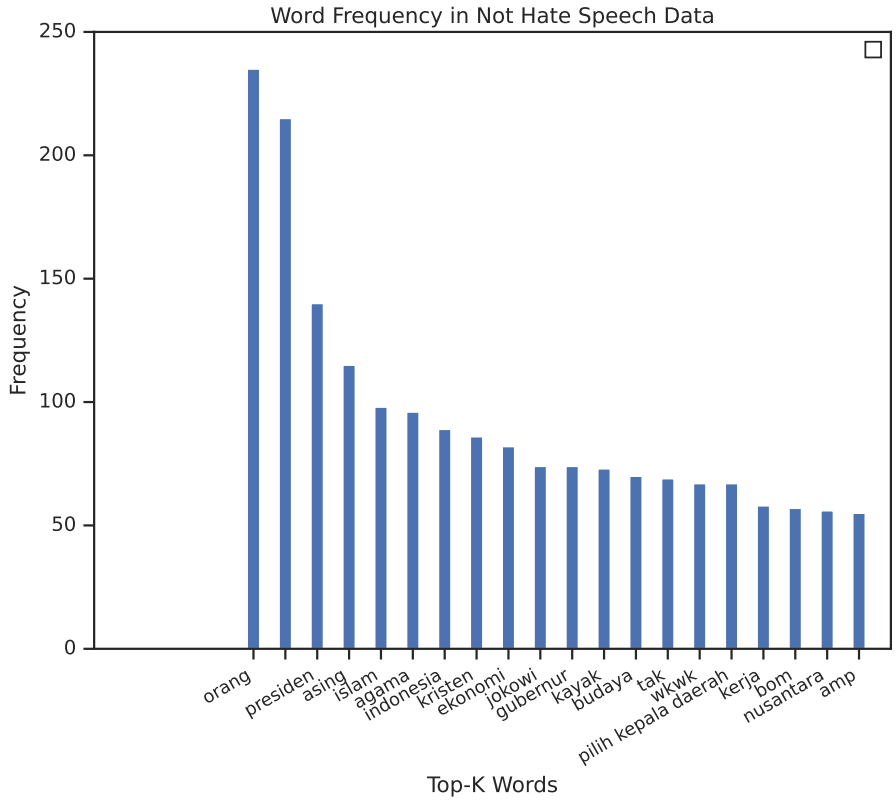
Gambar 3.2 menunjukkan urutan kata-kata dengan frekuensi tertinggi pada data berlabel ”*Abusive*”, didapatkan beberapa kata yang paling sering muncul dan menunjukkan sifat kasar adalah ”cebong”, ”kafir”, ”dungu”, dan ”kampret”. Gambar 3.4 menunjukkan urutan kata-kata dengan frekuensi tertinggi pada data berlabel ”*Hate Speech*”, dan didapatkan kata-kata yang bersifat kasar seperti yang ditemukan untuk data bahasa kasar, yaitu ”cebong”, ”kafir”, dan ”dungu”. Namun, ditemukan juga kata-kata yang tidak secara sendiri tidak memiliki unsur kebencian atau bersifat kasar seperti ”jokowi”, ”islam”, ”cina”, dan ”indonesia”.



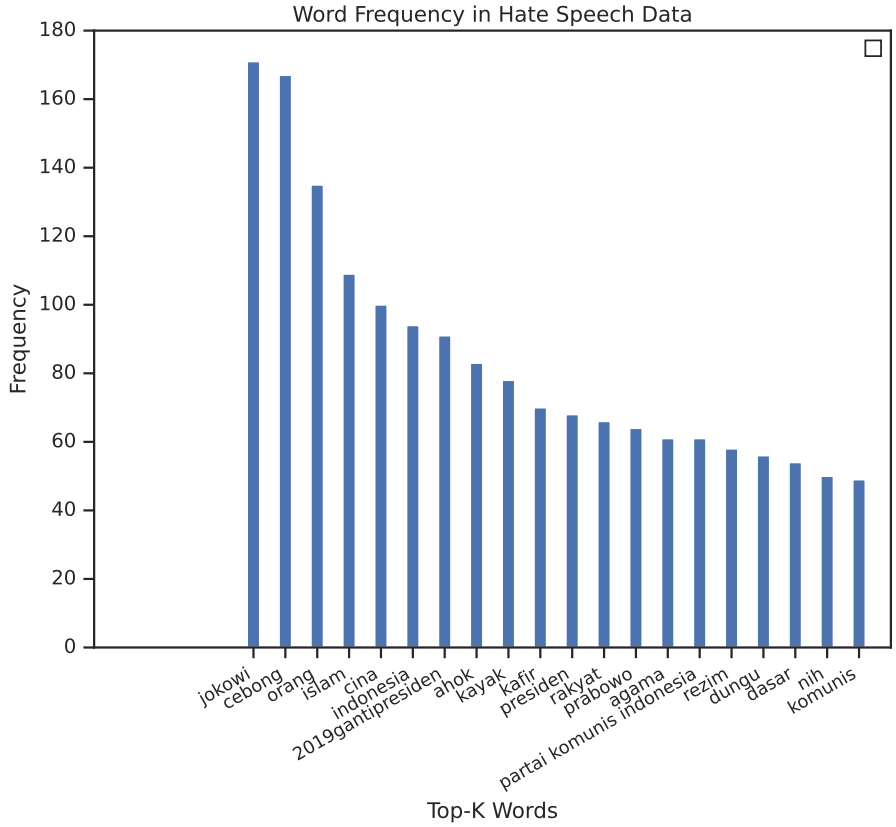
Gambar 3.1: Kata dengan Frekuensi Tertinggi pada Data Berlabel *Not Abusive*



Gambar 3.2: Top 20 kata dengan frekuensi tertingga pada data berlabel "*Abusive*"



Gambar 3.3: Top 20 kata dengan frekuensi tertingga pada data berlabel "*Not Hate Speech*"

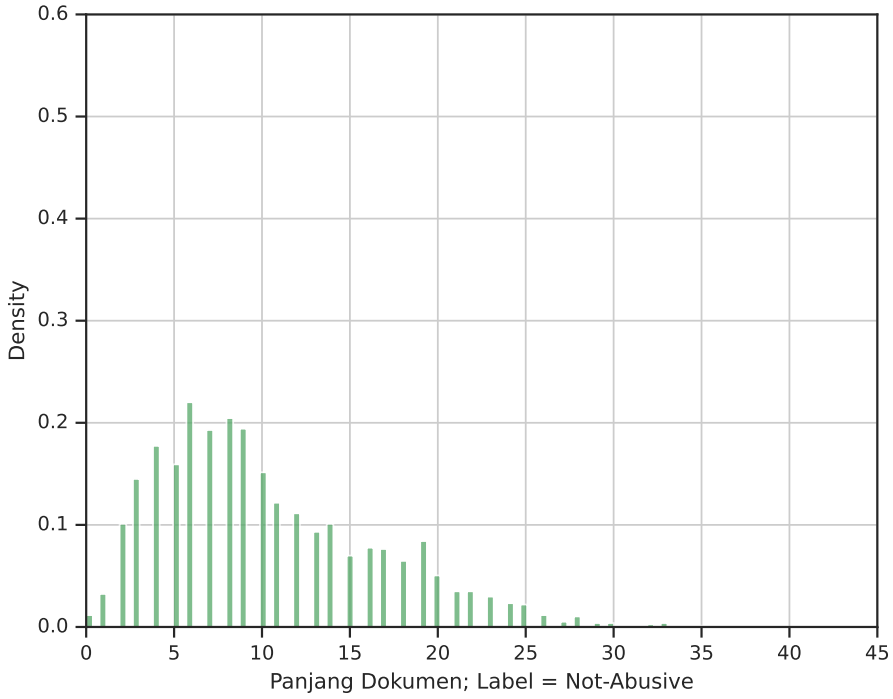


Gambar 3.4: Top 20 kata dengan frekuensi tertingga pada data berlabel "*Hate Speech*"

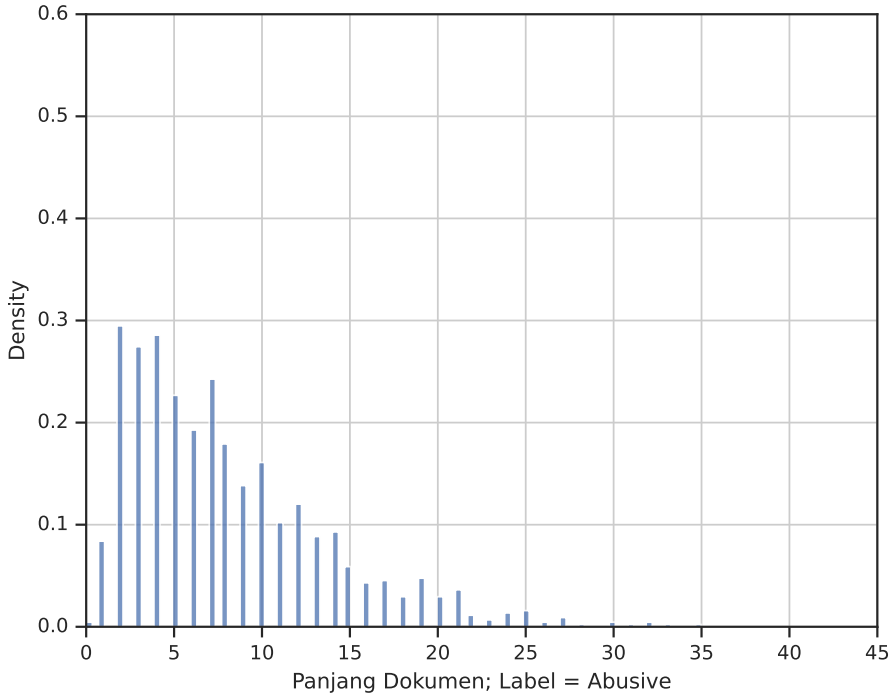
Kemudian, analisis distribusi panjang dokumen dari masing-masing label dilakukan menggunakan visualisasi diagram balok yang memetakan kepadatan dari jumlah kata (tanpa *stopwords*) pada pesan. Perlu dicatat bahwa panjang dokumen dihitung berdasarkan jumlah kata yang bukan merupakan *stopwords*. Distribusi panjang dari data yang berlabel "*Not Abusive*" ditampilkan pada Gambar 3.5. Dapat dilihat bahwa pada data yang berlabel "*Not Abusive*", terdapat lebih banyak pesan dengan panjang sekitar lima sampai sepuluh kata. Sedangkan untuk data yang berlabel "*Abusive*", dapat dilihat pada Gambar 3.6 bahwa terdapat lebih banyak pesan dengan panjang sekitar satu sampai tujuh kata. Kedua grafik menunjukkan *right skewness*, tetapi Gambar 3.5 menunjukkan grafik yang lebih tersebar dibandingkan grafik pada Gambar 3.6 yang cenderung lebih padat di sisi kiri. Hasil ini menunjukkan bahwa data berlabel "*Abusive*" memiliki panjang pesan yang cenderung lebih pendek dibandingkan data yang berlabel "*Not Abusive*".

Distribusi panjang dari data yang berlabel "*Not Hate Speech*" ditampilkan pada Gambar 3.7. Pada data yang berlabel "*Not Hate Speech*", terdapat lebih banyak pesan dengan panjang sekitar dua sampai sepuluh kata. Selain itu, dapat dilihat juga bahwa panjang pesan memiliki distribusi yang cukup luas walaupun masih menunjukkan *right skewness*. Distribusi panjang dari data yang berlabel "*Hate Speech*" ditampilkan pada Gambar 3.8. Pada data yang berlabel "*Hate Speech*", terdapat lebih banyak juga pesan dengan panjang sekitar dua sampai sepuluh kata. Namun, panjang pesan memiliki distribusi yang lebih padat di kiri, walaupun masih menunjukkan *right skewness*. Oleh karena itu, data berlabel "*Hate Speech*" dapat disimpulkan memiliki panjang pesan yang cenderung lebih pendek dibandingkan data yang berlabel "*Not Hate Speech*".

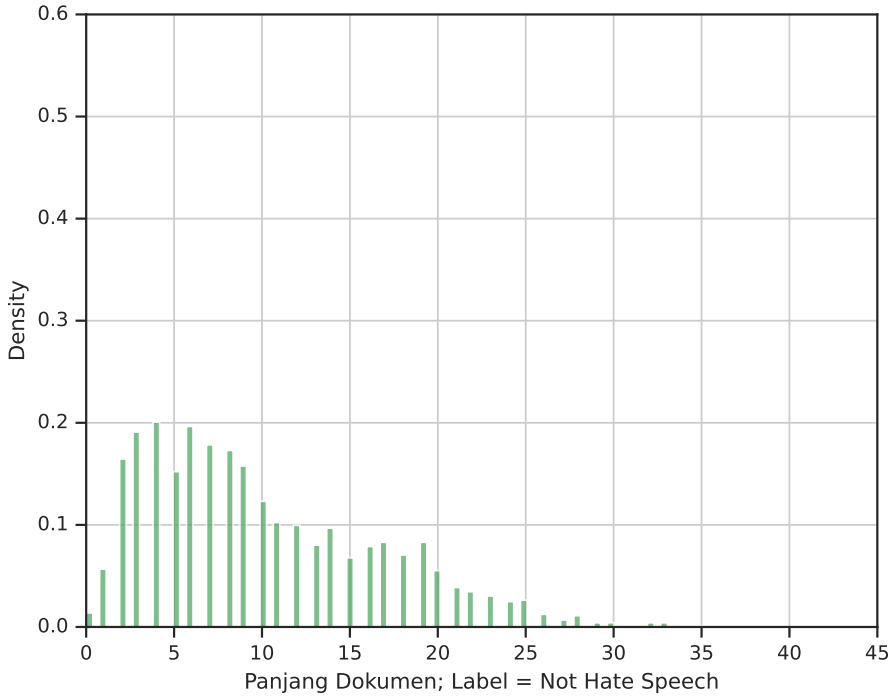
Berdasarkan eksplorasi yang telah dilakukan, diusulkan fitur-fitur tekstual terkait keberadaan kata dan struktur pesan. Kumpulan kata-kata unik yang terdiri dari 300 kata-kata yang paling sering muncul secara keseluruhan; kata-kata yang ada di dalam *abusive lexicon*; dan 19 kata-kata yang memiliki potensi bias, masing-masing diusulkan sebagai fitur terkait keberadaan kata. Selanjutnya, diusulkan fitur-fitur terkait struktur pesan berupa panjang pesan berdasarkan banyak kata dan panjang pesan secara keseluruhan berdasarkan banyak karakter. Selain itu, diusulkan juga fitur-fitur berupa jumlah huruf kapital pada pesan; penanda jika pesan seluruhnya dituliskan menggunakan huruf kapital; penanda keberadaan tanda seru pada pesan; dan penanda keberadaan kata yang ada di *abusive lexicon* (tidak spesifik kata apa seperti fitur sebelumnya). Secara total, tahap ini menghasilkan 389 usulan fitur yang akan digunakan untuk tahap-tahap selanjutnya.



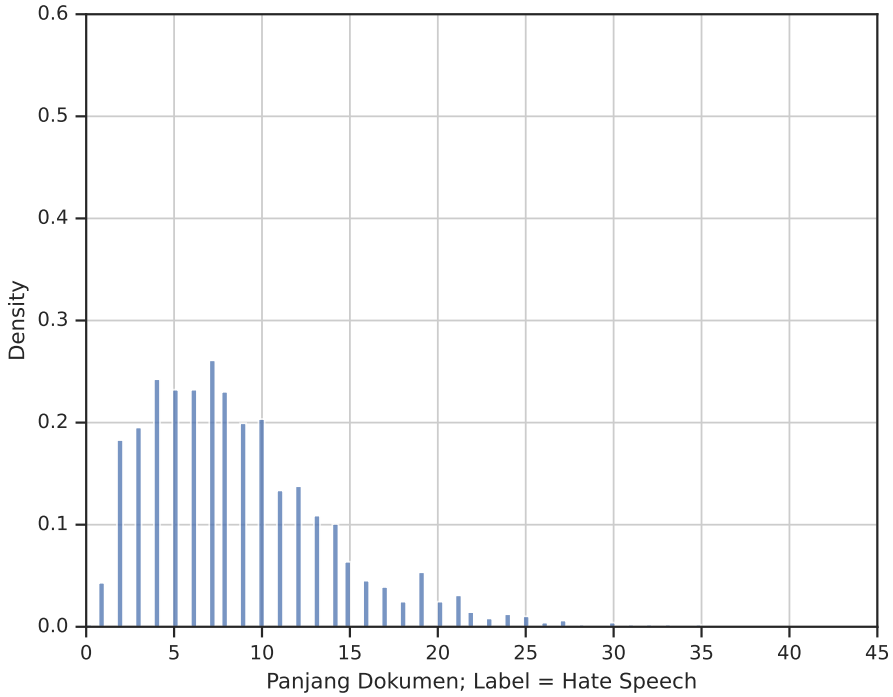
Gambar 3.5: Distribusi panjang data berlabel *Not Abusive*



Gambar 3.6: Distribusi panjang data berlabel *Abusive*



Gambar 3.7: Distribusi panjang data berlabel *Not Hate Speech*



Gambar 3.8: Distribusi panjang data berlabel *Hate Speech*

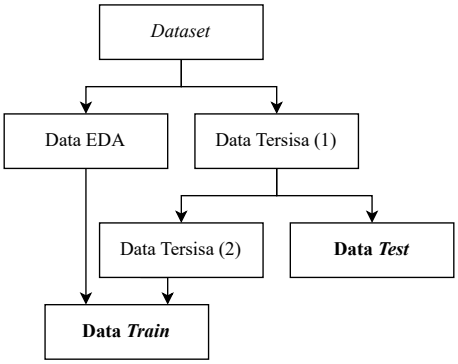
3.3 Feature Selection

Pada tahap EDA, diusulkan sebanyak 389 fitur untuk diekstraksi. Fitur-fitur tersebut merupakan kumpulan dari fitur ortografi dan fitur leksikal, terutama berdasarkan keberadaan suatu kata. Fitur ortografi diusulkan karena seringkali teks yang menunjukkan emosi negatif mengandung tanda baca tertentu dan kapitalisasi berlebih, seperti penulisan yang sepenuhnya menggunakan huruf kapital dan adanya tanda seru untuk menyampaikan emosi marah dari penulis. Fitur leksikal seperti panjang dokumen berdasarkan banyak kata dan panjang dokumen berdasarkan banyak karakter digunakan karena dalam tahap EDA terlihat bahwa data dengan label *Abusive* dan data dengan label *Hate Speech* cenderung memiliki panjang dokumen yang lebih pendek. Fitur keberadaan kata diusulkan salah satunya karena ditemukan banyak kata-kata yang secara sendiri seharusnya bersifat netral tapi muncul sebagai salah satu *top-k* kata dengan frekuensi tertinggi pada dokumen-dokumen berlabel "*Hate Speech*" tapi "*Not Abusive*". Tujuannya adalah untuk melihat seberapa berpengaruh keberadaan kata tersebut dalam pembuatan keputusan label pada dokumen.

Selanjutnya, *ANOVA f-test* digunakan untuk memberi nilai pada masing-masing fitur dan membuat *list* fitur yang terurut berdasarkan *f score* terbesar hingga terkecil. Setelah didapat urutan fitur, digunakan SHAP dan *Logistic Regression Analysis* sebagai metode menjelaskan *effect size* dari 30 fitur dengan *f score* terbesar. Dalam penggunaan SHAP, analisis *effect size* top 30 fitur dilakukan dalam pengembangan model-model berbeda: *Logistic Regression*, *Extreme Gradient Boosting*, *CatBoost*, *Multi-layer Perceptron*, dan *Support Vector Machine*. Perlu ditekankan bahwa penggunaan LRA dan SHAP bukan untuk seleksi fitur, tetapi untuk mengkaji aspek kontribusi dari fitur dan hanya dilakukan terhadap 30 fitur dengan *f score* terbesar. Kemudian, model-model tersebut juga digunakan untuk mencari jumlah optimal fitur dengan cara memetakan penggunaan fitur dengan jumlah yang meningkat terhadap akurasi model. Semua proses ini dilakukan dua kali untuk kedua tugas klasifikasi, yaitu kalsifikasi biner untuk label '*Hate Speech*'/'*Not Hate Speech*' dan '*Abusive*'/'*Not Abusive*'.

