IJSA (2021), Vol. 5, No. 1, 26-38
DOI: https://doi.org/10.29244/ijsa.v5i1p26-38

Indonesian Journal of Statistics and Its Applications
Available at https://ijsa.stats.id
e-ISSN: 2599-0802

Improving Classification Model Performances Using an Active Learning Method to Detect Hate Speech in Twitter*

Peningkatan Kinerja Model Klasifikasi dengan Pembelajaran Aktif dalam Mendeteksi Ujaran Kebencian di Twitter

Muhammad Ilham Maulady Abidin[1], Khairil Anwar Notodiputro[2‡], and Bagus Sartono[3]

[1]Department of Statistics, IPB University, Indonesia
[2]Department of Statistics, IPB University, Indonesia
[3]Department of Statistics, IPB University, Indonesia
[‡]corresponding author: khairil@apps.ipb.ac.id

Abstract

Efforts from the police to address hate speech on social media such as Twitter will not be sufficient to rely solely on manual checks. Therefore, it is necessary to use statistical modelling like the classification model to detect hate speech automatically. Classification is a type of predictive modelling to produce accurate predictions based on labelled data. Generally, the available data are usually unlabelled implying that the labelling process needs to be done beforehand. Data labelling is time consuming, high cost, and often fails to produce correct labels. This research aims to improve the performances of classification models by adding a small amount of data through the so called active learning method. The results showed that there was no significant difference in the performances of logistic regression and naïve bayes classification models in detecting hate speech. However, the results also showed that adding data through the active learning method substantially improved the logistics regression performance in detecting hate speech when compared to data addition based on a simple random sampling method. Therefore, the performances of classification models in detecting hate speech on Twitter could be improved by using an active learning method.

Keywords: active learning, classification models, hate speech, logistics regression, naïve bayes.

* Received: Aug 2020; Reviewed: Nov 2020; Published: Mar 2021

1. Pendahuluan

Indonesian Journal of Statistics and Its Applications: diterbitkan berkala 3 (tiga) kali dalam setahun yang memuat tulisan ilmiah yang berhubungan dengan bidang statistika dan aplikasinya. Artikel yang dimuat berupa hasil penelitian bidang statistika dan aplikasinya dengan topik (tapi tidak terbatas): rancangan dan analisis percobaan, metodologi survey dan analisis, riset operasi, data mining, pemodelan statistika, komputasi statistika, time series dan ekonometrika, serta pendidikan statistika.

Menurut Kuhn & Johnson (2013) pemodelan prediktif merupakan proses pengembangan alat prediksi berbasis model matematika yang dapat menghasilkan prediksi dengan akurasi tinggi. Model matematika seperti ini dikenal sebagai model predifktif. Pemodelan klasifikasi banyak digunakan di berbagai bidang, seperti klasifikasi *email* spam, sistem rekomendasi dari data historis pelanggan, dan pendeteksian penyakit yang diderita pasien. Dalam praktiknya, seringkali model matematika tidak cukup baik jika digunakan sebagai alat untuk memprediksi suatu kejadian atau klasifikasi. Hal tersebut dikarenakan model matematika bersifat deterministik, sementara itu hasil prediksi seringkali bersifat probabilistik atau tidak pasti. Karena alasan tersebut, dalam praktik pemodelan klasifikasi digunakan model statistika yang mengakomodir ketidakpastian dalam memprediksi suatu kejadian. Model klasifikasi statistika inilah yang dikaji di dalam penelitian ini.

Penyiapan data merupakan salah satu tantangan yang terdapat pada pemodelan klasifikasi. Penyiapan data ini dalam istilah *big data analytics* dikenal sebagai prapemrosesan data atau *data pre-processing* yang membutuhkan investasi waktu yang banyak dengan tingkat kecermatan yang tinggi. Selain itu, dalam pembuatan model klasifikasi dibutuhkan data latih yang memiliki label kelas peubah respon yang menjadi target penelitian. Bila gugus data tidak memiliki label, perlu dilakukan proses pelabelan pada gugus data tersebut. Namun, proses pelabelan tersebut tidak mudah karena terdapat kendala biaya, waktu, dan kerap muncul kesalahan pelabelan. Di sisi lain, data latih yang tersedia seringkali kurang representatif, sehingga model yang dibangun kurang tepat untuk menarik kesimpulan secara umum.