3.4 Perancangan Model

Langkah pertama yang dilakukan sebelum memulai pemodelan adalah membagi data menjadi data *training* dan data *test* secara *stratified*. Untuk memastikan kredibilitas



Gambar 3.9: Pembagian data *train-test*

data *test*, data *test* diambil dari data yang tersisa pada proses sampling untuk data EDA. Dapat dilihat di Gambar 3.9, data training terdiri dari data yang tersisa dari kedua proses sampling digabungkan dengan data yang sebelumnya dipakai dalam proses EDA. Berdasarkan peringkat fitur dari ANOVA f-*test* dan proses pemetaan fitur dengan jumlah yang meningkat terhadap akurasi model, didapatkan hasil akurasi yang konsisten meningkat pada semua model ketika digunakan sampai 360 fitur dengan *f score* terbesar. Sehingga, pengembangan model untuk tugas klasifikasi biner dilakukan dengan menggunakan 360 fitur tersebut untuk masing-masing label '*Hate Speech*'/'*Not Hate Speech*' dan '*Abusive*'/'*Not Abusive*'. Untuk mendapatkan gambaran umum dari performa model, dilakukan *training* menggunakan seluruh data latih, lalu akurasi masing-masing model dinilai menggunakan data *test*. Seperti yang telah dinyatakan di Subbab 3.3, penelitian ini menggunakan lima model pemelajaran mesin berupa: *Logistic Regression*, *Extreme Gradient Boosting*, *CatBoost*, *Multi-layer Perceptron*, dan *Support Vector Machine*. Penelitian ini mengedepankan isu *explainability* model pemelajaran mesin, sehingga tidak menggunakan model pemelajaran mesin *state of the art* seperti BERT (*Bidirectional Encoder Representations from Transformerst*) yang bersifat *black-box*.

3.5 Evaluasi Model

Untuk memastikan hasil performa yang didapat dapat dipercaya, dilakukan beberapa metode evaluasi. Pertama, metode *bootstrap with replacement* digunakan pada data latih untuk mendapatkan data sampel *bootstrap* yang berjumlah sebanyak data latih asli. Data sampel *bootstrap* digunakan dalam pengembangan lima model menggunakan kelima algoritma model yang dipakai di tahap sebelumnya. Kemudian, data *test* digunakan untuk menilai performa seluruh model hasil pelatihan menggunakan data sampel *bootstrap*.

Proses ini dilakukan sebanyak 30 kali dan rata-rata akurasi masing-masing model dari seluruh iterasi adalah yang dijadikan sebagai acuan performa suatu model terhadap model lainnya. Selain bootstrap, dilakukan juga uji statistik menggunakan metode *paired t-test* untuk membandingkan apakah suatu model terbukti memiliki performa lebih baik dari model lainnya.

3.6 Pengembangan Dataset untuk Deteksi Bias

Dalam pengembangan model, keberadaan bias berpotensi menyebabkan hasil prediksi yang tidak adil dan tidak akurat (Bansal, 2022). Oleh karena itu, diperlukan kajian terkait bias untuk memastikan model dapat membuat keputusan secara adil dan akurat; tidak memiliki akurasi yang tinggi hanya di *dataset* tertentu karena pengaruh bias terhadap pelatihan model. Pada penelitian ini, diusulkan sebuah metode pendeteksian bias berupa pengembangan *dataset* untuk perbandingan performa model. Penelitian yang dilakukan oleh Dixon et al. (2018) melakukan generalisasi istilah identitas pada data yang digunakan di penelitian tersebut. Hal serupa dilakukan di penelitian ini untuk melakukan generalisasi pada data *test*, dengan mengganti kata-kata yang diidentifikasi memiliki potensi bias menjadi kata-kata lain yang tidak berpotensi bias.

3.6.1 Dataset Sintetis

Sebelum melakukan pengembangan *dataset* sintetis, perlu dilakukan identifikasi kata-kata yang memiliki potensi bias. Identifikasi kata-kata ini dilakukan menggunakan hasil EDA dari data sampel yang diambil secara *stratified* untuk mendapatkan *top-100* kata dengan frekuensi kemunculan tertinggi, tidak termasuk stopwords, pada data yang sudah difilter memiliki label *Hate Speech* dan *Not Abusive*. Dari *top-100* kata tersebut, diusulkan beberapa syarat untuk memutuskan kata-kata mana saja yang dapat disebut potensi bias. Syarat potensi bias adalah kata yang merupakan: nama orang, jabatan, organisasi, suku, agama, ras atau etnis, dan negara.

Selain itu, diterapkan pengecualian untuk memutuskan kata-kata yang tidak bisa disebut potensi bias. Kata-kata yang tidak dianggap sebagai potensi bias berdasarkan pengecualian adalah:

1. kata sifat karena mengandung kata-kata yang maknanya memang negatif seperti ”diktator” dan ”teroris”;

2. kata kerja karena mengandung kata-kata yang maknanya memang negatif seperti ”korupsi” dan ”bunuh”;

3. kata-kata yang mereferensi ideologi komunisme, karena di Indonesia penggunaan kata-kata tersebut relatif memiliki konotasi negatif, seperti ”pki” dan ”komunis”;

4. kata pribumi karena penggunaannya relatif untuk diskriminasi ras;

5. kata saracen karena di Indonesia merupakan nama sebuah website yang menyediakan konten kebencian/diskriminatif/rasis;

6. kata benda di luar syarat yang diusulkan karena bisa memiliki makna yang memang negatif seperti ”rezim” dan ”antek”.

Proses ini menghasilkan 19 usulan kata-kata potensi bias. Dataset sintetis dikembangkan dengan cara membuat 5 salinan dari setiap pesan blog mikro yang ada pada data *test*. Untuk masing-masing salinan, apabila ada kata potensi bias pada pesan, dilakukan penggantian kata potensi bias menjadi salah satu kata pengganti berdasarkan Tabel 3.1, sehingga didapat 5 versi dari pesan yang tidak mengandung kata bias. Untuk kasus di mana tidak ada kata bias dalam pesan, disimpan 5 salinan pesan tanpa perubahan. Untuk kasus di mana ada lebih dari satu kata bias, dilakukan pengacakan kombinasi kata pengganti dan tetap didapat 5 versi berbeda dari pesan yang tidak mengandung kata bias.

Tabel 3.1: Tabel kata-kata potensi bias dengan usulan kata pengganti

Kata Potensi Bias	Pengganti 1	Pengganti 2	Pengganti 3	Pengganti 4	Pengganti 5
jokowi	soekarno	habibi	kelvin	arie	nathan
islam	bahai	shinto	tenrikyo	animisme	shamanisme
indonesia	jepang	korea	india	australia	perancis
prabowo	arthur	filbert	dimas	khariz	gabe
cina	vietnam	rusia	meksiko	amerika	inggris
ahok	doni	baraz	usman	zaidan	adi
presiden	menteri	pemimpin	bupati	walikota	pegawai
kpk	ky	kpu	polri	bi	tni
gubernur	menteri	pemimpin	bupati	walikota	pegawai
sby	soeharto	megawati	gusdur	angga	sean
anies	adit	fathan	sulthan	anto	reza
dpr	mpr	dpd	mk	ma	bpk
muslim	agnostik	jainisme	sikh	mormon	henoteis
sandi	farrel	zidan	yazid	rifqy	ditya
anies-sandi	soekarno-hatta	hamzah-agum	amien-siswono	wiranto-salahuddin	megawati-hasyim
rohingya	melayu	mulato	mestis	serani	tionghoa
kristen	helenisme	neopaganisme	mahayana	taoisme	politeisme
buddha	deisme	zoroaster	shinto	konfusianisme	chondoisme
myanmar	filipina	afganistan	brasil	jerman	malaysia

3.6.2 Dataset yang Dikembangkan secara Manual

Dengan tujuan mencari bentuk pesan yang terpengaruh bias ketika diprediksi dan seberapa berpengaruh kata-kata berpotensi bias pada pesan-pesan yang bukan merupakan ujaran kebencian dan bukan bahasa kasar, dikembangkan sebuah *dataset* yang terdiri dari 100 pesan. Berbeda dengan data nyata yang telah dikembangkan oleh Ibrohim dan Budi (2019), di mana kata-kata dengan potensi bias dapat ditemukan dalam pesan yang merupakan ujaran kebencian dan bahasa kasar, seluruh pesan pada *dataset* ini tidak memiliki unsur kebencian atau mengandung kata-kata kasar dan diberi label ”*Not Hate Speech*” dan ”*Not Abusive*”. Beberapa contoh kalimat yang ada pada *dataset* ini adalah sebagai berikut:

1. ”GO JOKOWI GO!!”
2. ”Foto Prabowo sama Bobby lucu banget!!!”
3. ”Pak Jokowi keren bgt di Asian Games”

Pada *dataset* yang dikembangkan secara manual, terdapat kalimat seperti Contoh 1 yang menggunakan kata potensi bias ”jokowi” bersamaan dengan huruf kapital, tanda seru, dan memiliki panjang pesan yang pendek. Contoh 2 menggunakan kata potensi bias ”prabowo” dalam konteks positif dan menunjukkan kegembiraan. Contoh 3 menggunakan kata potensi bias ”jokowi” dalam kalimat yang tidak memiliki tanda seru dan memiliki panjang pesan yang tidak panjang tapi tidak pendek. Kalimat-kalimat dibuat berbeda untuk merepresentasikan pengaruh fitur lain dalam akurasi prediksi pesan yang mengandung kata-kata potensi bias. Kalimat-kalimat lain pada *dataset* yang dikembangkan secara manual dapat dilihat secara lengkap di Lampiran 6. Untuk melihat apakah *dataset* buatan dapat mendeteksi bias atau tidak, dilakukan proses yang sama dengan pengembangan *dataset* sintesis untuk mendapatkan *dataset* yang dikembangkan secara manual yang tidak mengandung kata bias.

3.7 Evaluasi Model Menggunakan Dataset Deteksi Bias

Keberadaan bias dideteksi berdasarkan penurunan atau peningkatan akurasi model ketika digunakan untuk memprediksi data *test* dibandingkan dengan versi sintetis dari data *test*.

Perbandingan juga dilakukan pada akurasi hasil prediksi *dataset* yang dikembangkan secara manual dengan versi sintetisnya. Apabila model mempelajari kata-kata potensi bias sebagai faktor yang memengaruhi keputusan padahal seharusnya kata-kata tersebut bersifat netral, diasumsikan akan terjadi perubahan berupa penurunan atau peningkatan pada hasil akurasi; membandingkan data *test* dengan data *test* sintetis, dan *dataset* yang dikembangkan secara manual dengan versi sintetisnya. Perbandingan ini dapat dilakukan karena *dataset* sintesis berukuran tepat lima kali dari ukuran *dataset* sumbernya, lalu masing-masing pesan direpresentasikan dengan lima salinan yang tidak mengandung kata potensi bias, termasuk pesan yang aslinya tidak mengandung kata potensi bias.

BAB 4
PERANCANGAN DAN IMPLEMENTASI

Bab 4 mendiskusikan perancangan dan implementasi yang dilakukan untuk melakukan eksperimen berdasarkan metodologi yang di bahas di Bab 3. Pertama, Subbab 4.1 membahas kode untuk melakukan *feature selection* menggunakan metode *ANOVA f-test*, LRA, dan SHAP. Subbab 4.2 membahas kode untuk melakukan ekstraksi fitur secara otomatis. Kemudian, Subbab 4.3 membahas kode untuk mengembangkan *dataset* deteksi bias menggunakan *dataset* sumbernya. Terakhir, Subbab 4.4 membahas kode pengembangan model dan evaluasi performa model.

4.1 Feature Selection

Feature selection dilakukan menggunakan modul-modul *SelectKBest* dan *f_classif* dari *sklearn* untuk *ANOVA f-test*, *statsmodels* untuk LRA, dan *tools* SHAP untuk *feature importance* dan *summary plot* berdasarkan nilai *SHAPley*.

```
"""
Original file is located at
    https://colab.research.google.com/drive/1klYJVNuyzf8JoEvfpO3lSDx4mrEeVHg8 (Abusive)
    https://colab.research.google.com/drive/1z3pswYfAMDQrRcfE0cLtHBLX3nQGGeVJ (Hate
    Speech)
"""
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import SelectKBest, f_classif

data = df.values
X = data[:, :-1]
y = data[:,-1]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33,
    random_state=1)

def select_features(X_train, y_train, X_test):
    fs = SelectKBest(score_func=f_classif, k='all')
    fs.fit(X_train, y_train)
    X_train_fs = fs.transform(X_train)
    X_test_fs = fs.transform(X_test)
    return X_train_fs, X_test_fs, fs

```

```
X_train_fs, X_test_fs, fs = select_features(X_train, y_train, X_test)
# f score
fs_list = []
for i in range(len(fs.scores_)):
    print('Feature %s: %f' % (features_df.columns[i], fs.scores_[i]))
    if fs.scores_[i] > 0:
        fs_list.append((features_df.columns[i], fs.scores_[i]))

fs_list.sort(key = lambda i:i[1], reverse = True)
top_30_features = [fs_list[i] for i in range(30)]

"""# LRA"""
import statsmodels.formula.api as smf
lra_model = smf.logit("""
    y ~          # nama target <- 'HS' atau 'A'
          X      # nama fitur
""", data = features_df)
results = lra_model.fit()
results.summary()

"""# SHAP"""
!pip install --quiet shap

f30 = [i for i, j in top_30_features]
top30X = df[df.columns.intersection(f30)]

X_train, X_test, y_train, y_test = train_test_split(top30X, y, test_size = 0.2)
model = Model().fit(X_train, y_train)
explainer = shap.Explainer(model.predict, X_test)
shap_values = explainer(X_test)

shap.plots.bar(shap_values)
shap.summary_plot(shap_values)
```

Kode 4.1: *Feature selection* pada 389 fitur yang diusulkan

4.2 Ekstraksi Fitur

Ektraksi fitur dilakukan dengan cara melakukan iterasi pada data, untuk masing-masing baris pada data dilakukan tokenisasi dari pesan menjadi sebuah *list* kata-kata. *List* tersebut kemudian diiterasi untuk mendapatkan keberadaan tanda seru (fitur *Exclamation*) dan memeriksa keberadaan kata yang merupakan bagian dari kata-kata potensi bias, kata-kata yang paling sering muncul, dan kata-kata *abusive lexicon* yang telah dikembangkan pada penelitian yang dilakukan oleh Ibrohim dan Budi (2019) (fitur *Abusive*). Kemudian, dihitung dan disimpan panjang dari pesan (fitur *Length*), banyak kata dalam pesan (fitur

Words), banyak huruf kapital dalam pesan (fitur *Uppercase*), dan diperiksa apakah pesan sepenuhnya menggunakan huruf kapital (fitur *All Caps*).

```
"""
Original file is located at

    https://colab.research.google.com/drive/1nzoA8k1cy-cy--oXKPCvDM0w2UxwGCSp#scrollTo=sGxY7ihCQexO
"""
for i, row in df.iterrows():
    words = remove_stops_and_lemmatize(tokenize(row["Tweet"]))
    ex_words = remove_stops_and_lemmatize_keep_exclamation(tokenize(row["Tweet"]))

    for xword in ex_words:
        if xword == "!":
            df.loc[i, 'Exclamation'] = 1
            break

    for word in words:
        if word.lower() in abusive_lexicon.values:
            df.loc[i, 'Abusive'] = 1

    length = len(row["Tweet"])
    df.loc[i, 'Length'] = length

    word_count = len(words)
    df.loc[i, 'Words'] = word_count

    uppercase = sum(1 for c in row["Tweet"] if c.isupper())
    df.loc[i, 'Uppercase'] = uppercase
    df.loc[i, 'All Caps'] = 1 if row["Tweet"].isupper() else 0

    tokens = tokenize(row["Tweet"])
    for token in words:
            token = token.lower()
            if token in potensi_bias.values:
                df.loc[i, token] = 1
            if token in abusive_lexicon["ABUSIVE"].values:
                df.loc[i, token] = 1
            if token in most_freq["0"].values:
                df.loc[i, token] = 1
```

Kode 4.2: Ekstraksi fitur untuk 389 fitur yang diusulkan

4.3 Pengembangan *Dataset* untuk Deteksi Bias

Dataset sintetis untuk deteksi bias dikembangkan menggunakan sebuah program dengan bahasa pemrograman Python. Input program adalah sebuah *dataset* yang ingin dilakukan

penggantian kata-kata potensi bias menjadi kata pengganti dari pilihan yang ditampilkan di Tabel 3.1. Cara kerja program adalah dengan melakukan iterasi pada *dataset* input, untuk masing-masing baris dilakukan iterasi untuk mengecek kata potensi bias apa saja yang ada di dalam pesan. Kemudian, dibuat 5 salinan dari pesan, lalu dilakukan iterasi pada kata-kata potensi bias yang terdeteksi untuk mengganti kata potensi bias pada salinan pesan dengan kata pengganti dalam urutan yang diacak. Terakhir, kelima salinan yang sudah tidak mengandung kata-kata potensi bias ditambahkan ke dalam *dataset output*.

```
"""
Original file is located at

    https://colab.research.google.com/drive/1nzoA8k1cy-cy--oXKPCvDM0w2UxwGCSp#scrollTo=Q-qOXXccQTCl
"""

import random
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)

for i, row in custom_data.iterrows():
  new_row = row
  words = remove_stops_and_lemmatize(tokenize(row["Tweet"]))
  bias_list = []
  pool = ['r1', 'r2', 'r3', 'r4', 'r5']

  sentence = replace_bias_alay(row.Tweet)

  tweet_1 = sentence.lower()
  tweet_2 = sentence.lower()
  tweet_3 = sentence.lower()
  tweet_4 = sentence.lower()
  tweet_5 = sentence.lower()

  copy_row_1 = row.copy()
  copy_row_2 = row.copy()
  copy_row_3 = row.copy()
  copy_row_4 = row.copy()
  copy_row_5 = row.copy()

  for word in words:
    if word.lower() in replacement_data["0"].values:
      bias_list.append(word)
    elif word.lower() == "komisi berantas korupsi":
      bias_list.append("kpk")
    elif word.lower() == "susilo bambang yudhoyono":
      bias_list.append("sby")
```

```
    elif word.lower() == "dewan wakil rakyat":
      bias_list.append("dpr")

  for bias in bias_list:
    bias_row = replacement_data[replacement_data['0'] == bias.lower()]
    random.shuffle(pool)

    tweet_1 = tweet_1.replace(bias, str(bias_row[pool[0]].values[0]))
    tweet_2 = tweet_2.replace(bias, str(bias_row[pool[1]].values[0]))
    tweet_3 = tweet_3.replace(bias, str(bias_row[pool[2]].values[0]))
    tweet_4 = tweet_4.replace(bias, str(bias_row[pool[3]].values[0]))
    tweet_5 = tweet_5.replace(bias, str(bias_row[pool[4]].values[0]))

  copy_row_1.Tweet = tweet_1
  copy_row_2.Tweet = tweet_2
  copy_row_3.Tweet = tweet_3
  copy_row_4.Tweet = tweet_4
  copy_row_5.Tweet = tweet_5

  syn_df = syn_df.append(copy_row_1)
  syn_df = syn_df.append(copy_row_2)
  syn_df = syn_df.append(copy_row_3)
  syn_df = syn_df.append(copy_row_4)
  syn_df = syn_df.append(copy_row_5)
```

Kode 4.3: Pengembangan *dataset* sintetik

4.4 Eksperimen

Pada Subbab 4.4, dijelaskan kode bahasa pemrograman Python yang digunakan untuk melakukan evaluasi model dengan sampel *bootstrap* dari data latih, membandingkan hasil uji statistik dari *list* akurasi sampel *bootstrap*, dan membandingkan performa model pada *dataset* berbeda.

4.4.1 Evaluasi menggunakan Sampel *Bootstrap*

Evaluasi model menggunakan sampel *bootstrap* dilakukan dengan cara mengambil sampel dari data latih sejumlah panjang data latih itu sendiri menggunakan pengganti. Dari sampel *bootstrap* yang didapat, masing-masing model pemelajaran mesin dilatih dan digunakan untuk memprediksi data *test* asli, dan nilai akurasi dari prediksi disimpan pada suatu *list*. Pada penelitian ini, pengambilan sampel *bootstrap* dilakukan sebanyak 30 iterasi.

```
"""
Original file is located at

    https://colab.research.google.com/drive/1_uutSpda61LZqGdPKjgUfqBsottu8XdK#scrollTo=mqSJnkzFwOks
"""
# https://insidelearningmachines.com/implement-the-bootstrap-method-in-python/
def generate_bootstrap_sample(source_df):
    bwr = source_df.copy()[0:0]
    idx = source_df.index.values
    sample_ids = np.random.choice(idx,replace=True,size=len(idx))
    for i in sample_ids:
        row = source_df.loc[source_df.index == i]
        bwr = bwr.append(row)
    return bwr

def get_x_y(df):
    x = df.loc[:, df.columns != 'Tweet']
    x = x.loc[:, x.columns != 'HS']
    x = x.loc[:, x.columns != 'A']
    y = df.loc[:, df.columns == label]      # label <- 'HS' atau 'A'
    return x, y

acc_list = []
for n in range(30):
    bwr = generate_bootstrap_sample(all_train_data)
    X, y = get_x_y(bwr)
    model = Classifier().fit(X, y)
    acc_list.append(accuracy_score(y_test, model.predict(X_test)))

rata_rata = sum(acc_list)/len(acc_list)
```

Kode 4.4: Pelatihan model menggunakan sampel *bootstrap*

4.4.2 Evaluasi menggunakan uji statistik

Menggunakan *list* akurasi yang didapat dari pelatihan model menggunakan data sampel *bootstrap*, uji statistik dilakukan dengan memanfaatkan fungsi *ttest_rel* dari *library scipy.stats*. Perbandingan dilakukan sebanyak pasangan model yang bisa dibentuk. Pada penelitian ini, kelima model membentuk 10 pasangan yang dibandingkan dengan satu sama lain.

```
from scipy.stats import ttest_rel

modelA_to_modelB = ttest_rel(modelA_accuracy_list, modelB_accuracy_list)
```

Kode 4.5: Evaluasi *paired t-test*

4.4.3 Deteksi Bias menggunakan *Dataset* Sintetis

Untuk melakukan uji deteksi bias, model yang sudah dikembangkan menggunakan data latih asli digunakan untuk memprediksi data *test* asli, data sintetik hasil pengembangan dengan input data *test*, data yang dikembangkan secara manual, dan data sintetik hasil pengembangan dengan input data yang dikembangkan secara manual.

```
"""
Original file is located at
    https://colab.research.google.com/drive/1xSgIxE0xlVKx8q9XtUSWA6YCdSsXreso
"""
"""# Dilakukan untuk masing-masing model"""
model_hs = Classifier().fit(X_train_HS, y_train_HS)
model_ab = Classifier().fit(X_train_Ab, y_train_Ab)

hasil_hs_test_asli = accuracy_score(
                        y_test_HS,
                        model_hs.predict(X_test_HS)
                        )
hasil_hs_test_sintetis = accuracy_score(
                        y_test_sintetis_HS,
                        model_hs.predict(X_test_sintetis_HS)
                        )
hasil_hs_data_olahan_asli = accuracy_score(
                        y_olahan_hs,
                        model_HS.predict(X_olahan_HS)
                        )
hasil_hs_data_olahan_sintetis = accuracy_score(
                        y_olahan_sintetis_HS,
                        model_hs.predict(X_olahan_sintetis_HS)
                        )

hasil_ab_test_asli = accuracy_score(
                        y_test_Ab,
                        model_ab.predict(X_test_Ab)
                        )
hasil_ab_test_sintetis = accuracy_score(
                        y_test_sintetis_Ab,
                        model_ab.predict(X_test_sintetis_Ab)
                        )
hasil_ab_data_olahan_asli = accuracy_score(
                        y_olahan_Ab,
                        model_ab.predict(X_olahan_Ab)
                        )
hasil_ab_data_olahan_sintetis = accuracy_score(
                        y_olahan_sintetis_Ab,
                        model_ab.predict(X_olahan_sintetis_Ab)
                        )
```

Kode 4.6: Pengembangan model dan uji deteksi bias

BAB 5
EKSPERIMEN DAN ANALISIS

Bab 5 mendiskusikan hasil dan analisis dari eksperimen yang dilakukan. Pertama, Subbab 5.1 membahas seberapa berpengaruh fitur-fitur yang diusulkan terhadap pelatihan model yang dikembangkan. Kemudian, Subbab 5.2 membahas performa rata-rata dari model yang dikembangkan dan membandingkan performa model pada masing-masing *dataset* untuk *testing*. Terakhir, Subbab 5.3 membahas metode deteksi bias yang diusulkan.

5.1 Hasil *Feature Selection*

Feature selection dilakukan untuk mendapatkan urutan fitur berdasarkan pengaruh. Hasil perhitungan *f score* dari penggunaan metode ANOVA *f-test* beserta koefisien dan *p values* dari penggunaan *logistic regression analysis* untuk 30 fitur dengan *f score* tertinggi ditampilkan secara lengkap di Lampiran 2 dan Lampiran 3. Beberapa fitur dengan *f score* tertinggi untuk masing-masing tugas klasifikasi adalah sebagai berikut:

Tabel 5.1: Tabel 5 fitur dengan *f score* tertinggi untuk deteksi ujaran kebencian

Fitur	***F Score***	***coef***	***p***
cebong	110,348	3,717	0,000
2019gantipresiden	75,901	3,367	0,000
Exclamation	70,554	1,437	0,000
jokowi	63,371	1,347	0,000
prabowo	54,861	3,353	0,000

Pada Tabel 5.1, berdasarkan *f score*, keberadaan kata ”cebong” memiliki pengaruh terbesar dalam keputusan sebuah pesan terdeteksi sebagai ujaran kebencian. Hal ini didukung dengan hasil LRA yang membuktikan bahwa ketika pesan mengandung kata ”cebong”, maka *odds ratio* pesan tersebut dideteksi sebagai ujaran kebencian (”*Hate Speech*”) meningkat sebesar $exp(3{,}717)$ atau 41,149 kali lipat daripada ketika pesan tidak mengandung kata ”cebong”. *Dataset* ini memiliki domain politik, sehingga penggunaan

kata cebong[1] tidak memiliki makna anak kodok, tetapi merepresentasikan terminologi ujaran kebencian terkait kontestasi politik di tahun 2019. Selain itu, Tabel 5.1 menunjukkan bahwa dari 5 fitur dengan *f score* tertinggi, sebagian besar memiliki keterkaitan dengan kontestasi politik di tahun 2019. Namun, hasil tersebut juga menunjukkan bahwa keberadaan kata ”jokowi” dan ”prabowo” memiliki pengaruh terhadap pendeteksian suatu pesan sebagai ujaran kebencian. Hasil *ANOVA f-test* dan LRA untuk deteksi ujaran kebencian dapat dilihat selengkapnya di Lampiran 2.

Tabel 5.2: Tabel 5 fitur dengan *f score* tertinggi untuk deteksi bahasa kasar

Fitur	***F Score***	***coef***	***p***
Abusive	863,491	2,662	0,000
cebong	148,459	4,835	0,000
Length	79,293	-0,006	0,000
Words	59,505	-0,059	0,000
goblok	55,097	-	-

Pada Tabel 5.2, berdasarkan *f score*, fitur keberadaan kata yang ada di *abusive lexicon* dari penelitian yang dilakukan oleh Ibrohim dan Budi (2019) memiliki pengaruh terbesar dalam keputusan sebuah pesan terdeteksi sebagai bahasa kasar. Hal ini didukung dengan hasil LRA yang membuktikan bahwa ketika fitur ”*Abusive*” bernilai 1, maka *odds ratio* pesan tersebut dideteksi sebagai bahasa kasar (”*Abusive*”) meningkat sebesar $exp(2{,}662)$ atau 14,325 kali lipat daripada ketika fitur ”*Abusive*” bernilai 0. Perlu dicatat bahwa fitur ”*Abusive*” di sini bukanlah target klasifikasi, tetapi merupakan fitur bernilai biner yang mencatat apakah pesan mengandung kata yang ada di *abusive lexicon*. Hasil *ANOVA f-test* dan LRA untuk deteksi bahasa kasar dapat dilihat selengkapnya di Lampiran 3.

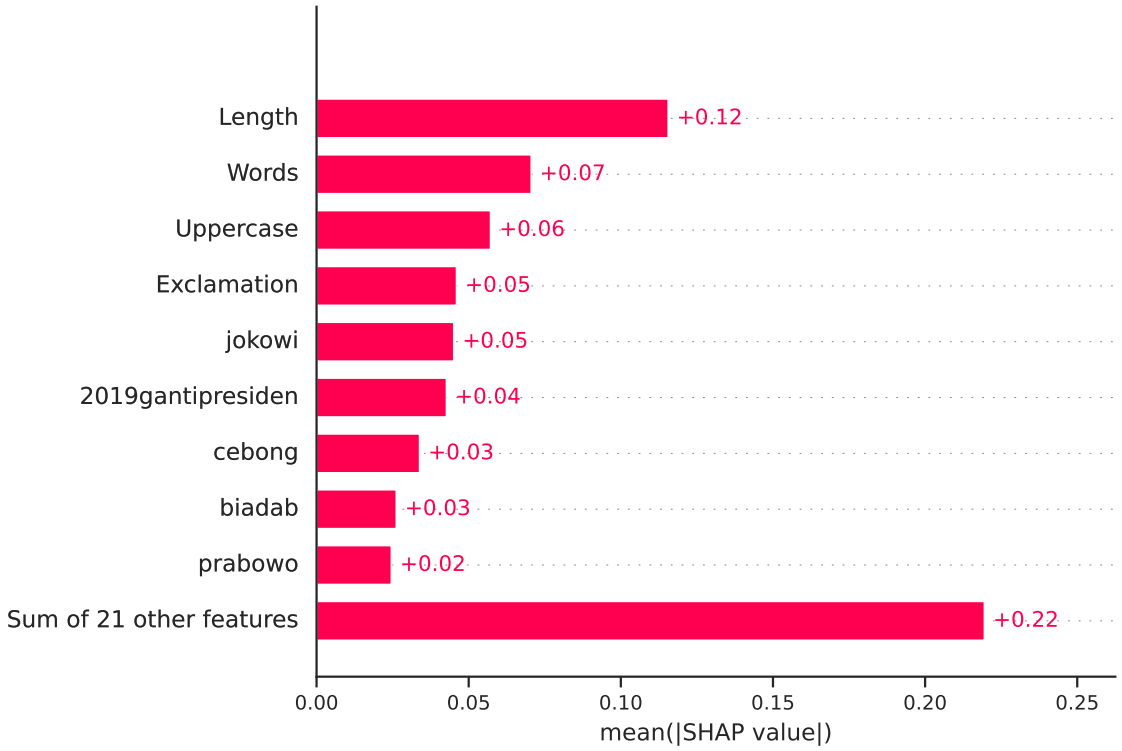
Selain itu, teknik lain yang digunakan untuk menilai pengaruh fitur terhadap keputusan model adalah penggunaan modul visualisasi SHAP. *Tools* SHAP digunakan untuk melakukan visualisasi *feature importance* dan *summary plot* dari 30 fitur dengan *f score* tertinggi untuk masing-masing tugas klasifikasi. Visualisasi SHAP dilakukan pada kelima model yang digunakan dalam penelitian ini untuk melakukan pendeteksian ujaran kebencian dan bahasa kasar ketika menggunakan 30 fitur dengan *f score* tertinggi.

[1]Berdasarkan KBBI daring, cebong adalah anak kodok yang masih kecil berwujud seperti ikan dan hidup di air; berudu

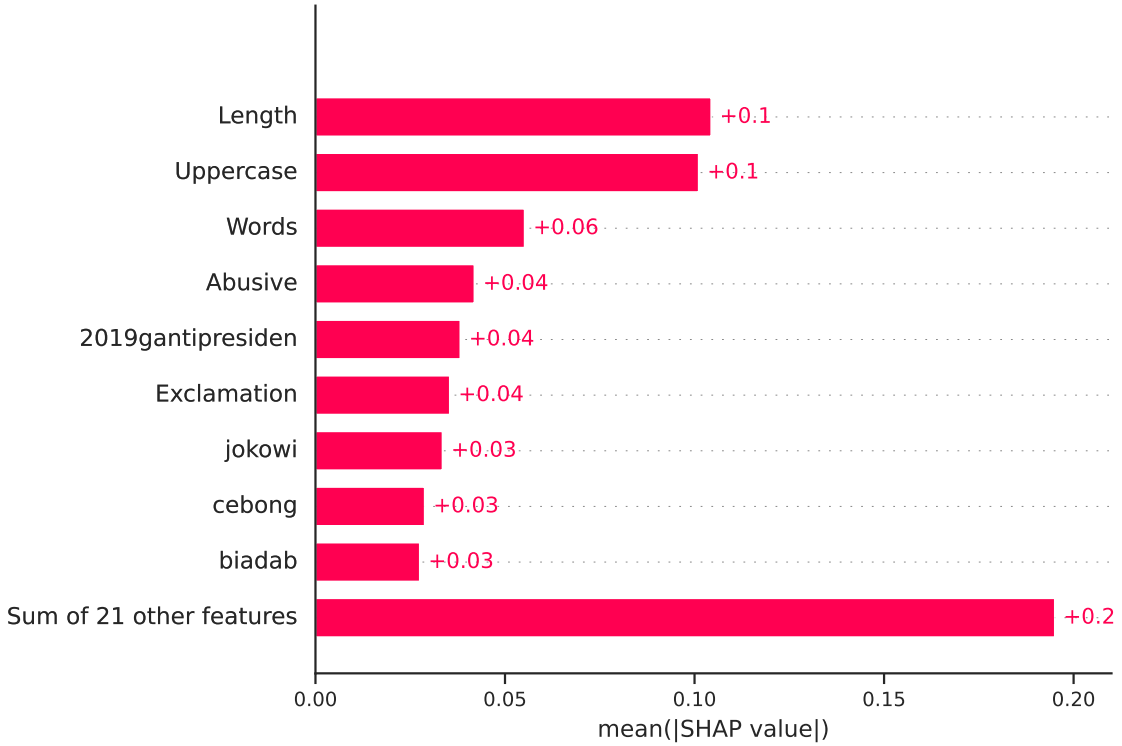
Gambar 5.1 menunjukkan bahwa fitur ”*Length*”, yaitu representasikan panjang pesan, merupakan fitur yang dianggap memiliki kontribusi terbesar pada klasifikasi ujaran kebencian menggunakan model *LogReg*. Hasil serupa dapat dilihat pada Gambar 5.2 untuk model *XGBoost*, Gambar 5.4 untuk model MLP, dan Gambar 5.5 untuk model SVM. Namun, Gambar 5.3 menunjukkan bahwa fitur dengan kontribusi terbesar untuk model *CatBoost* adalah fitur ”*Uppercase*” yang merepresentasikan jumlah huruf kapital pada pesan. Sedangkan untuk pendeteksian bahasa kasar, Gambar 5.6 menunjukkan bahwa pada model *LogReg*, fitur dengan kontribusi terbesar dalam hasil prediksi adalah keberadaan tanda seru. Namun, visualisasi *feature importance* untuk model *XGBoost* pada Gambar 5.7 dan model SVM pada Gambar 5.10 menunjukkan bahwa fitur dengan kontribusi terbesar adalah panjang pesan. Kemudian, Gambar 5.8 untuk model *CatBoost* dan Gambar 5.9 untuk MLP menunjukkan bahwa fitur dengan kontribusi terbesar adalah jumlah huruf kapital.

Kesimpulan yang dapat ditarik dari hasil-hasil tersebut adalah berdasarkan rata-rata nilai *SHAPley* absolut, struktur kalimat memiliki kontribusi keseluruhan yang lebih besar daripada fitur keberadaan kata. Namun, perlu dicatat bahwa fitur keberadaan kata ”jokowi” merupakan salah satu fitur penting dalam pendeteksian ujaran kebencian dan bahasa kasar pada seluruh model. Hasil ini menunjukkan bahwa keberadaan kata ”jokowi” memiliki kontribusi terhadap hasil prediksi. Rincian kontribusi fitur berdasarkan *feature importance* dan *feature effects* pada setiap model ditampilkan di lampiran 4 dan 5.

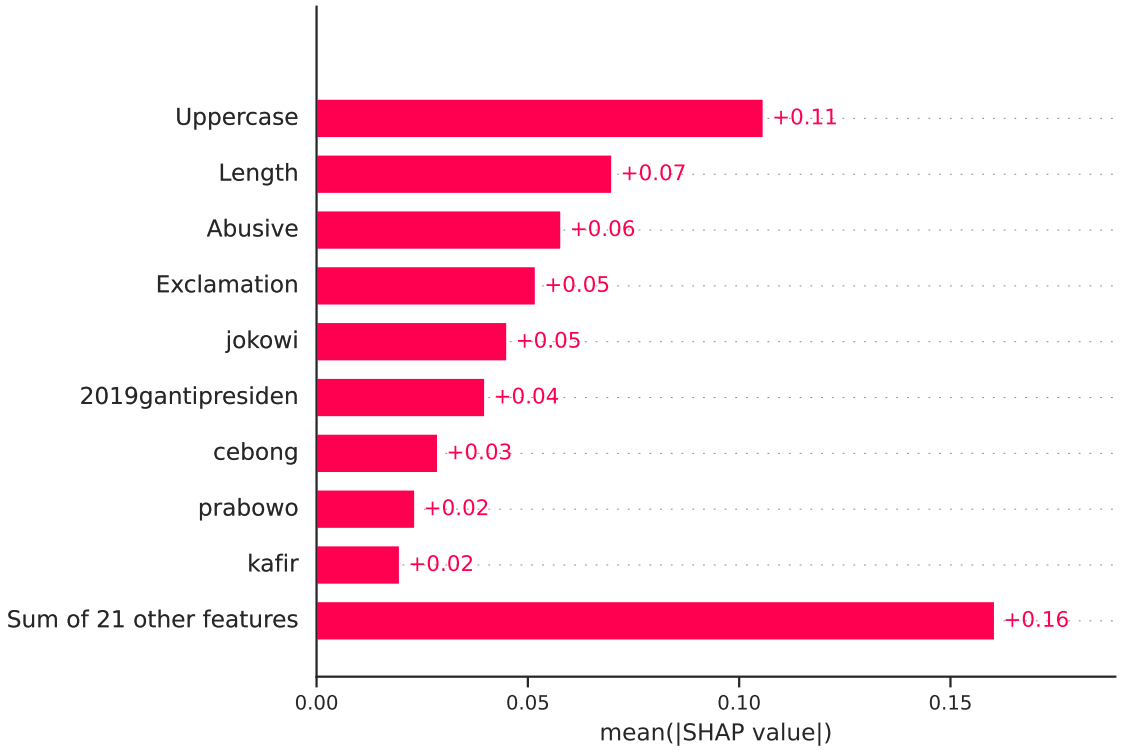
Berdasarkan ketiga teknik *feature selection* yang digunakan, didapatkan fitur-fitur penting untuk kedua tugas klasifikasi. Untuk tugas klasifikasi pendeteksian ujaran kebencian, didapatkan fitur-fitur yang dianggap penting berupa keberadaan kata-kata terkait dorongan supaya presiden mengundurkan diri seperti ”2019gantipresiden” dan ”lengser”; keberadaan kata-kata terminologi politik yang memiliki unsur kebencian seperti ”cebong” dan ”kampret”; keberadaan kata-kata yang bersifat kasar seperti ”tolol”, ”goblok”, ”dungu”, dan ”kafir”; dan fitur terkait struktur pesan seperti keberadaan tanda seru, panjang pesan, banyak huruf kapital, dan jumlah kata dalam pesan. Untuk tugas klasifikasi pendeteksian bahasa kasar, didapatkan fitur-fitur yang dianggap penting berupa keberadaan kata-kata yang ada di dalam *abusive lexion* hasil pekerjaan yang dilakukan oleh Ibrohim dan Budi (2019) dan bersifat kasar seperti ”cebong”, ”kontol”, ”anjing”, ”bacot” dan sebagainya; dan fitur terkait struktur pesan seperti panjang pesan, jumlah kata dalam pesan, dan banyak huruf kapital.