Masalah-masalah tersebut dapat diatasi dengan menggunakan *active learning*. *Active learning* merupakan proses untuk mendapatkan data tambahan dari data yang tidak memiliki label dengan cara memilih data yang paling informatif untuk dilabeli (Hu, 2011). Data yang dipilih merupakan data yang paling diragukan oleh algoritma pembelajaran, sehingga proses pelabelan akan lebih efektif dan efisien.

Salah satu kasus yang dapat diselesaikan oleh pemodelan klasifikasi adalah mendeteksi ujaran kebencian di media sosial. Menurut Sudut Hukum (2016) ujaran kebencian merupakan tindakan komunikasi yang dilakukan oleh individu atau kelompok dalam bentuk hasutan, provokasi, atau hinaan kepada individu atau kelompok dalam aspek ras, etnis, *gender*, disabilitas, agama, orientasi seksual, dan lain-lain. Terdapat banyak kasus ujaran kebencian di Indonesia. Selama tahun 2017 tercatat sebanyak 3325 kasus ujaran kebencian yang ditangani oleh pihak kepolisian. Angka tersebut meningkat dari tahun sebelumnya yang berjumlah 1829 kasus (Medistiara, 2017). Dalam praktiknya, prosedur penanganan kasus ujaran kebencian tidaklah mudah. Sebelum menangani kasus, pihak kepolisian perlu menentukan apakah suatu kasus dapat digolongkan sebagai kasus ujaran kebencian atau tidak Seiring dengan meningkatnya jumlah kasus, pemeriksaan laporan secara manual

memakan waktu dan biaya yang besar. Hal tersebut membuktikan bahwa perlu adanya pembuatan model prediktif yang dapat mendeteksi tulisan ujaran kebencian secara otomatis.

Data ujaran kebencian yang diperoleh dari Twitter belum memiliki keterangan label. Oleh karena itu, perlu dilakukan proses pelabelan terlebih dahulu. Data yang dikumpulkan dari berbagai macam *tweets* dikategorikan sebagai data besar. Hal tersebut dikarenakan jumlah data yang besar mengakibatkan waktu untuk prapemrosesan bertambah. Dikarenakan jumlah data yang besar, proses pelabelan data tidak dapat dilakukan untuk seluruh data yang ada. Oleh karena itu, perlu dilakukan penarikan contoh yang efektif agar gugus data yang dikumpulkan representatif walau jumlahnya minim.

Dalam penelitian ini dilakukan kajian kinerja model klasifikasi regresi logistik dan *naïve bayes* dalam mendeteksi ujaran kebencian di Twitter. Selanjutnya kinerja model klasifikasi yang terbaik diperbaiki menggunakan metode penambahan data dengan penarikan contoh acak sederhana dan *active learning*.

2. Metodologi

2.1 Data

Data yang digunakan pada penelitian ini adalah data hasil *scraping* dari media sosial Twitter. *Scraping* data dilakukan sejak tanggal 24 Oktober hingga 13 November 2019. Pengambilan data dilakukan menggunakan beberapa kata kunci yang dianggap memiliki *tweets* ujaran kebencian cukup banyak. Adapun contoh data yang didapatkan dapat dilihat pada Tabel 1.

Tabel 1: Contoh tweets yang diambil pada proses scraping.

No	*Tweets*
1	@942428653KLBG03 @OrdinaryEric @JKFC23456789 Tuh, kemaren2 hewan anj*ng bawa ke Masjid , karang kotoran. tp nuduh radikal tetep ke Islam. Heran,,
2	@Budiawan1985 @MCAOps Menghadapi gerombolan b*bi cina dan antek pki seperti lu, emang gk blh pake cara2 lembut.
3	@Nyongbae__ Apalagi anak kecil, karena gay BUKAN paedophile. Orientasi Seksual itu BERBEDA dengan Perilaku Seksual.
4	@54bd4_Palon @gopebagitiga @AnkerGear @Restichayah @aniesbaswedan Kalo gak g*blok bukan cebong namanya.
5	@mohmahfudmd Ini baru statement kelas mentri..., yg begini ini dibanyakin prof... bs bikin suasana adem, tolong jg yg koment hindari yg bersifat menghujat personal, kritik itu hrs substantial..., stop istilah kampret, cebong, d*ngu, kadal, radikal dll...

Data yang diperoleh dari proses *scraping* tidak dapat digunakan untuk pemodelan secara langsung. Pra-pemrosesan data, seperti pembersihan data, pembuatan *document-term matrix*, dan pelabelan perlu dilakukan terlebih dahulu untuk persiapan data. Tabel 2 menunjukkan data yang siap digunakan untuk pemodelan.