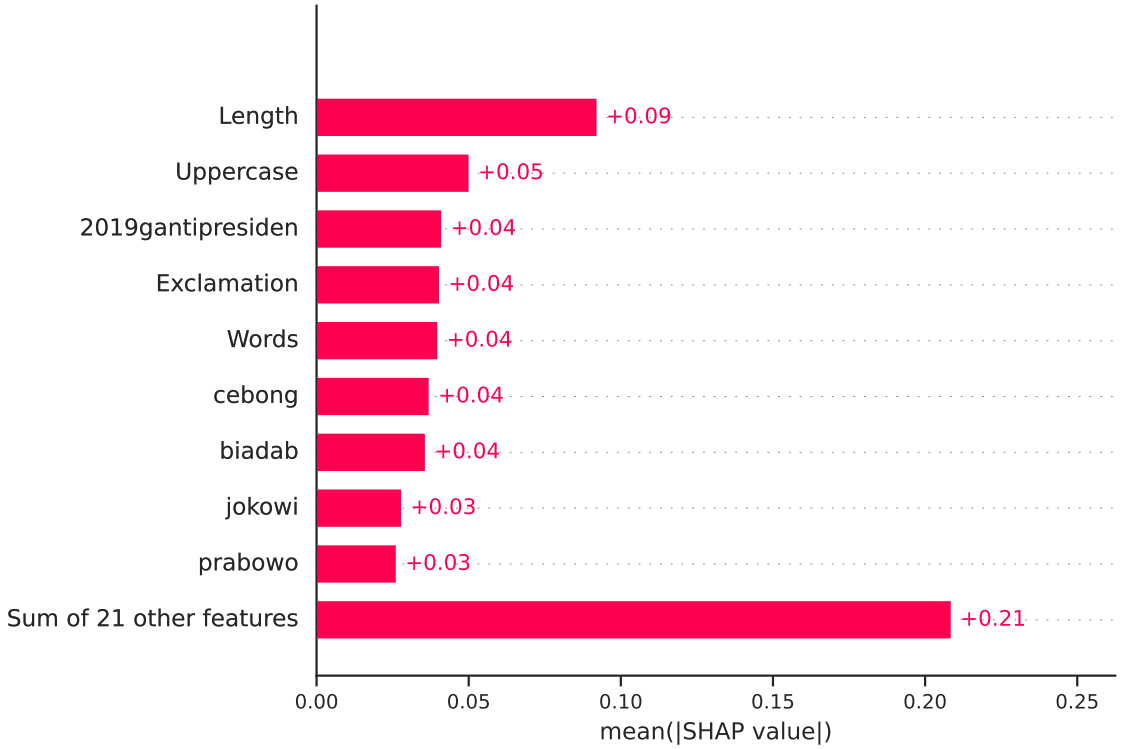
Gambar 5.1: Visualisasi *feature importance* dari model *LogReg* untuk deteksi ujaran kebencian



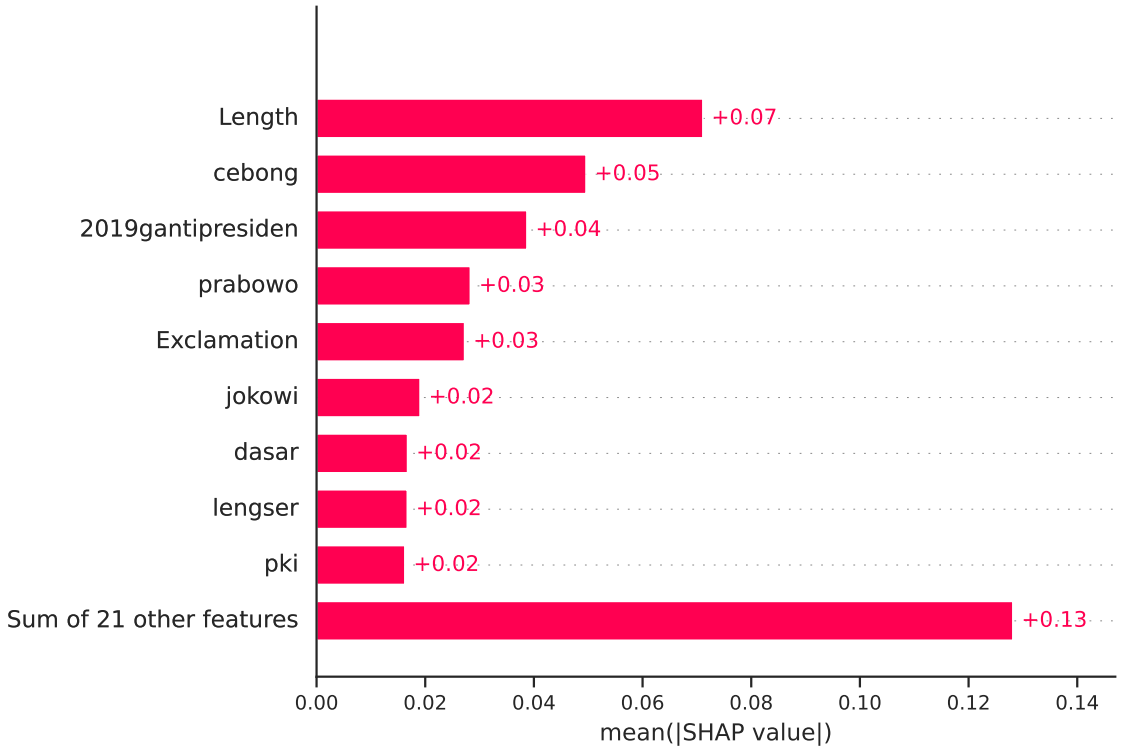
Gambar 5.2: Visualisasi *feature importance* dari model *XGBoost* untuk deteksi ujaran kebencian



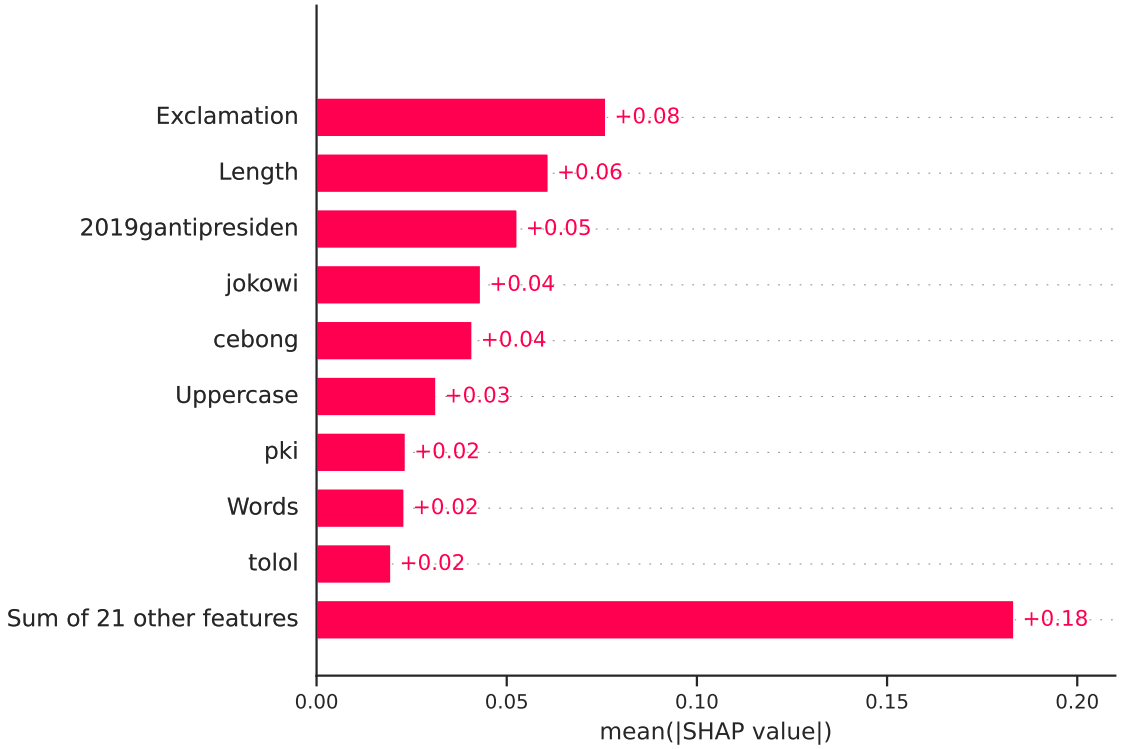
Gambar 5.3: Visualisasi *feature importance* dari model *CatBoost* untuk deteksi ujaran kebencian



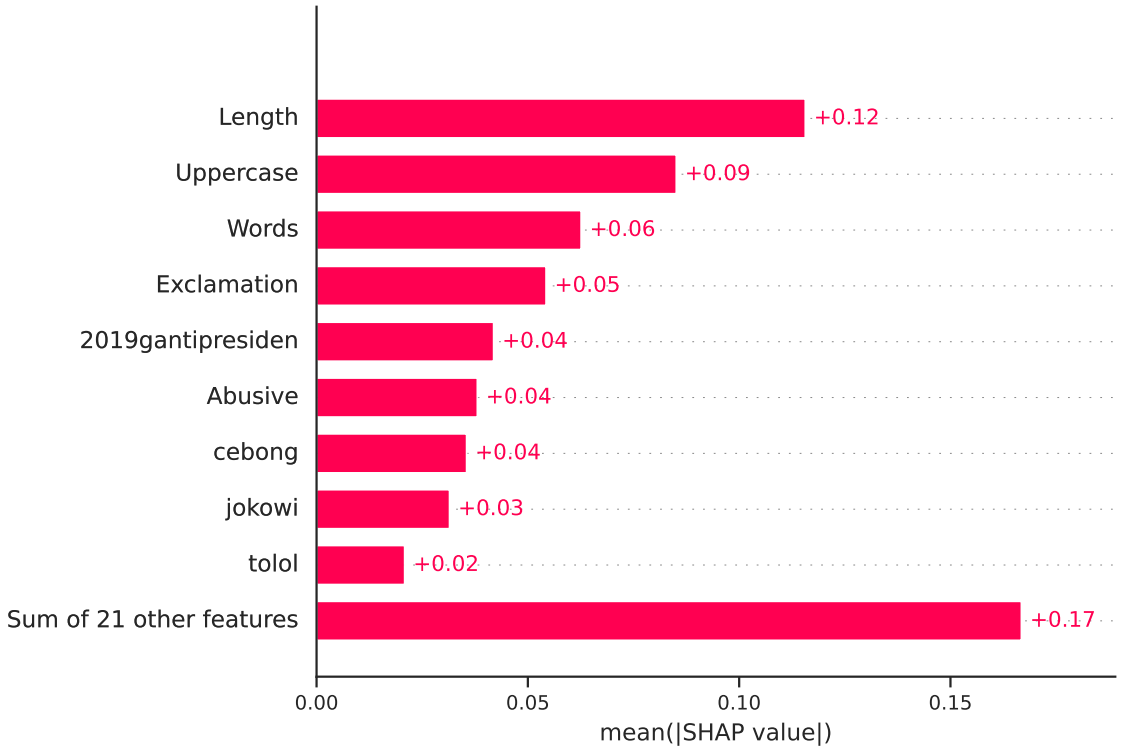
Gambar 5.4: Visualisasi *feature importance* dari model MLP untuk deteksi ujaran kebencian



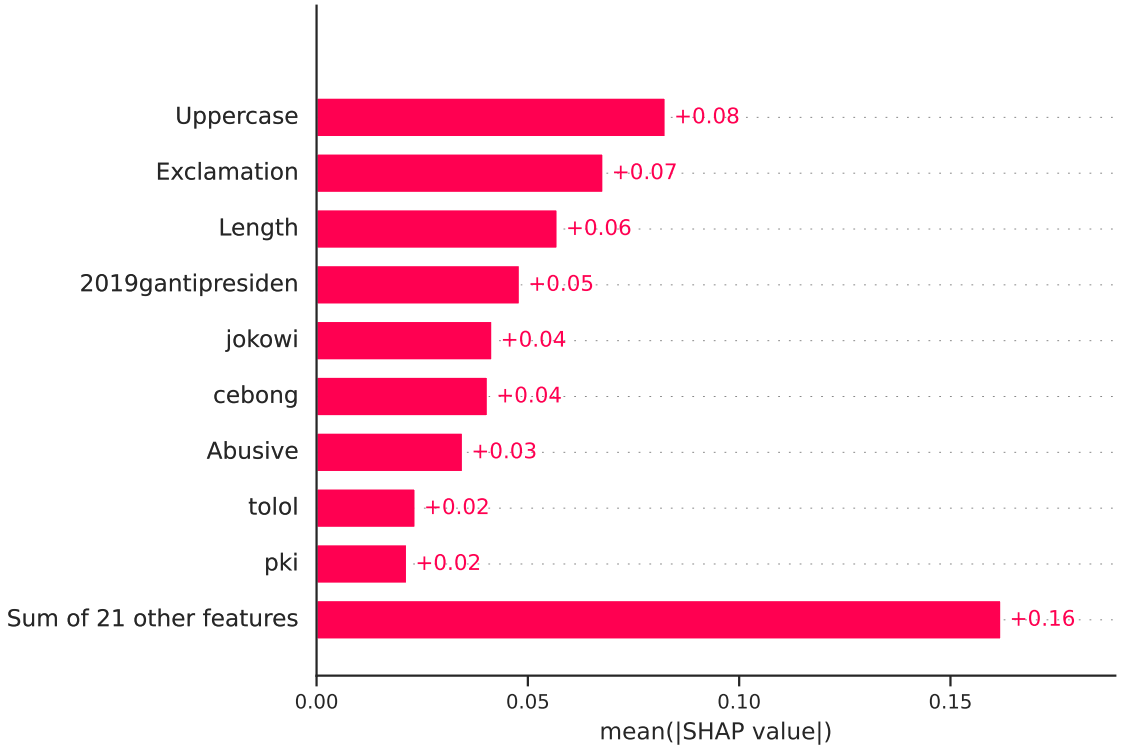
Gambar 5.5: Visualisasi *feature importance* dari model SVM untuk deteksi ujaran kebencian



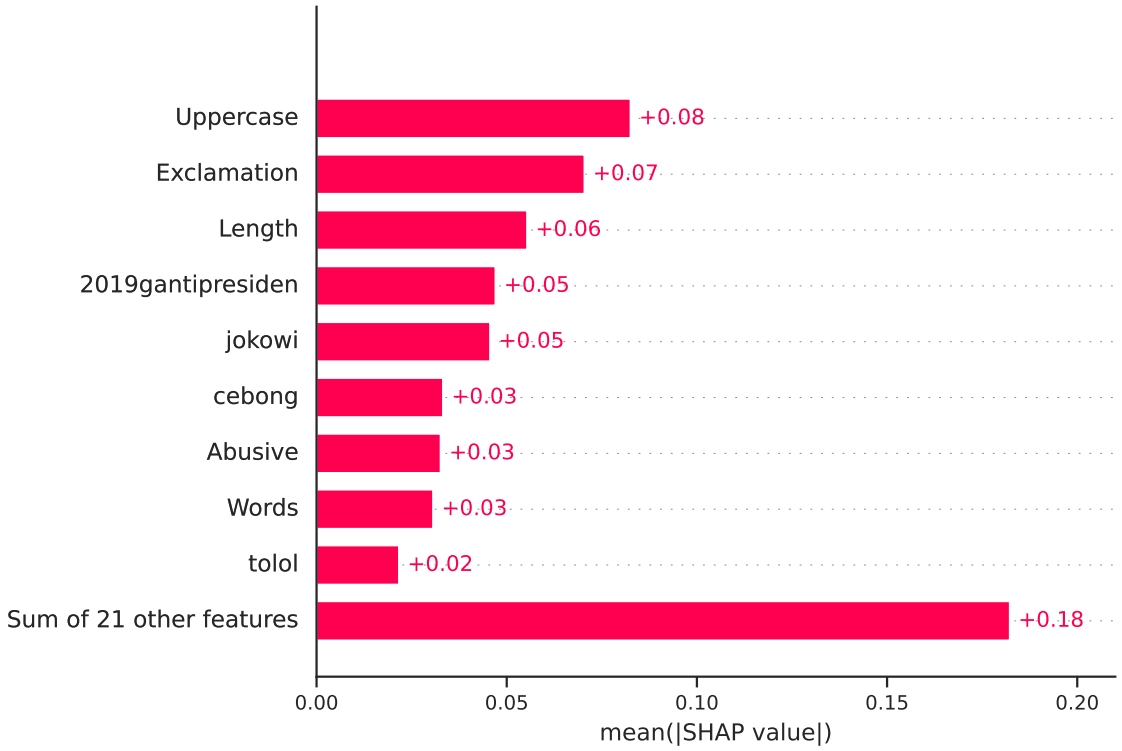
Gambar 5.6: Visualisasi *feature importance* dari model *LogReg* untuk deteksi bahasa kasar



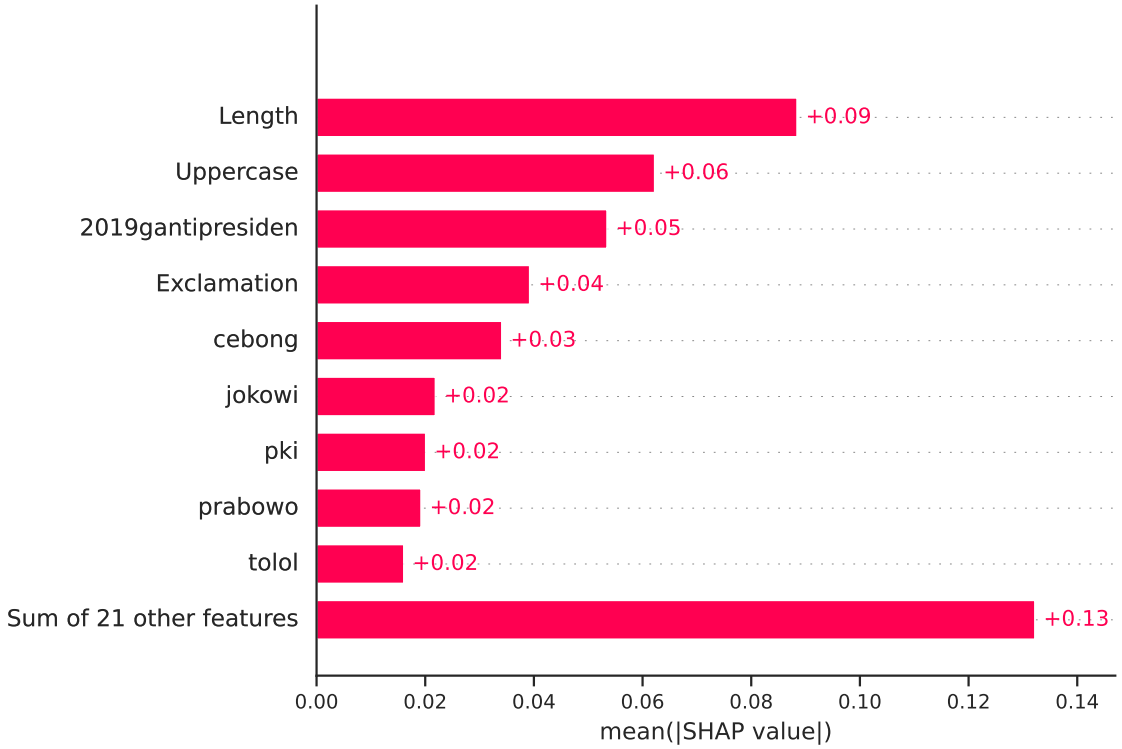
Gambar 5.7: Visualisasi *feature importance* dari model *XGBoost* untuk deteksi bahasa kasar



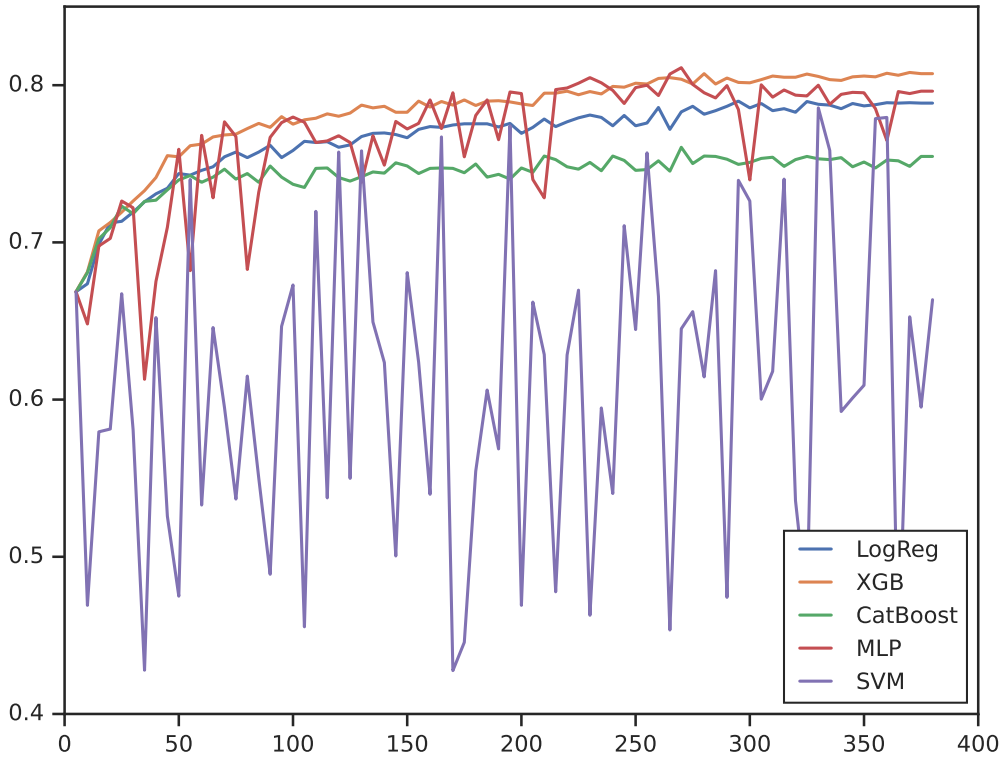
Gambar 5.8: Visualisasi *feature importance* dari model *CatBoost* untuk deteksi bahasa kasar



Gambar 5.9: Visualisasi *feature importance* dari model MLP untuk deteksi bahasa kasar



Gambar 5.10: Visualisasi *feature importance* dari model SVM untuk deteksi bahasa kasar

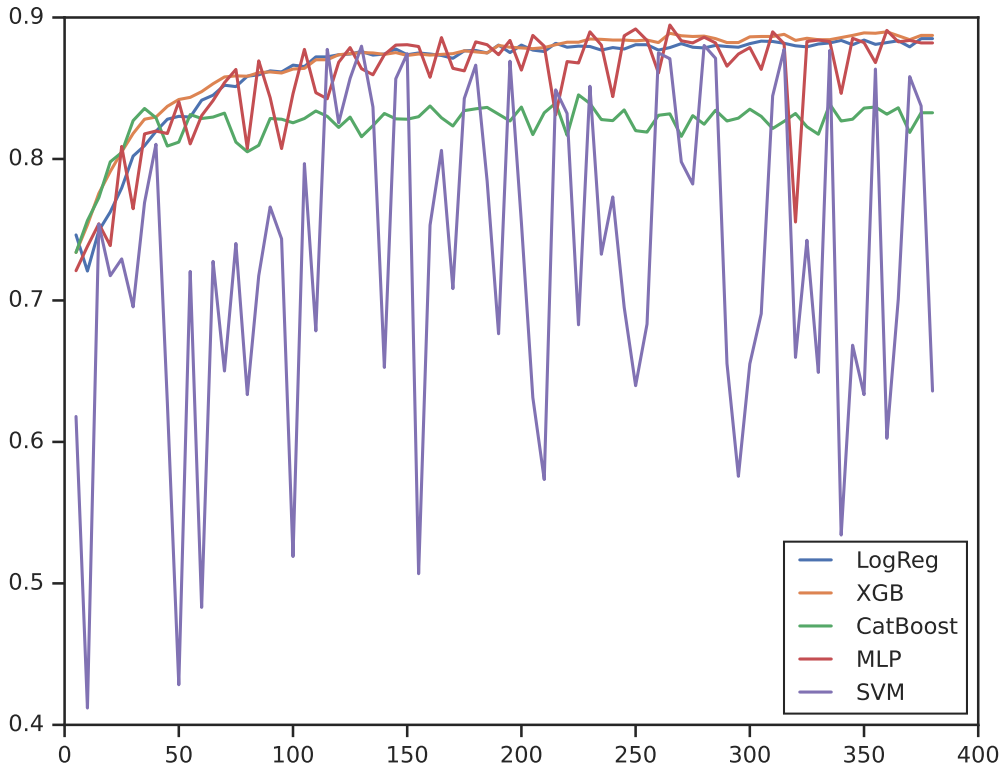


Gambar 5.11: Hasil pemetaan jumlah fitur yang dipakai terhadap akurasi model deteksi ujaran kebencian

Terakhir, dilakukan pemetaan akurasi model berdasarkan jumlah fitur yang digunakan dari usulan fitur-fitur yang sebelumnya sudah diurutkan. Berdasarkan Gambar 5.11 dan Gambar 5.12, dapat dilihat bahwa model *Logistic Regression* dan *XGBoost* memiliki peningkatan akurasi seiring bertambahnya fitur, dan model *CatBoost* memiliki peningkatan akurasi tapi hanya di awal penambahan fitur. Berdasarkan kedua gambar, model SVM bisa dikatakan tidak *robust* terhadap penambahan fitur secara bertahap, sama dengan model MLP walaupun ada kecenderungan peningkatan akurasi. Menggunakan hasil dari pemetaan fitur terhadap akurasi model, dipilih 360 fitur dari urutan fitur-fitur yang diusulkan. Alasan hanya dipilih 360 fitur dari 389 fitur yang diusulkan adalah ketika dilihat secara detail, hampir tidak ada perubahan pada akurasi model ketika dilakukan penambahan fitur untuk melatih model menggunakan lebih dari 360 fitur, tanpa mempertimbangkan model yang tidak *robust*.

5.2 Hasil Evaluasi Model

Evaluasi model dilakukan menggunakan dua metode, metode sampel *bootstrap* dengan pengganti dan metode uji statistik. Berdasarkan hasil pada Tabel 5.3 dapat dilihat bahwa



Gambar 5.12: Hasil pemetaan jumlah fitur yang digunakan terhadap akurasi model deteksi bahasa kasar

model *XGBoost* memiliki performa rata-rata tertinggi yaitu akurasi 0.791 dalam mendeteksi ujaran kebencian dan akurasi 0.876 dalam mendeteksi bahasa kasar, diikuti oleh model *LogReg* dengan akurasi 0.779 dalam mendeteksi ujaran kebencian dan 0.875 dalam mendeteksi bahasa kasar. Namun ketika kedua model dibandingkan menggunakan uji statistik, dapat dilihat pada Tabel 5.4 bahwa model *LogReg* terbukti lebih unggul daripada model *XGBoost* untuk pendeteksian ujaran kebencian, dan hasil pada Tabel 5.5menunjukkan bahwa tidak ada kesimpulan untuk pendeteksian bahasa kasar ($pvalue > 0.05$).

Tabel 5.3: Tabel akurasi rata-rata model pada *dataset* sampel *bootstrap*

	Deteksi Ujaran Kebencian	**Deteksi Bahasa Kasar**
LogReg	0,779	0,875
XGBoost	0,791	0,876
CatBoost	0,745	0,813
MLP	0,768	0,859
SVM	0,659	0,795

Tabel 5.4: Tabel hasil uji statistik untuk deteksi ujaran kebencian

	XGBoost	***CatBoost***	***MLP***	***SVM***
LogReg	-10,229	22,296	4,921	7,985
XGBoost		30,761	11,229	8,690
CatBoost			-10,601	5,773
MLP				7,182

Tabel 5.5: Tabel hasil uji statistik untuk deteksi bahasa kasar

	XGBoost	***CatBoost***	***MLP***	***SVM***
LogReg	-	24,045	11,973	4,382
XGBoost		23,594	13,889	4,490
CatBoost			-14,618	-
MLP				3,441

5.3 Hasil Uji Deteksi Bias

Berdasarkan hasil pada proses *feature selection*, ditemukan bias di tingkat *dataset* sebab keberadaan kata-kata potensi bias terbukti memiliki pengaruh terhadap hasil prediksi. Dapat dilihat pada Tabel 5.6 besar pengaruh kata-kata potensi bias. Berdasarkan koefisien LRA, dari seluruh kata-kata potensi bias, keberadaan kata ”prabowo” memiliki keterkaitan terbesar dengan label ”*Hate Speech*”. Artinya, keberadaan kata ”prabowo” pada suatu pesan membuat kemungkinan pesan tersebut dideteksi sebagai ujaran kebencian meningkat sekitar 28 kali daripada kalau pesan tersebut tidak mengandung kata ”prabowo”. Di sisi lain, koefisien LRA menunjukkan bahwa kata-kata potensi bias tidak tebukti memiliki pengaruh terhadap meningkatnya kemungkinan suatu pesan dideteksi sebagai bahasa kasar.

Uji deteksi bias di tingkat model dilakukan berdasarkan performa hasil pengembangan model menggunakan data latih yang kemudian diuji performanya menggunakan data *test* dan data yang dikembangkan secara manual. Setelah itu hasil yang didapat dibandingkan dengan versi yang sudah tidak mengandung kata-kata potensi bias dari masing-masing data. Penulis memiliki asumsi bahwa apabila terdapat bias pada suatu model, ketika

Tabel 5.6: Tabel pengaruh kata bias berdasarkan *ANOVA f-test* dan LRA

	Deteksi Ujaran Kebencian			**Deteksi Bahasa Kasar**		
	ANOVA f-test	**LRA**		*ANOVA f-test*	**LRA**	
		coef	***p-value***		***coef***	***p-value***
jokowi	63,371	1,347	0,000	37,646	-1,411	0,000
islam	6,009	0,544	0,001	14,118	-0,831	0,000
indonesia	0,863	0,462	0,005	25,119	-1,030	0,000
prabowo	54,861	3,352	0,000	7,543	-0,884	0,005
cina	14,316	1,070	0,000	4,016	-0,511	0,017
ahok	53,126	2,545	0,000	0,095	0,308	0,179
presiden	1,470	-0,427	0,000	27,984	-1,400	0,000
kpk	10,985	1,671	0,000	7,957	-2,137	0,004
gubernur	10,152	-0,739	0,001	38,462	-2,139	0,000
sby	12,228	1,471	0,001	5,979	-1,692	0,006
anies	12,228	1,626	0,000	3,837	-1,323	0,015
dpr	6,424	0,964	0,000	5,979	-1,364	0,012
muslim	1,410	0,634	0,025	1,640	-0,431	0,162
sandi	8,114	2,338	0,002	7,765	-27,329	1,000
anies-sandi	12,127	20,299	0,997	2,973	-1,925	0,066
rohingya	5,361	20,299	0,997	2,573	-19,505	0,998
kristen	17,574	-1,551	0,000	31,534	-3,313	0,000
buddha	0,599	0,197	0,686	5,810	-27,329	1,000
myanmar	8,059	23,878	1,000	3,865	-1,471	0,169

model tersebut diuji menggunakan data yang bebas dari faktor-faktor bias, akan ada perubahan pada performa model.

Pada Tabel 5.7 dan Tabel 5.9, akurasi model dalam memprediksi data *test* dicatat sebagai hasil ”Data *Test* Asli” dan akurasi model dalam memprediksi data sintetis yang dikembangan secara otomatis menggunakan data *test* dicatat sebagai hasil ”Data *Test* Sintetis”. Pada Tabel 5.8 dan Tabel 5.10, akurasi model dalam memprediksi data yang dikembangkan secara manual dicatat sebagai hasil ”Data Manual Asli” dan akurasi model dalam memprediksi data sintetis yang dikembangkan menggunakan data tersebut dicatat sebagai hasil ”Data Manual Sintetis”. Jika hasil akurasi di data asli lebih besar dari data sintetis, dapat disimpulkan bahwa model belajar dari keberadaan kata-kata potensi bias dan ketika kata-kata tersebut digantikan ditemukan pesan-pesan yang sebelumnya dapat diprediksi dengan benar menjadi salah. Sebaliknya, jika hasil akurasi di data asli lebih kecil dari data sintetis, dapat disimpulkan bahwa model masih belajar dari keberadaan kata-kata potensi bias, hanya saja ditemukan pesan-pesan yang sebelumnya gagal diprediksi dengan benar karena keberadaan kata-kata tersebut menjadi dapat diprediksi dengan benar. Hasil

yang ideal adalah ketika tidak ada perbedaan antara hasil akurasi di data asli dan sintetis, karena artinya keberadaan kata-kata potensi bias tidak memiliki pengaruh besar terhadap pelatihan model.

Tabel 5.7: Tabel akurasi model untuk deteksi ujaran kebencian pada data *test*

	Data *Test* Asli	**Data *Test* Sintetis**
LogReg	0,789	0,701
XGBoost	0,808	0,710
CatBoost	0,772	0,700
MLP	0,765	0,685
SVM	0,654	0,621

Tabel 5.8: Tabel akurasi model untuk deteksi ujaran kebencian pada data yang dikembangkan secara manual

	Data Manual Asli	**Data Manual Sintetis**
LogReg	0,430	0,960
XGBoost	0,580	0,960
CatBoost	0,750	1,000
MLP	0,460	0,962
SVM	0,850	0,998

Berdasarkan hasil yang ditampilkan di Tabel 5.7 dan Tabel 5.8, dapat dilihat bahwa model yang dilatih menggunakan data latih yang mengandung kata-kata potensi bias memiliki akurasi cukup baik pada data *test* asli. Ketika kata-kata bias dihilangkan dari data *test*, terjadi penurun akurasi. Hal ini menandakan bahwa keberadaan kata-kata tersebut mempengaruhi keputusan model. Untuk melakukan uji deteksi bias lebih lanjut, penulis mengembangkan *dataset* yang terdiri atas 100 pesan yang masing-masing mengandung kata-kata potensi bias tapi seluruhnya memiliki label ”*Not Hate Speech*” dan ”*Not Abusive*”. Ketika model digunakan untuk memprediksi ”*Hate Speech*”/”*Not Hate Speech*” pada *dataset* tersebut, akurasi beberapa model sangatlah rendah. Hasil tersebut menunjukkan bahwa keberadaan kata-kata potensi bias membuat pesan yang sebenarnya bukan merupakan ujaran kebencian dideteksi sebagai ujaran kebencian. Ini didukung dengan peningkatan performa ketika kata-kata potensi bias dihilangkan dari *dataset* yang dikembangkan secara manual.

Sebagai kesimpulan, hasil akurasi di ”*Test* Asli” didapatkan lebih besar dari hasil di akurasi di ”*Test* Sintetis”, sehingga dapat diartikan bahwa model belajar dari keberadaan kata-kata potensi bias dalam melakukan prediksi dan ketika kata-kata tersebut digantikan ditemukan pesan yang sebelumnya dapat diprediksi dengan benar menjadi salah, sehingga terjadi penurunan akurasi. Sedangkan dalam memprediksi *dataset* yang dikembangkan secara manual, ketika hasil akurasi di ”Manual Asli” lebih kecil dari hasil di akurasi di ”Manual Sintetis”, dapat diartikan bahwa ditemukan pesan yang seharusnya bukan merupakan ujaran kebencian tapi diprediksi menjadi ujaran kebencian karena keberadaan kata-kata potensi bias. Hal ini dikarenakan *dataset* yang dikembangkan secara manual hanya mengandung pesan yang bukan merupakan ujaran kebencian dan bukan merupakan bahasa kasar, kegagalan memprediksi menandakan terdapat pesan yang dideteksi sebagai ujaran kebencian. Hasil tersebut menunjukkan keberadan bias seleksi ketika model memprediksi data yang dikembangkan secara manual, karena model dilatih menggunakan data latih di mana kata-kata berpotensi bias banyak munculnya dalam pesan yang berupa ujaran kebencian.

Tabel 5.9: Tabel akurasi model untuk deteksi bahasa kasar pada data *test*

	Data *Test* Asli	**Data *Test* Sintetis**
LogReg	0,882	0,860
XGBoost	0,890	0,863
CatBoost	0,867	0,842
MLP	0,891	0,863
SVM	0,878	0,855

Tabel 5.10: Tabel akurasi model untuk deteksi bahasa kasar pada data yang dikembangkan secara manual

	Data Manual Asli	**Data Manual Sintetis**
LogReg	0,990	0,990
XGBoost	0,980	0,980
CatBoost	0,990	0,990
MLP	0,980	0,980
SVM	0,990	0,980

Berdasarkan hasil yang ditampilkan di Tabel 5.9 dan Tabel 5.10, dapat dilihat bahwa kata-kata potensi bias tidak memiliki pengaruh besar dalam pendeteksian bahasa kasar.

Satu-satunya pesan yang konsisten salah diprediksi adalah kalimat ”pengen ke resto cina makan babi panggang” karena mengandung nama hewan yang umum dipakai sebagai kata kasar.

BAB 6
PENUTUP

Bab 6 mendiskusikan kesimpulan penelitian dan saran untuk penelitian berikutnya. Subbab 6.1 membahas kesimpulan yang dapat ditarik berdasarkan hasil analisa dari eksperimen yang telah dilakukan. Kemudian, Subbab 6.2 membahas saran terkait hal yang dapat dikembangkan dan hal-hal yang dapat dilakukan pada penelitian selanjutnya.

6.1 Kesimpulan

Penelitian ini melanjutkan pekerjaan yang sudah dilakukan oleh Ibrohim dan Budi (2019) terkait pengembangan *dataset* dan model untuk prediksi ujaran kebencian dan bahasa kasar bahasa Indonesia menggunakan pesan blog mikro. Sebagai kontribusi, pekerjaan ini menghasilkan *dataset* untuk melakukan uji bias berupa *dataset* yang dikembangkan secara otomatis dan yang dikembangkan secara manual. Terdapat dua isu utama yang dibahas dalam penelitian ini. Isu pertama adalah kajian *effect size* dari fitur-fitur, terutama fitur tekstual berbasis kata-kata, terhadap kemampuan deteksi ujaran kebencian dan bahasa kasar (Rumusan Masalah 1). Untuk menjawab Rumusan Masalah 1 tersebut, kajian *effect size* dilakukan dengan menggunakan tiga teknik *feature selection*: ANOVA *f-test*, *Logistic Regression Analysis*, dan analisis kontribusi dengan nilai *Shapley*. Masih terkait isu pertama, penelitian ini mengembangkan model prediksi dengan pemelajaran mesin berbasis fitur rancangan manual dan dengan melibatkan model yang pernah diuji oleh Ibrohim dan Budi (2019), yaitu SVM, dan model lain yang berbasis *ensemble* dan *neural networks* (Rumusan Masalah 2). Isu yang kedua adalah kajian terkait bias yang muncul pada model terhadap fitur kata-kata tertentu yang seharusnya netral, seperti ”jokowi”, ”islam”, dan ”cina” (Rumusan Masalah 3). Pembahasan terkait bias memiliki fokus kepada deteksi potensi kata-kata penyebab bias pada *dataset* dan pengembangan *dataset* untuk menguji bias yang muncul pada model akibat dilatih pada *dataset* yang dikembangkan oleh Ibrohim dan Budi (2019).

Analisis terkait *effect size* dilakukan berdasarkan hasil *feature selection* menggunakan tiga metode: *Analysis of Variance f-test* untuk mendapatkan urutan fitur berdasarkan *f*

score, *Logistic Regression Analysis* untuk menghitung keterkaitan fitur terhadap label, dan *Shapley Additive Explanations* untuk menghitung kontribusi fitur terhadap hasil prediksi model. Berdasarkan ketiga teknik *feature selection* yang digunakan, didapatkan fitur-fitur penting untuk kedua label. Untuk label "*Hate Speech*", didapatkan fitur-fitur yang dianggap penting berupa keberadaan kata-kata terkait dorongan supaya presiden mengundurkan diri seperti "2019gantipresiden" dan "lengser"; keberadaan kata-kata terminologi politik yang memiliki unsur kebencian seperti "cebong" dan "kampret"; keberadaan kata-kata yang bersifat kasar seperti "tolol", "goblok", "dungu", dan "kafir"; dan fitur terkait struktur pesan seperti keberadaan tanda seru, panjang pesan, banyak huruf kapital, dan jumlah kata dalam pesan. Untuk label "*Abusive*", didapatkan fitur-fitur yang dianggap penting berupa keberadaan kata-kata yang ada di dalam *abusive lexion* hasil pekerjaan yang dilakukan oleh Ibrohim dan Budi (2019) dan bersifat kasar seperti "cebong", "kontol", "anjing", "bacot" dan sebagainya; dan fitur terkait struktur pesan seperti panjang pesan, jumlah kata dalam pesan, dan banyak huruf kapital. Urutan fitur berdasarkan *f score* kemudian digunakan untuk memetakan performa model berdasarkan jumlah fitur yang digunakan secara meningkat. Dari proses ini didapatkan bahwa ketika digunakan lebih dari 360 fitur, performa model berhenti mengalami peningkatan.