Tabel 2: Contoh tweets yang sudah dilakukan prapemrosesan.

Dokumen	Kata							Label
	kemarin	hewan	anj*ng	bawa	...	cebong	radikal	
1	2	1	1	1	...	0	1	Non Hate
2	0	0	0	0	...	0	0	Hate
3	0	0	0	0	...	0	0	Non Hate
4	0	0	0	0	...	1	0	Hate
5	0	0	0	0	...	1	1	Non Hate

2.2 Model

Dalam penelitian ini dicobakan lima jenis model, yaitu regresi logistik (RL NO), regresi logistik dengan regularisasi gulud (RL GULUD), regresi logistik dengan regularisasi LASSO (RL LASSO), *multinomial naïve bayes* (NB NO), dan *multinomial naïve bayes* dengan pemulusan aditif (NB PEMULUSAN). Model tersebut digunakan karena model regresi logistik dan *naïve bayes* memiliki keluaran peluang duga kelas yang digunakan saat proses *active learning*. Model pertama yang digunakan adalah regresi logistik. Bentuk model regresi logistik yang dibangun adalah sebagai berikut:

$$P(y_i = 1) = \pi_i = \frac{\exp(\beta_0 + \boldsymbol{x_i}^T\boldsymbol{\beta})}{1 + \exp(\beta_0 + \boldsymbol{x_i}^T\boldsymbol{\beta})} \quad (1)$$

Y_i merupakan kelas dari peubah respon dokumen ke-i, dimana $y_i = 1$ untuk dokumen yang terdapat ujaran kebencian (*Hate*) dan $y_i = 0$ untuk untuk dokumen yang tidak terdapat ujaran kebencian *(Non Hate)*. $P(y_i = 1)$ atau π_i merupakan peluang dokumen ke-i masuk ke dalam kelas *Hate* (dokumen yang mengandung ujaran kebencian). $\boldsymbol{x_i}$ merupakan vektor berukuran $1 \times p$ (p adalah banyaknya kata yang terdapat pada seluruh dokumen) yang berisikan frekuensi kemunculan kata pada dokumen ke-i. β_0 merupakan nilai intersep regresi dan $\boldsymbol{\beta}$ merupakan vektor yang berisikan koefisien regresi. Untuk menduga parameter β_0 dan $\boldsymbol{\beta}$ pada model regresi logistik, digunakan pendugaan kemungkinan maksimum. Fungsi kemungkinan yang digunakan adalah sebagai berikut:

$$L(\beta_0, \boldsymbol{\beta}) = \prod_{i=1}^{n} [\pi_i{}^{y_i} \, (1 - \pi_i)^{1-y_i}] \quad (2)$$

Model berikutnya yang digunakan adalah regresi logistik dengan regularisasi *ridge* (gulud) dan *Least Absolute Shrinkage and Selection Operator* (LASSO). Alasan digunakannya kedua jenis regularisasi adalah kondisi data teks yang digunakan dalam penelitian memiliki dimensi yang sangat besar. Penggunaan regularisasi ditunjukan untuk mengurangi dimensi data, sehingga dapat mengatasi multikolinearitas dengan menyusutkan parameter $\boldsymbol{\beta}$. Model regresi logistik dengan regularisasi gulud dapat menyusutkan parameter $\boldsymbol{\beta}$ yang berkorelasi mendekati nol, sedangkan model regresi logistik dengan regularisasi LASSO dapat menyusutkan parameter $\boldsymbol{\beta}$ yang berkorelasi mendekati nol atau hingga menjadi nol (James et al., 2013). Fungsi log-kemungkinan model regresi logistik dengan regularisasi gulud terdapat pada persamaan (3) dan regularisasi LASSO terdapat pada persamaan (4) (Hastie et al., 2009).

$$l_\lambda^R(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \boldsymbol{x_i}^T\boldsymbol{\beta}) - ln\left(1 + exp\left(\beta_0 + \boldsymbol{x_i}^T\boldsymbol{\beta}\right)\right)\right] - \lambda \sum_{j=1}^{p} \beta_j^2 \quad (3)$$

$$l_{\lambda}^{L}(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \boldsymbol{x_i}^T \boldsymbol{\beta}) - ln\left(1 + exp\left(\beta_0 + \boldsymbol{x_i}^T \boldsymbol{\beta}\right)\right)\right] - \lambda \sum_{j=1}^{p} \left|\beta_j\right| \quad (4)$$

Nilai λ merupakan nilai penalti yang terdapat pada regularisasi gulud dan LASSO. Semakin besar nilai λ, parameter $\boldsymbol{\beta}$ pada model regresi logistik akan mendekati nol (untuk regularisasi gulud) atau menjadi nol (untuk regularisasi LASSO). Nilai λ optimum ditentukan menggunakan metode validasi silang, yaitu dengan cara membangkitkan beberapa nilai λ. Nilai λ yang dibangkitkan akan dicobakan pada validasi silang. Nilai λ optimum adalah nilai λ yang menghasilkan model dengan akurasi tertinggi pada metode validasi silang.

Model berikutnya yang digunakan adalah *multinomial naïve bayes*. Dalam *multinomial naïve bayes,* dokumen merupakan serangkaian kemunculan n buah kata x ke-i. Illustrasi dokumen dapat dilihat pada persamaan (5).

$$X = < x_1, \dots, x_i, \dots, x_{n_d} >,\ x_i \in V \quad (5)$$

V merupakan himpunan kata unik dalam seluruh dokumen dan n_d adalah jumlah kata unik yang terdapat pada suatu dokumen. Model *multinomial naïve bayes* memperhitungkan frekuensi kemunculan kata dalam dokumen. Fungsi peluang *multinomial naïve bayes* untuk kelas *Hate* adalah sebagai berikut:

$$P(y = Hate|X) \propto \prod_{i=1}^{n_d} P(x_i|y = Hate)P(y = Hate) \quad (6)$$

$\mathrm{P}(y = Hate|X)$ merupakan peluang dokumen *X* masuk ke dalam kelas *Hate*, $P(y = Hate)$ peluang kemunculan kelas *Hate* pada seluruh dokumen, $P(x_i|y = Hate)$ adalah peluang kata x ke-i muncul pada dokumen kelas *Hate*, dan p adalah jumlah kata unik pada dokumen. Pendugaan parameter $\hat{P}(y = Hate)$ dan $\hat{P}(x_i|y = Hate)$ adalah sebagai berikut:

$$\hat{P}(y = Hate) = \frac{N_{y=Hate}}{N} \quad (7)$$

$$\hat{P}(x_i|y = Hate) = \frac{N_{y=Hate;xi}}{\sum_{x \in V} N_{y=Hate;xi}} \quad (8)$$

$N_{y=Hate}$ adalah banyaknya dokumen latih pada kelas *Hate*, sedangkan N adalah total dokumen latih. $N_{y=Hate;x_i}$ adalah banyaknya dokumen yang mengandung kata x ke-i dalam dokumen latih kelas *Hate*.

Model terakhir yang digunakan adalah model *multinomial naïve bayes* dengan pemulusan aditif. Model *multinomial naïve bayes* dapat menghasilkan pendugaan $\hat{P}(x_i|y = Hate)$ yang bernilai nol akibat kata x ke-i yang terdapat di dokumen X tidak ada di himpunan V (Schütze et al., 2008). Masalah tersebut dapat diatasi dengan menggunakan *additive smoothing*, sehingga pendugaan $\hat{P}(x_i|y = Hate)$ menjadi:

$$\hat{P}(x_i|y = Hate) = \frac{N_{y=Hate;xi} + \alpha}{\sum_{x \in V} N_{y=Hate;xi} + \alpha|V|} \quad (9)$$

Nilai α merupakan nilai pemulusan yang terdapat pada model *naïve bayes*. Penentuan nilai α yang optimum dapat dilakukan dengan metode validasi silang. Beberapa nilai α yang dibangkitkan dicobakan pada validasi silang. Nilai α yang optimum adalah nilai α yang menghasilkan model dengan akurasi tertinggi pada metode validasi silang.

2.3 Tahapan Kegiatan

Secara garis besar, terdapat tiga tahapan kegiatan yang dilakukan, yaitu penarikan

data, prapemrosesan data, dan analisis data. Secara lebih rinci, berikut penjelasan dari setiap tahapan kegiatan tersebut.

Penarikan Data

Tahapan penarikan data dalam penelitian ini dibagi menjadi dua tahap, yaitu penentuan kata kunci dan penarikan data Twitter.

1. Penentuan kata kunci

   Kata kunci merupakan suatu *query* yang digunakan dalam proses pengambilan data dari Twitter. Pada penelitian ini digunakan empat topik terkait *tweets* ujaran kebencian, yaitu agama, orientasi seksual, politik, serta ras dan etnis. Secara lebih rinci, daftar kata kunci tertera pada Tabel 3.

Tabel 3: List kata kunci yang digunakan.