Penelitian ini melakukan pendeteksian ujaran kebencian dan bahasa kasar menggunakan lima model pemelajaran mesin, yaitu: *Logistic Regression*, *Extreme Gradient Boosting*, *CatBoost*, *Multi-layer Perceptron*, dan *Support Vector Machine*. Pendeteksian ujaran kebencian dan bahasa kasar dilakukan secara terpisah menjadi dua tugas klasifikasi biner dengan target "*Hate Speech*"/"*Not Hate Speech*" dan "*Abusive*"/"*Not Abusive*". Evaluasi dilakukan pada model pemelajaran mesin yang digunakan dengan metode *bootstrap* dan uji statistik. Hasil evaluasi pengembangan model menggunakan data sampel *bootstrap* menunjukkan bahwa *XGBoost* dan *LogReg* adalah dua model dengan performa terbaik berdasarkan akurasi untuk kedua tugas klasifikasi. Selanjutnya, uji statistik digunakan untuk menentukan model terbaik dan didapatkan bukti bahwa model *LogReg* lebih baik daripada *XGBoost* untuk tugas klasifikasi deteksi ujaran kebencian tapi tidak didapat bukti yang menentukan mana di antara kedua model yang lebih baik untuk tugas klasifikasi deteksi bahasa kasar. Akurasi hasil pengembangan model menggunakan data latih dan data *test* adalah 0,779 untuk *LogReg* dan 0,791 untuk *XGBoost* dalam deteksi ujaran kebencian; dan 0,875 untuk *LogReg* dan 0,876 untuk *XGBoost* dalam deteksi bahasa kasar.

Terakhir, uji deteksi bias dilakukan menggunakan data *test* dan data yang dikem-

bangkan secara manual, dengan versi *dataset* yang kata-kata potensi bias di dalamnya sudah digantikan dengan kata-kata pengganti. Berdasarkan uji deteksi bias, bias ditemukan dalam tugas klasifikasi deteksi ujaran kebencian. Ketika model yang dikembangkan menggunakan data latih digunakan untuk memprediksi data *test*, terjadi penurunan akurasi ketika kata-kata potensi bias digantikan. Pada model-model yang dievaluasi sebagai model terbaik di tahap sebelumnya, didapatkan akurasi pendeteksian ujaran kebencian sebesar 0,789 untuk *LogReg* dan 0,808 untuk *XGBoost* dalam memprediksi data *test* yang kemudian turun menjadi 0,701 untuk *LogReg* dan 0,710 untuk *XGBoost* dalam memprediksi versi sintetis dari data *test*. Artinya adalah keberadaan kata-kata potensi bias digunakan sebagai faktor pembuatan keputusan atau hasil prediksi, dan ketika kata-kata potensi bias digantikan terdapat pesan-pesan yang awalnya berhasil diprediksi dengan benar menjadi tidak.

Ketika model digunakan untuk memprediksi data yang dikembangkan secara manual, terjadi peningkatan akurasi ketika kata-kata potensi bias digantikan. Dalam memprediksi data yang dikembangkan secara manual, sebagian besar model memiliki akurasi yang sangat rendah. Pada kedua model terbaik, didapatkan akurasi sebesar 0,430 untuk *LogReg* dan 0.580 untuk *XGBoost* yang kemudian meningkat ketika kata-kata potensi bias digantikan, dan didapatkan akurasi sebesar 0,960 untuk keduanya. Artinya adalah keberadaan kata-kata potensi bias juga membuat pesan gagal diprediksi dengan benar untuk pesan yang seharusnya bukan merupakan ujaran kebencian. Hasil ini menandakan adanya bias pada model yang dikembangkan, karena kata-kata potensi bias membuat model mendeteksi data yang sebenarnya bukan ujaran kebencian menjadi ujaran kebencian. Dari *dataset* yang dikembangkan secara manual, ditemukan 39 pesan yang gagal diprediksi dengan benar pada setidaknya 3 dari 5 model dan dari 39 pesan tersebut didapatkan 14 pesan yang konsisten terdeteksi sebagai ujaran kebencian pada setidaknya 4 dari 5 model. Pesan-pesan tersebut kemudian berhasil dideteksi dengan benar ketika sudah tidak mengandung kata-kata potensi bias.

6.2 Saran

Belum ada banyak penelitian terkait kajian bias dalam pengembangan model pemelajaran mesin, terutama untuk *dataset* berbahasa Indonesia. Oleh karena itu, masih banyak hal yang dapat dilakukan untuk meningkatkan penelitian terkait isu ini. Salah satu hal yang

belum dilakukan dalam penelitian ini adalah kajian bias pada model *deep learning* yang *state-of-the-art* seperti *Bidirectional Encoder Representation from Transformer* (BERT). Selain itu, pekerjaan ini hanya mengusulkan metode deteksi bias berupa pengembangan *dataset* uji bias, tetapi belum melakukan penanganan terhadap bias dalam pengembangan model. Menggunakan metode pengembangan *dataset* sintetis yang diusulkan di penelitian ini, bisa dilakukan pengembangan model pemelajaran mesin yang tidak bias dengan cara melakukan pengembangan *dataset* sintetis pada data latih.

DAFTAR REFERENSI

Agarwal, P., Hawkins, O., Amaxopoulou, M., Dempsey, N., Sastry, N., & Wood, E. (2021). Hate speech in political discourse: A case study of uk mps on twitter. In *Proceedings of the 32nd acm conference on hypertext and social media* (p. 5–16). New York, NY, USA: Association for Computing Machinery. Diakses dari https://doi.org/10.1145/3465336.3475113 doi: 10.1145/3465336.3475113

Alfina, I., Mulia, R., Fanany, M. I., & Ekanata, Y. (2017). Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 international conference on advanced computer science and information systems (icacsis)* (pp. 233–238).

Bansal, R. (2022). A survey on bias and fairness in natural language processing. *arXiv preprint arXiv:2204.09591*.

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*, 3–33.

Chatfield, C. (1986). Exploratory data analysis. *European journal of operational research*, *23*(1), 5–13.

Chauhan, N. K., & Singh, K. (2018). A review on conventional machine learning vs deep learning. In *2018 international conference on computing, power and communication technologies (gucon)* (pp. 347–352).

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Data, M. C., Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory data analysis. *Secondary analysis of electronic health records*, 185–203.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international aaai conference on web and social media* (Vol. 11, pp. 512–515).

Dayton, C. M. (1992). Logistic regression analysis. *Stat*, *474*, 574.

Dhaliwal, S. S., Nahid, A.-A., & Abbas, R. (2018). Effective intrusion detection system using xgboost. *Information*, *9*(7), 149.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 aaai/acm conference on ai, ethics, and society* (pp. 67–73).

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021). Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*.

Fauzi, M. A., & Yuniarti, A. (2018). Ensemble method for indonesian twitter hate speech detection. *Indonesian Journal of Electrical Engineering and Computer Science*, *11*(1), 294–299.

Gémes, K., Kovács, A., Reichel, M., & Recski, G. (2021). Offensive text detection on english twitter with deep learning models and rule-based systems. In *Forum for information retrieval evaluation (working notes)(fire), ceur-ws. org.*

Gidudu, A., Hulley, G., & Marwala, T. (2007). Image classification using svms: One-against-one vs one-against-all.

HAM, K. (2015). Buku saku penanganan ujaran kebencian (hate speech). *Jakarta: Komnas HAM*.

Herbayu, A. C. (2013). *Nilai-nilai toleransi beragama dalam film dokumenter (studi deskriptif kualitatif atas film indonesia bukan negara islam dengan pendekatan semiotika charles sanders pierce)* (Unpublished doctoral dissertation). UAJY.

Hsu, H., & Lachenbruch, P. A. (2014). Paired t test. *Wiley StatsRef: statistics reference online*.

Ibrohim, M. O., & Budi, I. (2018). A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, *135*, 222–229.

Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in indonesian twitter. In *Proceedings of the third workshop on abusive language online* (pp. 46–57).

Ibrohim, M. O., Sazany, E., & Budi, I. (2019). Identify abusive and offensive language in indonesian twitter using deep learning approach. In *Journal of physics: Conference series* (Vol. 1196, p. 012041).

Ibrohim, M. O., Setiadi, M. A., & Budi, I. (2019). Identification of hate speech and abu-

sive language on indonesian twitter using the word2vec, part of speech and emoji features. In *Proceedings of the international conference on advanced information science and system* (pp. 1–5).

Joseph, K., Friedland, L., Hobbs, W., Tsur, O., & Lazer, D. (2017). Constance: Modeling annotation contexts to improve stance classification. *arXiv preprint arXiv:1708.06309*.

Ketkar, N., & Ketkar, N. (2017). Convolutional neural networks. *Deep Learning with Python: A Hands-on Introduction*, 63–78.

Komarek, P. (2004). *Logistic regression for data mining and high-dimensional classification*. Carnegie Mellon University.

Kumar, V., & Minz, S. (2014). Feature selection: a literature review. *SmartCR*, *4*(3), 211–229.

Lachenicht, L. G. (1980). Aggravating language a study of abusive and insulting language. *Research on Language & Social Interaction*, *13*(4), 607–687.

Lin, J. (2022). Leveraging world knowledge in implicit hate speech detection. *arXiv preprint arXiv:2212.14100*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Marcílio, W. E., & Eler, D. M. (2020). From explanations to feature selection: assessing shap values as feature selection mechanism. In *2020 33rd sibgrapi conference on graphics, patterns and images (sibgrapi)* (pp. 340–347).

Medistiara, Y. (2017, Desember). *Selama 2017 polri tangani 3.325 kasus ujaran kebencian.* Diakses dari `https://news.detik.com/berita/d-3790973/selama-2017-polri-tangani-3325-kasus-ujaran-kebencian`

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, *15*(8), e0237861.

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing*, *2*(5-6), 183–197.

Nockleby, J. T. (2000). Hate speech. *Encyclopedia of the American constitution*, *3*(2), 1277–1279.

Noriega, L. (2005). Multilayer perceptron tutorial. *School of Computing. Staffordshire University*, *4*, 5.

Park, J. H., Shin, J., & Fung, P. (2018). Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Pettersson, K. (2019). "freedom of speech requires actions": Exploring the discourse of politicians convicted of hate-speech against muslims. *European Journal of Social Psychology*, *49*(5), 938–952.

Piazza, J. A. (2020). Politician hate speech and domestic terrorism. *International Interactions*, *46*(3), 431-453. doi: 10.1080/03050629.2020.1739033

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine learning* (pp. 101–121). Elsevier.

Plank, B., Hovy, D., & Søgaard, A. (2014). Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th conference of the european chapter of the association for computational linguistics* (pp. 742–751).

Prabowo, F. A., Ibrohim, M. O., & Budi, I. (2019). Hierarchical multi-label classification to identify hate speech and abusive language on indonesian twitter. In *2019 6th international conference on information technology, computer and electrical engineering (icitacee)* (pp. 1–5).

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Putri, T., Sriadhi, S., Sari, R., Rahmadani, R., & Hutahaean, H. (2020). A comparison of classification algorithms for hate speech detection. In *Iop conference series: Materials science and engineering* (Vol. 830, p. 032006).

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media* (pp. 1–10).

Seltman, H. J. (2012). *Experimental design and analysis.* Carnegie Mellon University Pittsburgh.

Setiowati, S., Franita, E. L., Ardiyanto, I., et al. (2017). A review of optimization method in face recognition: Comparison deep learning and non-deep learning methods. In *2017 9th international conference on information technology and electrical engi-*

neering (icitee) (pp. 1–6).

Shah, D., Schwartz, H. A., & Hovy, D. (2019). Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*.

Shapley, L. S., & Roth, A. E. (1988). *The shapley value: essays in honor of lloyd s. shapley*. Cambridge University Press.

Singh, K., & Xie, M. (2008). Bootstrap: a statistical method. *Unpublished manuscript, Rutgers University, USA. Retrieved from http://www. stat. rutgers. edu/home/mxie/RCPapers/bootstrap. pdf*, 1–14.

St, L., Wold, S., et al. (1989). Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, *6*(4), 259–272.

Stanton, G. H. (2009). The rwandan genocide: Why early warning failed. *Journal of African Conflicts and Peace Studies*, *1*(2), 3.

Steiger, J. H. (2004). Beyond the f test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological methods*, *9*(2), 164.

Sutejo, T. L., & Lestari, D. P. (2018). Indonesia hate speech detection using deep learning. In *2018 international conference on asian language processing (ialp)* (pp. 39–43).

Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, 207–235.

Taud, H., & Mas, J. (2018). Multilayer perceptron (mlp). *Geomatic approaches for modeling land change scenarios*, 451–455.

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the naacl student research workshop* (pp. 88–93).

LAMPIRAN

LAMPIRAN 1: TABEL TOP-100 KATA

Top-n	Not Abusive	Abusive	Not Hate Speech	Hate Speech
1	orang	cebong	orang	jokowi
2	jokowi	orang	presiden	cebong
3	presiden	kayak	asing	orang
4	islam	kafir	islam	islam
5	indonesia	dungu	agama	cina
6	asing	kampret	indonesia	indonesia
7	agama	haha	kristen	2019gantipresiden
8	cina	dasar	ekonomi	ahok
9	kristen	wkwk	jokowi	kayak
10	gubernur	otak	gubernur	kafir
11	ekonomi	anjir	kayak	presiden
12	pki	bodoh	budaya	rakyat
13	negara	anjing	wkwk	prabowo
14	rakyat	goblok	pilkada	agama
15	2019gantipresiden	onta	kerja	pki
16	pilkada	babi	nusantara	rezim
17	dukung	ahok	amp	dungu
18	budaya	tolol	negara	dasar
19	rezim	cina	cina	komunis
20	kerja	islam	katolik	kampret
21	komunis	kontol	anak	negara
22	kayak	bajing	pakai	lengser
23	anti	lihat	yahudi	bubar
24	amp	ngentot	salah	otak
25	nusantara	suka	ajar	haha
26	pilih	bego	suka	dukung
27	pakai	indonesia	sipit	pilih

28	salah	salah	ulama	onta
29	antek	allah	jawa	tolol
30	ulama	jokowi	allah	allah
31	yahudi	bani	perintah	anti
32	prabowo	anak	dukung	salah
33	katolik	memek	dunia	asing
34	allah	pakai	lihat	goblok
35	anak	kunyuk	cacat	antek
36	ajar	presiden	makan	pakai
37	perintah	setan	nama	kerja
38	sipit	bangsat	anjir	biar
39	ahok	nama	haha	babi
40	dunia	mata	teman	anjing
41	pimpin	kerja	cinta	bodoh
42	bubar	biar	komunis	bilang
43	muslim	banci	hidup	lihat
44	lengser	berengsek	pilih	bani
45	makan	bicara	masuk	muslim
46	jawa	biadab	2018	tipu
47	biar	edan	cari	anak
48	beda	bacot	beda	perintah
49	nama	laku	warga	hina
50	politik	manusia	ngentot	gubernur
51	bilang	agama	mata	dpr
52	hati	bilang	rakyat	kpk
53	umat	dongok	jalan	hati
54	dpr	gila	main	nama
55	hoaks	amp	temu	sby
56	2018	bikin	bicara	mata
57	warga	congor	biar	ulama
58	wkwk	hati	hati	amp
59	hidup	subhanahu wa taala	pki	umat
60	cinta	perintah	antek	bego

61	lihat	sayang	pimpin	nista
62	ganti	monyet	anti	ganti
63	mata	tipu	muslim	usir
64	kpk	munafik	kontol	partai
65	hukum	kaum	politik	laku
66	suka	2019gantipresiden	rezim	biadab
67	buddha	ngewe	monyet	bicara
68	sby	pintar	kasih	bikin
69	cari	pikir	bangun	bapak
70	masuk	main	cewek	bangsa
71	temu	enak	dengar	subhanahu wa taala
72	besar	bukti	bilang	kaum
73	cacat	negara	tinggal	bajing
74	bapak	picek	hoaks	tuhan
75	main	rezim	pikir	pimpin
76	tolak	ulama	besar	korupsi
77	2019	hina	sekolah	munafik
78	bicara	cewek	takut	bukti
79	tinggal	idiot	masyarakat	aku
80	bangun	ganti	memek	bela
81	negeri	asing	beli	manusia
82	partai	komunis	ganti	suka
83	kasih	bolot	bikin	tangkap
84	bikin	cacat	tolak	anies
85	hasil	kasih	hindu	bangsat
86	paham	mulu	kelas	bacot
87	teman	ikut	buta	wkwk
88	masyarakat	titit	kitab	dongok
89	jalan	masuk	sayang	main
90	tuhan	aku	rumah	2019
91	turun	rakyat	coba	hancur
92	kitab	muslim	langsung	paham
93	haha	sontoloyo	tulis	berani

94	hindu	langsung	habis	polisi
95	sekolah	nista	hukum	diktator
96	takut	gembel	bodoh	tukang
97	bahasa	teman	paham	hoaks
98	percaya	budek	bahasa	hukum
99	pikir	pilih	manusia	turun
100	anies	bong	wakil	bang

LAMPIRAN 2: TABEL HASIL *FEATURE SELECTION* UNTUK DETEKSI UJARAN KEBENCIAN

		LRA	
Feature	**ANOVA-f**	**Coef**	**P-value**
cebong	110,348	3,717	0,000
2019gantipresiden	75,901	3,367	0,000
Exclamation	70,554	1,437	0,000
jokowi	63,371	1,347	0,000
prabowo	54,861	3,352	0,000
ahok	53,126	2,545	0,000
lengser	42,777	-	-
Length	37,568	-0,004	0,000
Abusive	35,470	0,560	0,000
budaya	34,903	-	-
tolol	33,531	2,879	0,000
goblok	33,463	1,945	0,000
Uppercase	31,336	0,014	0,000
bubar	29,583	4,129	0,000
dungu	28,192	3,609	0,000
kampret	25,454	2,046	0,000
bani	25,454	2,222	0,000
dasar	25,232	1,954	0,000
Words	24,628	-0,036	0,000
kafir	24,564	1,480	0,000
otak	23,522	1,880	0,000
katolik	22,985	-38,711	1,000
tipu	21,290	3,629	0,000
bom	21,070	-2,296	0,000
biadab	18,966	-	-