Topik	Kata Kunci
Agama	HTI, Islam nusantara, Islam radikal, PKI, Yahudi
Orientasi Seksual	Banci, Bencong, LGBT agama, LGBT Indonesia, LGBT usir, Orientasi homo, Orientasi LGBT, Orientasi seksual
Politik	@aniesbaswedan, @Dennysiregar7, @PSI_ID, Ade Armando, Ahok, Anies, Cebong, Dewi Tanjung, Jokowi, Kadrun, Prabowo
Ras dan Etnis	Bacin , Cina, Papua, Pribumi, Wamena

2. Penarikan data Twitter

   Proses penarikan data dalam penelitian ini dilakukan menggunakan *library* Twython yang tersedia pada bahasa pemrograman Python 3.7. Penggunaan *library* ini membutuhkan akses OAuth 2 Twitter *Developer* untuk dapat menggunakan API jenis *search*. API tersebut digunakan untuk mendapatkan *tweets* yang memiliki kata kunci yang telah ditentukan terlebih dahulu.

Prapemrosesan Data

Dalam penelitian ini ada empat tahap prapemrosesan data, yaitu pembersihan data, pembuatan *document-term matrix,* penarikan contoh acak, dan pelabelan data.

1. Pembersihan data

   Pembersihan data merupakan tahapan penting yang harus dilakukan sebelum melakukan pemodelan klasifikasi. Pembersihan Data bertujuan untuk membersihkan data *tweets* dari hal-hal yang tidak diperlukan dalam penelitian ini serta mengubah *tweets* sesuai standar tata Bahasa Indonesia. Tahapan yang dilakukan adalah membuang *retweets*, menghapus link dan angka, mengubah huruf kapital menjadi *lowercase*, dan menghapus *stopwords*.

2. Pembuatan *document-term matrix*

   Setelah dilakukan proses pembersihan data, data tersebut dibentuk menjadi *document-term matrix*. Baris pada matriks tersebut menunjukkan dokumen *tweets* (selanjutnya disebut dokumen) dan kolom menunjukan frekuensi kemunculan kata yang terdapat di dokumen.

3. Penarikan contoh acak

   Setelah dilakukan proses pembuatan *document-term matrix*, dari data tersebut diambil contoh acak sederhana sebanyak 1000 data dokumen untuk dilakukan pelabelan.

4. Pelabelan data

   Hasil penarikan contoh acak pada tahap 3 diikuti dengan pelabelan secara manual. Label terdiri dari dua kelas, yaitu *Hate* dan *Non Hate*. Label *Hate* digunakan pada dokumen yang mengandung ujaran kebencian, sedangkan label *Non Hate* digunakan untuk melabeli dokumen yang tidak mengandung ujaran kebencian. Pelabelan dilakukan oleh mahasiswa Fakultas Psikologi Universitas Padjadjaran yang pernah mendapatkan materi mengenai ujaran kebencian saat kuliah. Pedoman yang digunakan untuk melabeli kelas dokumen adalah peraturan kebijakan perilaku kebencian yang dikeluarkan oleh Twitter.

Analisis Data

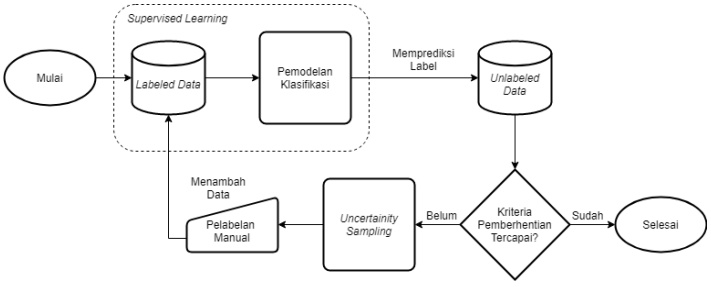
Tahapan analisis data dalam penelitian ini dibagi menjadi 2 tahap, yaitu pemodelan klasifikasi dan penambahan data.

1. Pemodelan klasifikasi

   Pemodelan klasifikasi dilakukan terhadap data yang telah melalui prapemrosesan. Peubah penjelas yang digunakan adalah setiap kata yang terdapat pada *document-term matrix* dan peubah respon adalah respon yang berisi label kelas *Hate* dan *Non Hate*. Pemodelan klasifikasi menggunakan 5-lipat validasi silang pada data latih dan lima metode klasifikasi, yaitu regresi logistik, regresi logistik dengan regularisasi LASSO, regresi logistik dengan regularisasi gulud, *multinomial naïve bayes*, dan *multinomial naïve bayes* dengan pemulusan aditif. Kemudian dilakukan evaluasi menggunakan nilai akurasi, spesifisitas, dan sensitivitas pada data uji berdasarkan model yang dibentuk.

2. Penambahan data

   Salah satu permasalahan yang terdapat pada pemodelan klasifikasi adalah *overfitting*. *Overfitting* terjadi ketika model yang dibangun memiliki kinerja yang tinggi pada data latih namun kinerja yang rendah pada data uji. Hal tersebut menandakan model yang dibangun kurang dapat menggeneralisasi data yang ada. Salah satu cara untuk mengatasi permasalahan tersebut adalah dengan menambah data latih yang baru (Ying, 2019). Penelitian ini menggunakan dua metode penambahan data, yaitu penambahan data dengan metode penarikan contoh acak sederhana dan metode *active learning*.



Gambar 1: Diagram alir *active learning* dengan *pool-based sampling.*

Dalam penelitian ini, skenario *active learning* yang digunakan adalah *pool-based sampling*. Ilustrasi *pool-based sampling* dapat dilihat pada Gambar 1. Sedangkan *query strategy frameworks* yang digunakan adalah *uncertainty*

sampling, yaitu memilih dokumen yang diragukan untuk diberi label. Metode yang digunakan dalam penentuan dokumen yang diragukan adalah metode *least confidence*. Bila peluang dari hasil dugaan klasifikasi dokumen tersebut mendekati 0.5, maka amatan tersebut dapat dikatakan sebagai dokumen yang diragukan. Skenario *active learning* akan berjalan secara iteratif hingga kriteria pemberhentian tercapai. Kriteria pemberhentian yang ditentukan pada penelitian kali ini adalah jumlah dokumen yang dilabeli mencapai 150.

Penelitian ini juga ingin mengetahui apakah perbedaan *learning curve* dapat mempengaruhi kinerja penambahan data dengan metode penarikan contoh acak sederhana dan *active learning*. *Learning curve* adalah kurva garis yang menunjukkan hubungan antara banyaknya data pada data latih dengan kinerja model klasifikasi. Secara garis besar, *learning curve* dibagi menjadi dua kondisi, yaitu saat kinerja model klasifikasi berkembang dan saat kinerja model klasifikasi sudah tidak berkembang. Pada penelitian ini dilakukan penambahan data pada kedua kondisi tersebut. Evaluasi hasil pada setiap penambahan data dilakukan menggunakan akurasi, spesifisitas, dan sensitivitas data uji.

3. Hasil dan Pembahasan

3.1 Deskripsi Data dan Metode Penarikan Contoh Acak

Data Twitter yang telah diambil dan disortir dari tanggal 24 Oktober hingga 13 November 2019 menghasilkan sebanyak 20287 *tweets*. Jumlah *tweets* yang terambil pada masing-masing topik serta *tweets* yang mengandung kata-kata kasar dapat dilihat pada Tabel 4. Terlihat *tweets* yang paling banyak berasal dari topik politik (14894 *tweets*) dan paling sedikit berasal dari topik agama (1477 *tweets*).

Tabel 4: Sebaran *tweets* berdasarkan topik.

Topik	Frekuensi	Persentase (%)
Agama	1477	7.28
Orientasi Seksual	1697	8.36
Politik	14894	73.42
Ras	2219	10.94

Penelitian ini menggunakan metode penarikan contoh acak sederhana sebanyak 1000 tweets. Setelah membangun gugus data, berikutnya dilakukan proses pelabelan. Hasil pelabelan dokumen dapat dilihat pada Tabel 5.

Tabel 5: Sebaran dokumen berdasarkan label kelas.

Topik	Label		Frekuensi
	Hate	*Non Hate*	
Agama	39	36	75
Orientasi	42	29	71
Politik	408	340	748
Ras	62	44	106
Total	551	449	1000