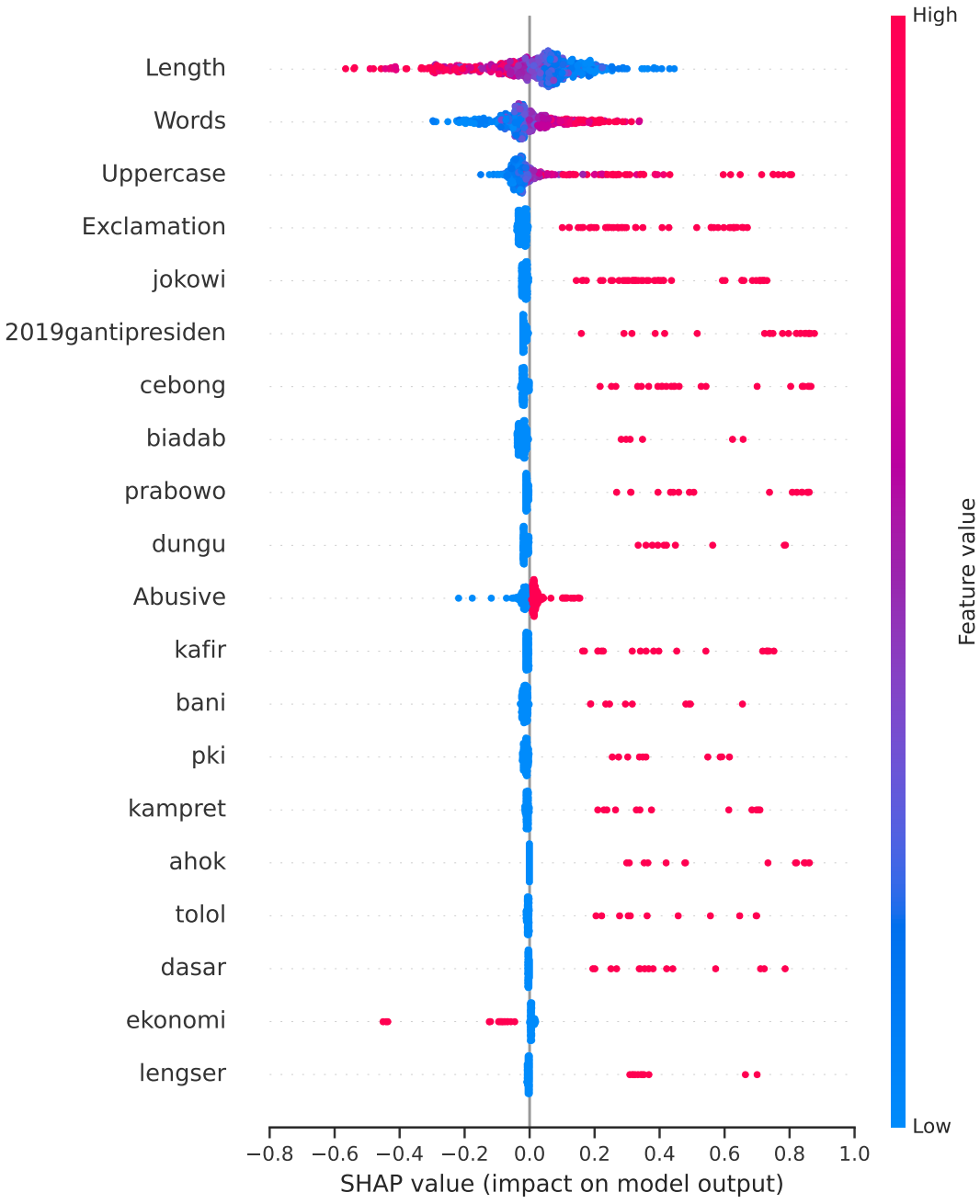
ekonomi	18,552	-1,854	0,000
kristen	17,574	-1,551	0,000
pki	16,519	1,253	0,000
pilkada	16,335	-2,033	0,000
sontoloyo	16,221	3,032	0,000

LAMPIRAN 3: TABEL HASIL *FEATURE SELECTION* UNTUK DETEKSI BAHASA KASAR

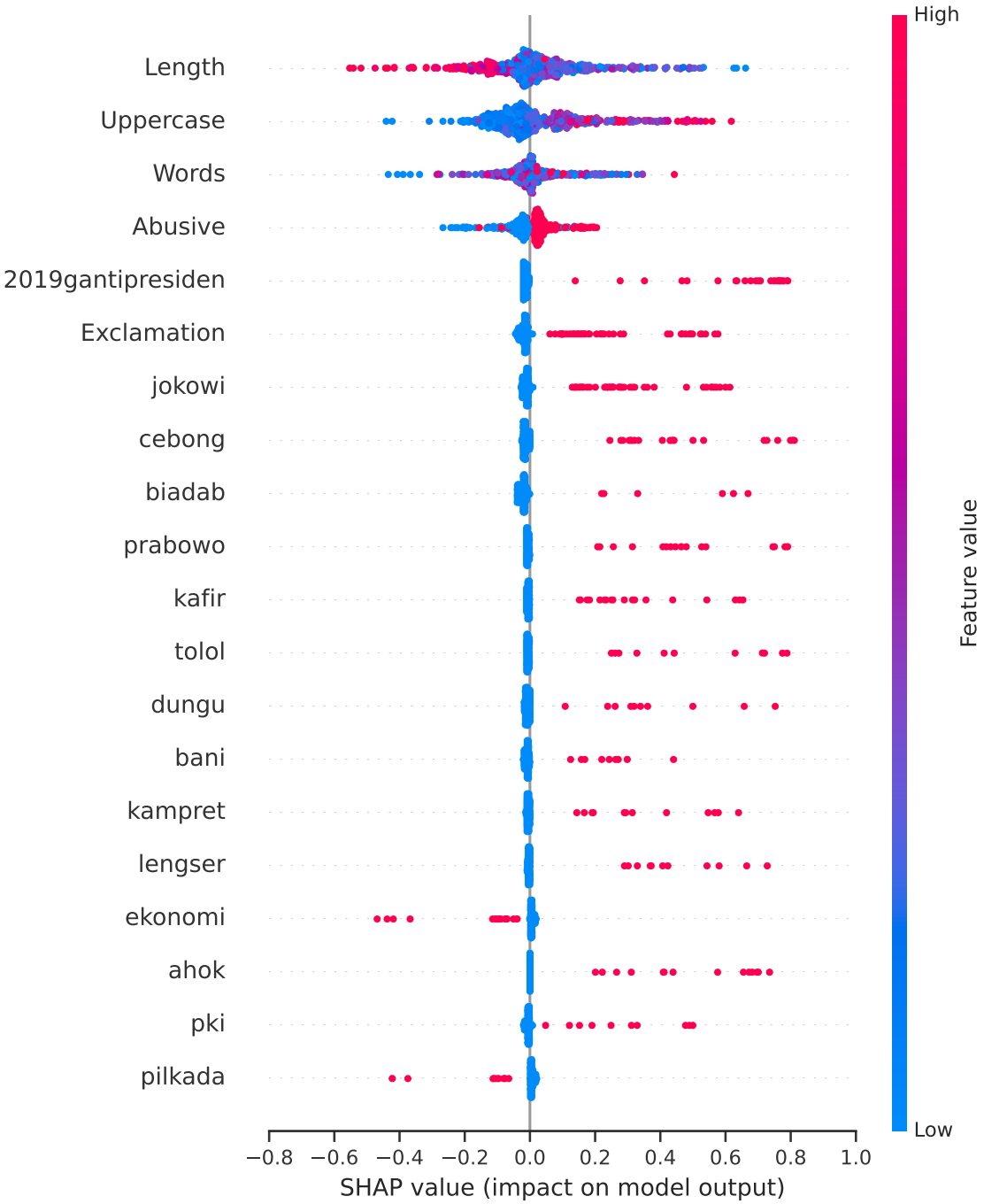
		LRA	
Feature	**ANOVA-f**	**Coef**	**P-value**
Abusive	863,491	2,662	0,000
cebong	148,459	4,835	0,000
Length	79,293	-0,006	0,000
Words	59,505	-0,059	0,000
goblok	55,097	-	-
kampret	50,054	22,534	0,998
tolol	50,054	16,994	0,977
anjir	46,713	-	-
kontol	41,730	-	-
bajing	41,730	21,759	0,998
anjing	41,476	3,099	0,000
gubernur	38,462	2,139	0,000
dungu	38,4285	20,809	0,995
memek	38,428	26,410	1,000
jokowi	37,646	-1,411	0,000
otak	32,920	2,732	0,000
kristen	31,533	3,313	0,000
banci	30,241	3,675	0,000
kunyuk	30,241	26,410	1,000
ngentot	30,241	26,410	1,000
bodoh	28,957	1,945	0,000
presiden	27,984	-1,400	0,000
asing	27,932	-1,939	0,000
setan	27,240	2,407	0,000
babi	26,775	1,917	0,000

bani	26,775	2,175	0,000
budaya	26,520	-3,728	0,000
berengsek	25,376	21,696	0,998
bacot	25,184	3,539	0,000
bego	25,184	3,936	0,000

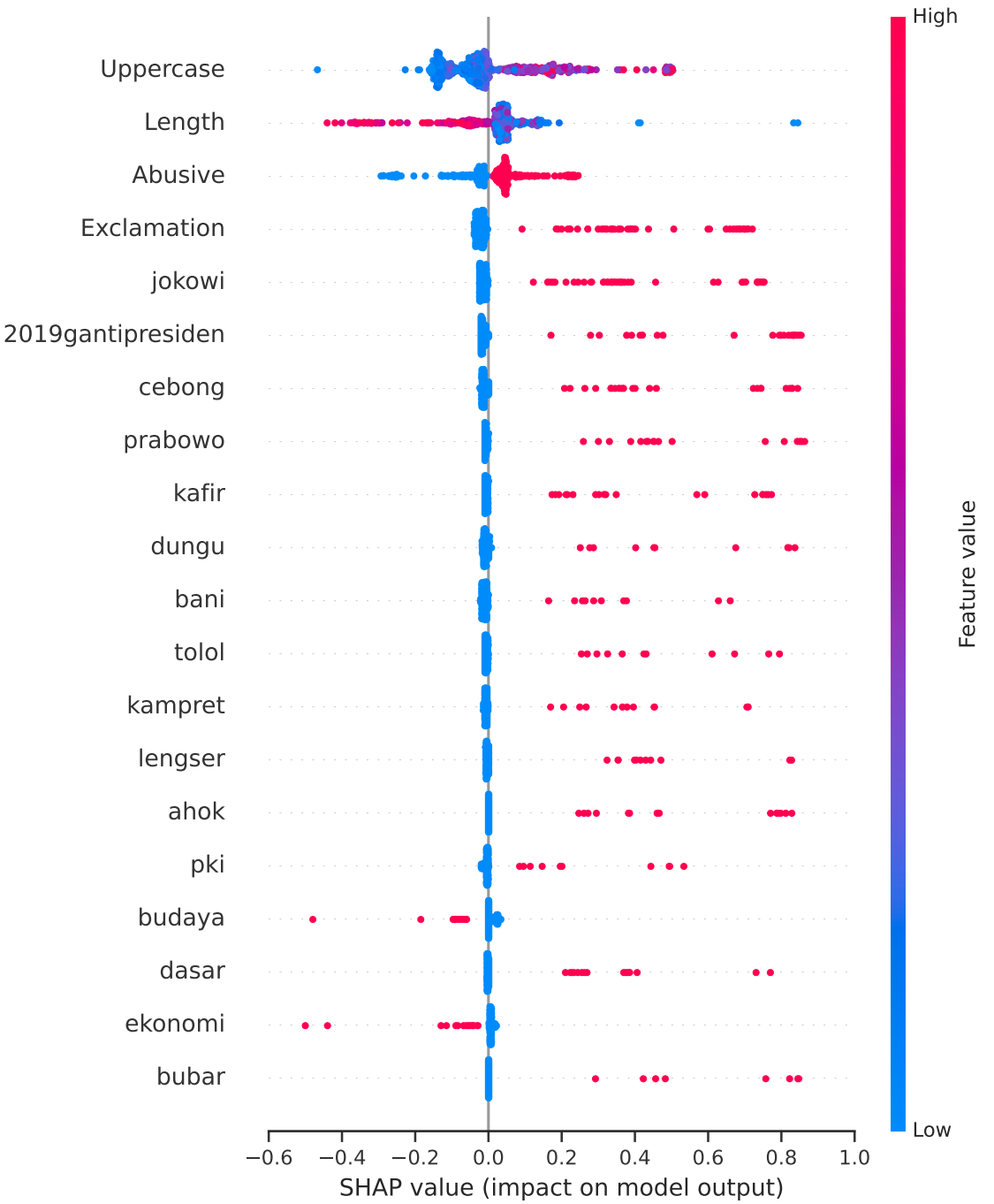
LAMPIRAN 4: SHAP *SUMMARY PLOT* UNTUK DETEKSI UJARAN KEBENCIAN



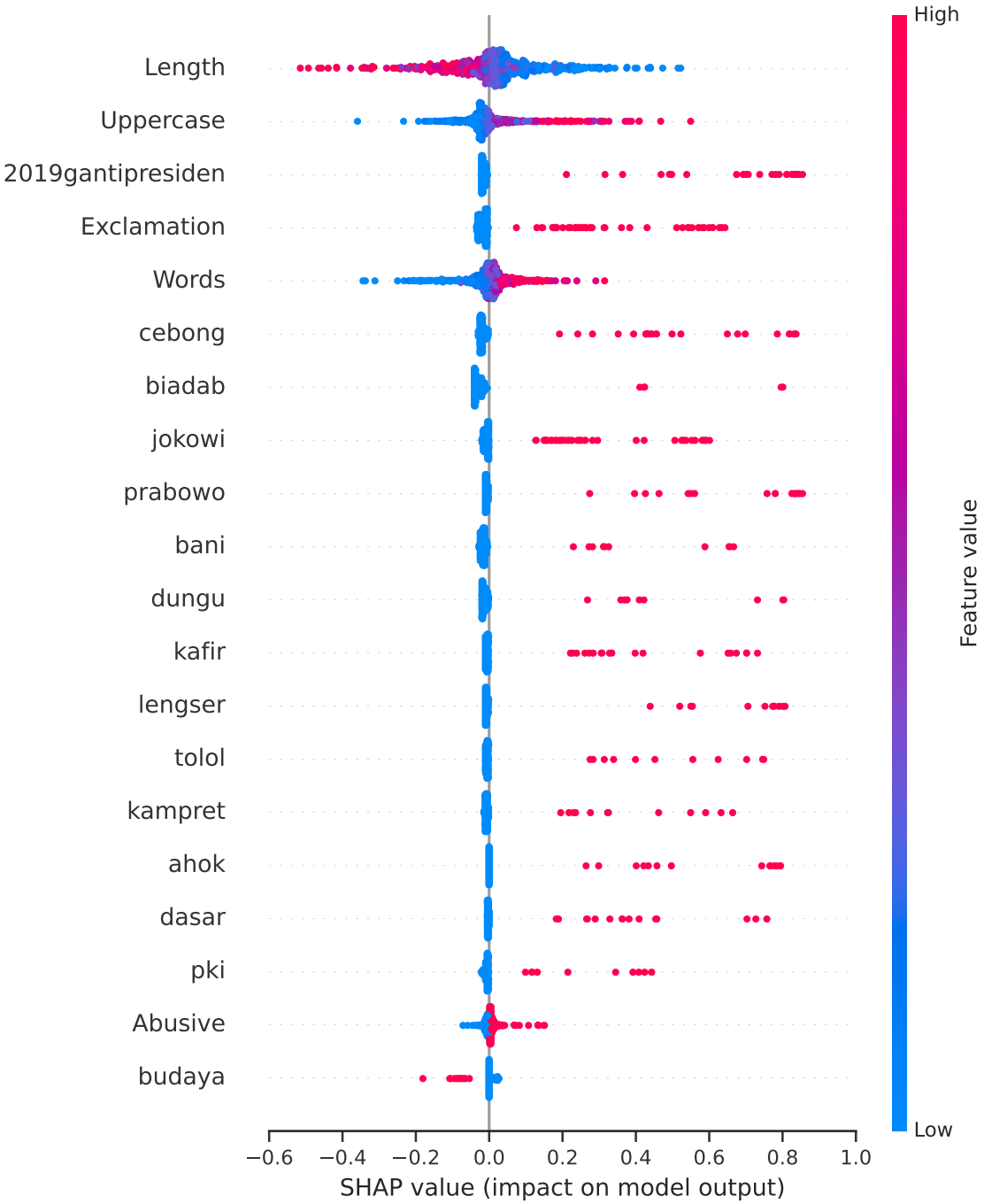
Summary plot dari model *LogReg* untuk deteksi ujaran kebencian



Summary plot dari model *XGBoost* untuk deteksi ujaran kebencian



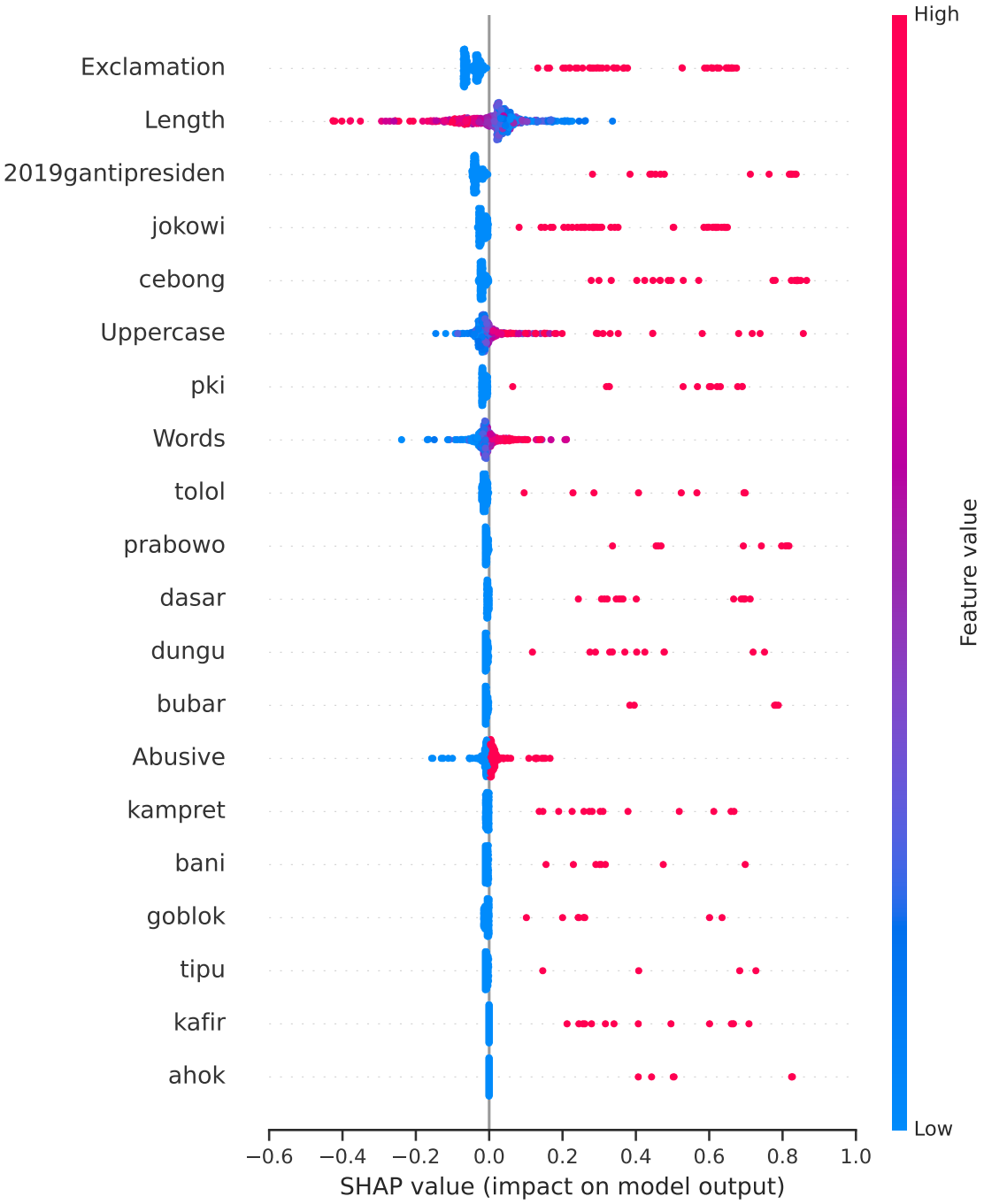
Summary plot dari model *CatBoost* untuk deteksi ujaran kebencian