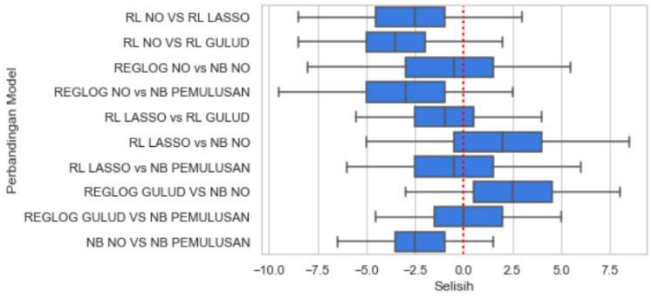
3.2 Evaluasi Pemodelan Klasifikasi

Sebelum membangun model, data dibagi menjadi data latih sebesar 80% dan data uji sebesar 20%. Pemodelan klasifikasi menggunakan 5-lipat validasi silang untuk mengetahui nilai λ terbaik pada regresi logistik dengan regularisasi LASSO dan gulud serta α terbaik pada *multinomial naïve bayes* dengan pemulusan aditif. Setelah mendapatkan nilai λ dan α terbaik, model akan dievaluasi menggunakan akurasi pada data uji. Seluruh proses tersebut diulang sebanyak 1000 kali dengan *seed* yang berbeda, sehingga diperoleh 1000 akurasi pada data latih dan pada data uji. Hal tersebut dilakukan untuk membandingkan masing-masing model dari sebaran yang dibentuk oleh 1000 akurasi tersebut.

Tabel 6: Evaluasi akurasi, sensitivitas, dan spesifisitas pada pemodelan klasifikasi.

Model	Data Latih		Data Uji	
	Rata-rata Evaluasi (%)	Simpangan Baku	Rata-rata Evaluasi (%)	Simpangan Baku
		Akurasi		
RL NO	70.15	1.42	70.27	3.02
RL LASSO	72.14	1.24	72.86	2.85
RL GULUD	73.06	1.16	73.65	2.9
NB NO	69.11	1.48	71.02	3.08
NB PEMULUSAN	73.39	1.25	73.41	2.84
		Sensitivitas		
RL NO	72.47	1.65	72.57	4.36
RL LASSO	69.54	1.77	70.78	4.53
RL GULUD	76.80	1.69	76.85	4.33
NB NO	74.44	1.93	76.29	4.13
NB PEMULUSAN	77.68	1.55	78.99	3.93
		Spesifisitas		
RL NO	67.30	2.26	67.52	4.89
RL LASSO	75.29	2.23	75.52	4.85
RL GULUD	68.44	2.41	69.92	5.04
NB NO	62.56	2.48	64.69	5.24
NB PEMULUSAN	68.12	2.06	66.69	4.86

Evaluasi akurasi model klasifikasi yang dibangun dari gugus data penarikan contoh acak sederhana dapat dilihat pada Tabel 6. Terlihat rata-rata akurasi data uji model klasifikasi berkisar 70.27%-73.65%. Model klasifikasi regresi logistik dan model regularisasi LASSO serta gulud menghasilkan rata-rata akurasi yang lebih tinggi daripada yang tidak menggunakan regularisasi. Model klasifikasi *multinomial naïve bayes*, model dengan pemulusan aditif menghasilkan rata-rata akurasi yang lebih tinggi daripada model yang tidak menggunakan pemulusan aditif.



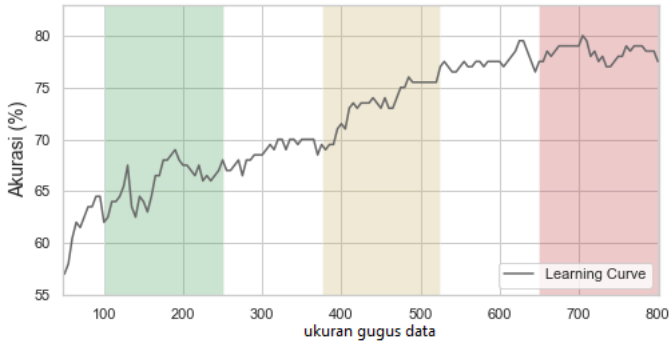
Gambar 2: Selang kepercayaan selisih akurasi data uji.

Untuk mengetahui apakah terdapat perbedaan akurasi data uji pada kelima model tersebut, disusun selang kepercayaan dari selisih akurasi data uji antar model. Langkah pertama yang dilakukan untuk mendapatkan selang kepercayaan tersebut adalah menghitung selisih akurasi data uji dari kedua model pada *seed* yang sama, proses ini menghasilkan 1000 nilai selisih akurasi data uji. Berikutnya mengurutkan 1000 nilai selisih dari yang terkecil sampai dengan yang terbesar. Untuk mendapatkan selang kepercayaan dengan taraf nyata 95%, 2.5% amatan pada bagian ekor kiri dan kanan dihilangkan.