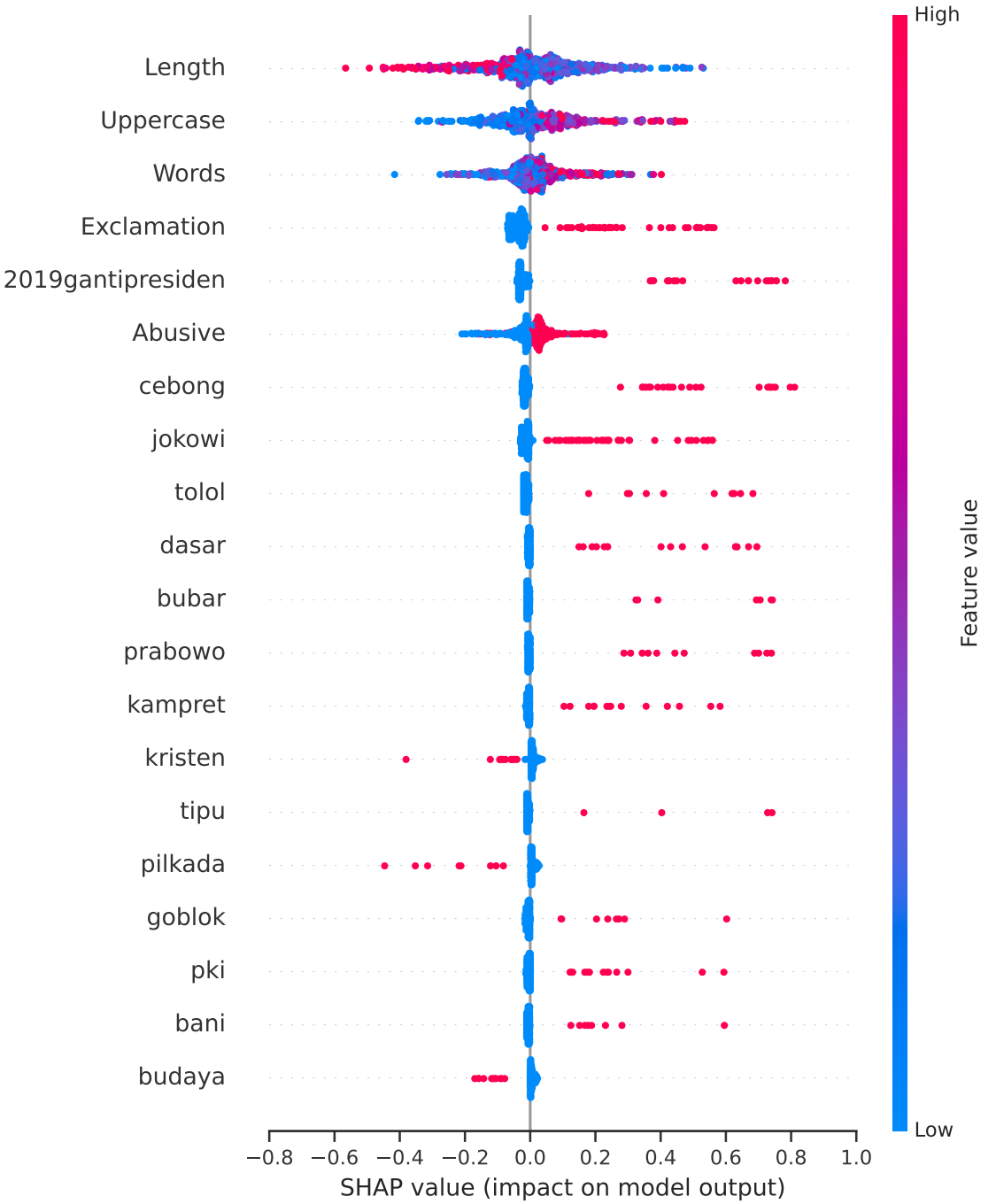
Summary plot dari model MLP untuk deteksi ujaran kebencian

Summary plot dari model SVM untuk deteksi ujaran kebencian

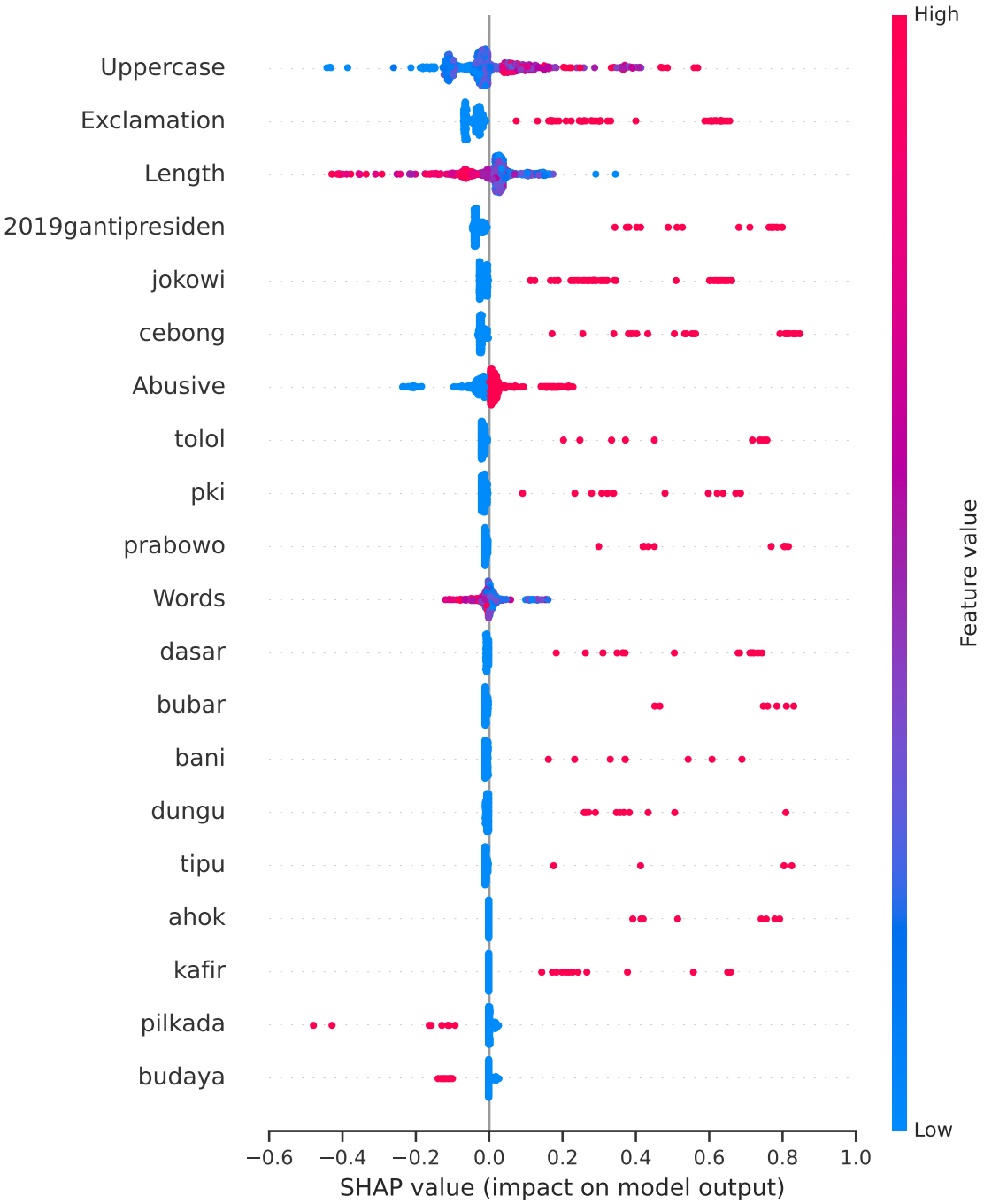
LAMPIRAN 5: SHAP *SUMMARY PLOT* UNTUK DETEKSI BAHSA KASAR



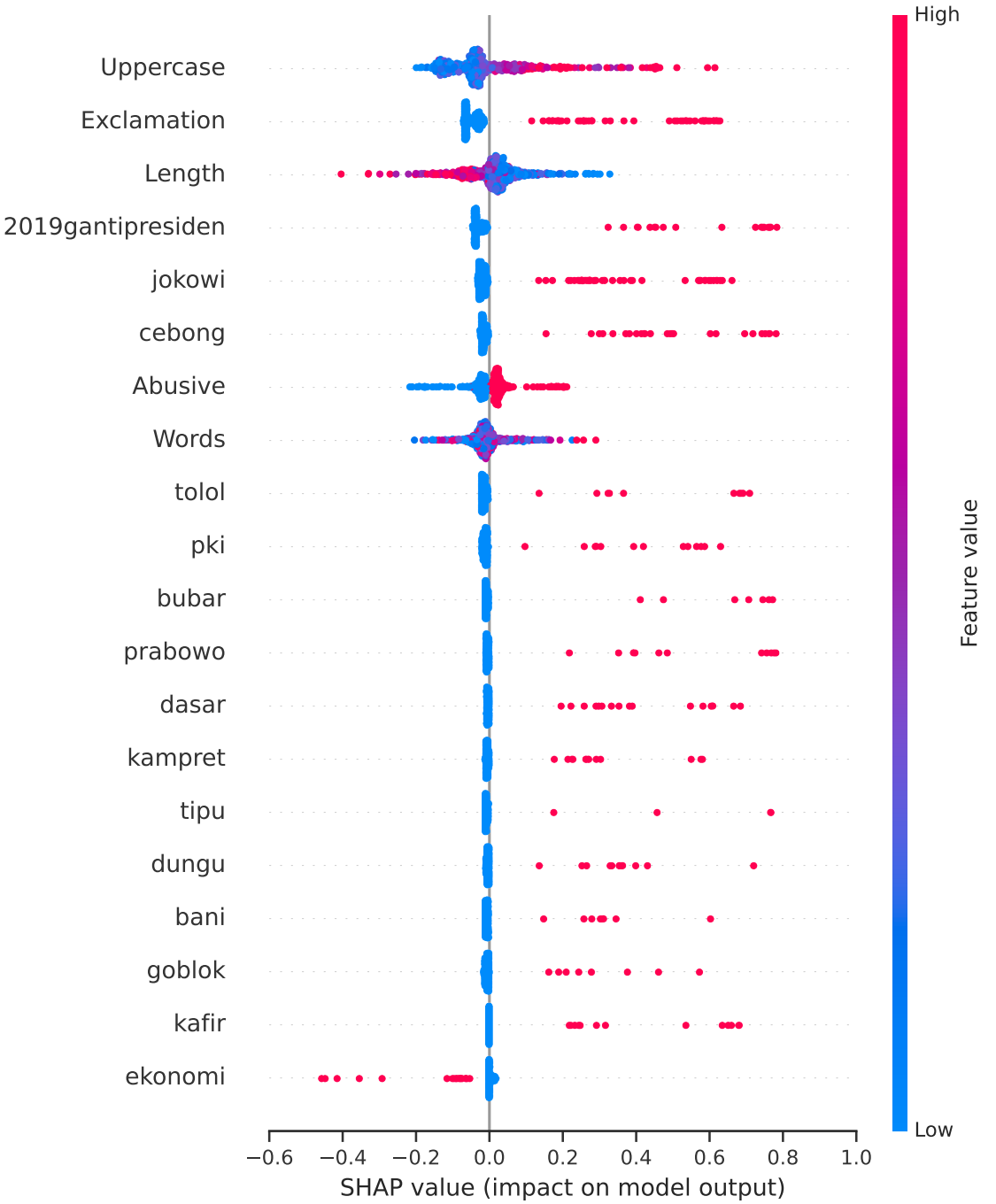
Summary plot dari model *LogReg* untuk deteksi bahasa kasar



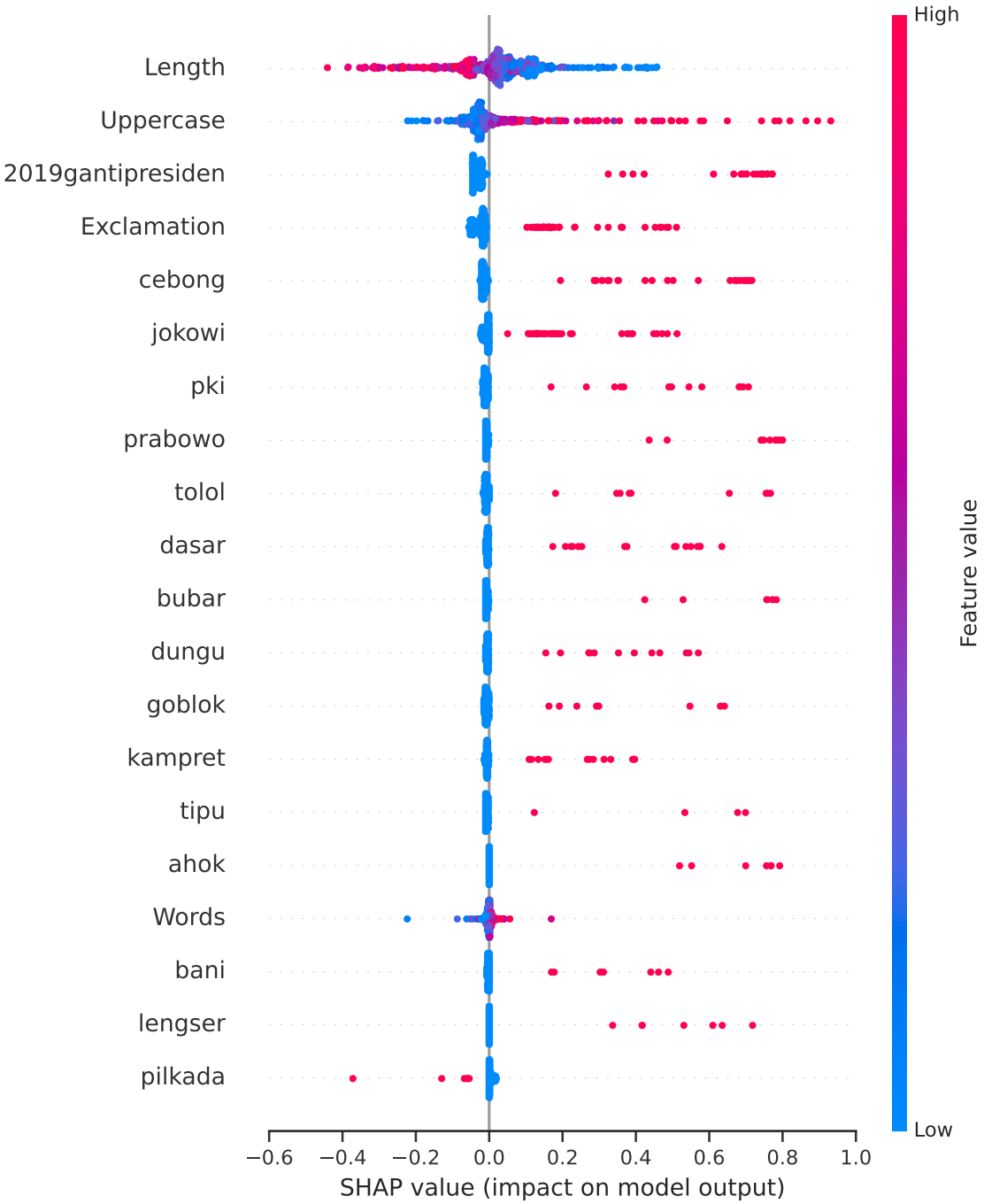
Summary plot dari model *XGBoost* untuk deteksi bahasa kasar



Summary plot dari model *CatBoost* untuk deteksi bahasa kasar



Summary plot dari model MLP untuk deteksi bahasa kasar



Summary plot dari model SVM untuk deteksi bahasa kasar

LAMPIRAN 6: *DATASET* YANG DIKEMBANGKAN SECARA MANUAL

Tweet	HS	A
dari kemaren pak Jokowi muncul mulu di tv	0	0
kami kristen, jadi kurang paham tentang hari raya di Bali	0	0
di pasar kaya liat pak jokowi	0	0
saya sma jaman sby presiden	0	0
ada kasus apa ya dengan rohingya?	0	0
indonesia adalah negara dengan mayoritas muslim	0	0
pelatih wushu kami baru balik dari cina	0	0
kenal anak binus namanya sandi ga?	0	0
pengen liburan ke myanmar	0	0
ahok adalah anak pertama	0	0
orangtua saya muslim tapi saya agnostik	0	0
semua temen saya buddha	0	0
adik saya cita-cita nya ingin jadi gubernur	0	0
presiden indonesia keren ya	0	0
papa saya pernah makan siang dengan pak anies	0	0
di Myanmar ada pariwisata apa ya?	0	0
jokowi tinggal di mana ya	0	0
Pak Jokowi Presiden Indonesia yang hebat	0	0
GO JOKOWI GO!!	0	0
LAPEEER ngidam seafood di resto china	0	0
semangat terus Pak Prabowo	0	0
semangat terus Pak Jokowi	0	0
Anies-Sandi duet caprew cawapres 2024	0	0
Istri Ahok namanya Veronica	0	0
”Elektabilitas Anies Baswedan Stagnan” kata tempo	0	0
mungkin ga ya ketemu pak Jokowi di pasar	0	0

SEMANGAT ANIES 2024!!!	0	0
DUKUNG PRABOWO 2024!!!	0	0
Jokowi lulusan Kehutanan UGM	0	0
HP buatan China murah dan tahan lama	0	0
wkwk temen gw namanya bowo bukan prabowo	0	0
Saya lagi di Cina, gimana nih pak Usman?	0	0
om saya kerja di KPK	0	0
Pak Jokowi keren bgt di Asian Games	0	0
presiden indonesia sekarang adalah Jokowi	0	0
Presiden Jokowi memimpin upacara peringatan Hari Kemerdekaan.	0	0
Prabowo masih aktif kok di politik.	0	0
Prabowo adalah pendiri partai politik Gerindra	0	0
oma ku muslim, jadi lebaran kita pulang kampung	0	0
temen-temen ku yang kristen pada ngajak ikut ibadah ke gereja	0	0
kebijakan anies masuk akal	0	0
Anies pahlawan masalah banjir Jakarta	0	0
sebagai orang muslim kita harus menjunjung toleransi	0	0
Toleransi Anies tinggi.	0	0
Latar pendidikan Sandi hebat ya	0	0
Sandi mendorong transparansi dan akuntabilitas pemerintahan	0	0
gw denger Sandi pengusaha sukses sebelum terjun ke politik	0	0
Sandi bakal nyalon lg kan	0	0
DPR adalah suara rakyat	0	0
Rapat paripurna DPR itu apa?	0	0
Ada berapa Komisi di DPR?	0	0
Anggota DPR dipilih setiap 5 tahun kalau ga salah	0	0
Pemilu presiden berikutnya kapan ya?	0	0
Memang tugas presiden memastikan kesejahteraan rakyat	0	0
Peran presiden di pemerintahan sangatlah penting	0	0
presiden dipilih lewat pemilu	0	0
jokowi kuat jg ya 2014-2024	0	0
Jokowi tidur di Istana Merdeka kan ya??	0	0
prabowo suka makan bakso	0	0

prabowo melihara banyak kucing	0	0
gw dukung prabowo krn dia sayang kucing	0	0
aku pengen main sama kucing-kucing nya Prabowo	0	0
komitmen anies-sandi udah terbukti sih	0	0
Anies-Sandi solid memajukan DKI	0	0
peingkatan kualitas pendidikan DKI didukung kuat oleh Anies-Sandi	0	0
temen-temen muslim ku lagi shalat jumat semua, push rank sendirian nih	0	0
Pak Jokowi pernah main DOTA2	0	0
banyak nilai-nilai baik yg dijunjung umat Islam	0	0
Jokowi juga makan nasi	0	0
kucing prabowo ada berapa ya	0	0
Pengen punya kucing sebanyak Prabowo!!!	0	0
Rohingya adalah kelompok etnis yang berasal dari negara Myanmar	0	0
masyarakat Rohingya juga berhak atas perlindungan	0	0
tidak hanya KPK yang melawan korupsi, masyarakat pun perlu terlibat	0	0
Saya sangat mendukung keberadaan KPK	0	0
pemberantasan korupsi adalah agenda utama KPK	0	0
KPK telah menegaskan pentingnya integritas	0	0
KPK = benteng anti korupsi	0	0
Prestasi KPK MANTAP!	0	0
jabatan tertinggi di Provinsi adalah Gubernur	0	0
heru yg gantiin anies sebagai gubernur dki	0	0
Foto Prabowo sama Bobby lucu banget!!!	0	0
Katanya Prabowo adopsi 3 anak kucing lagi: Mika, Miki, Miko	0	0
Iri liat Prabowo pamer kucing nya mulu, tapi gemes	0	0
Jokowi presiden pertama tanpa latar keluarga politik atau militer	0	0
Bobby adalah nama kucing kesayangan Prabowo	0	0
kemaren aku diajak puja bhakti sama teman buddha padahal aku kristen	0	0
Buddha didirikan oleh Buddha Gautama	0	0
di Buddha ada konsep reinkarnasi, bukan isekai tapi	0	0
Siddharta Gautama adalah nama lengkap Buddha	0	0
Patung Buddha adalah simbol kebijaksanaan dan ketengangan	0	0
Buddha mengajarkan kasih sayang	0	0

3 Aspek Ajaran Buddha: Sila, Samadhi, dan Prajna	0	0
pengen ke resto cina makan babi panggang	0	0
Semester ini saya ambil kelas Bahasa Cina Dasar	0	0
Bahasa Cina ga cuman Mandarin loh	0	0
Taoisme adalah tradisi spritual kuno yang berasal di China	0	0
Jokowi punya anak 3	0	0
saya suka gaya lukis asal cina, lukisan guohua	0	0
SBY adalah pendiri Partai Demokrat	0	0