Gambar 2 menunjukkan selang kepercayaan 95% untuk selisih akurasi data uji antar-model. Jika selang kepercayaan mencakup nilai nol maka dapat disimpulkan bahwa belum cukup bukti adanya perbedaan rata-rata akurasi data uji antar-model. Berdasarkan hasil diatas, terlihat semua selang kepercayaan mengandung angka nol sehingga dapat disimpulkan bahwa belum cukup bukti adanya perbedaan rata-rata akurasi di antara kelima model tersebut.

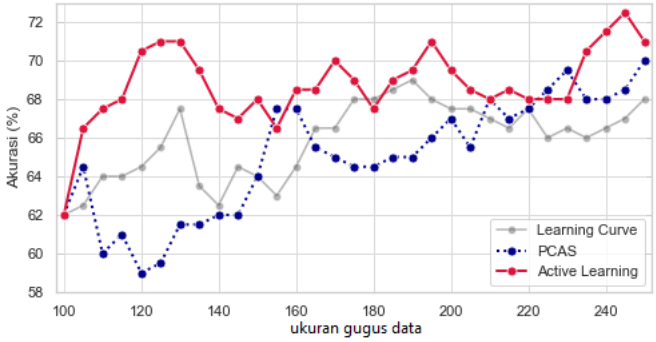
3.3 Penambahan Data

Tahapan penambahan data dilakukan dengan menggunakan dua metode, yaitu metode penarikan contoh acak sederhana (selanjutnya disebut PCAS) dan *active learning*. Model klasifikasi yang digunakan adalah regresi logistik dengan regularisasi gulud. Model klasifikasi tersebut digunakan karena menghasilkan kinerja terbaik dibandingkan model lainnya. Gambar 3 menunjukan plot *learning curve* dari akurasi data uji model regresi logistik gulud pada gugus data penarikan contoh acak sederhana. Terdapat tiga percobaan yang dilakukan, yaitu penambahan data menggunakan ukuran gugus data sebesar 100, 375, dan 650. Ukuran gugus data tersebut digunakan karena kondisi *learning curve* model klasifikasi yang berbeda-beda. Saat ukuran gugus data sebesar 100, *learning curve* model klasifikasi masih berkembang. Saat ukuran gugus data sebesar 375, *learning curve* model klasifikasi masih berkembang namun tidak sefluktuatif saat ukuran gugus data sebesar 100. Sedangkan saat ukuran gugus data sebesar 650, terlihat *learning curve* model klasifikasi sudah tidak berkembang. Percobaan dilakukan dengan menambahkan 150 data baru untuk kedua metode penambahan data. Proses tersebut dilakukan secara iteratif dengan menambahkan data baru sebanyak lima data setiap iterasinya. Total iterasi yang digunakan dalam penelitian ini adalah 30.

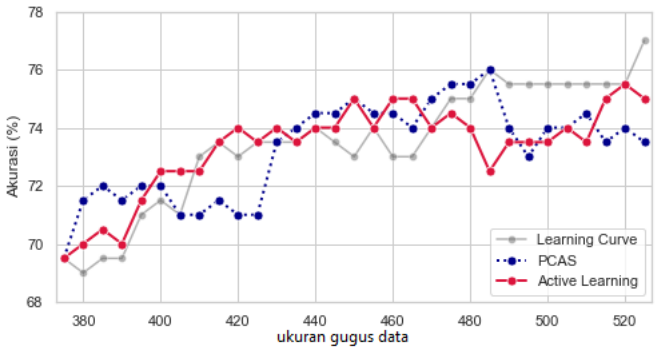


Gambar 3: Plot *learning curve* dari model regresi logistik gulud.

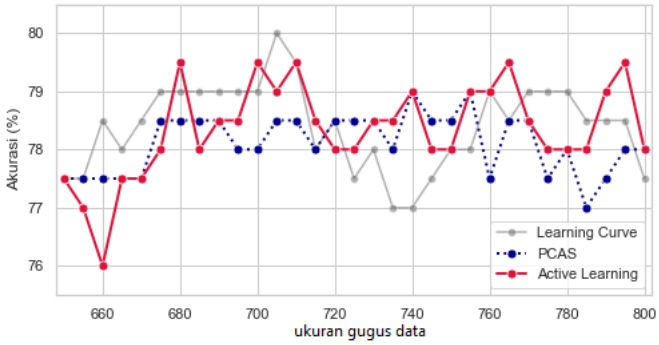
Hasil penambahan data pada ukuran gugus data sebesar 100 dapat dilihat pada Gambar 4. Terlihat penambahan data dengan metode *active learning* menghasilkan akurasi data uji yang lebih tinggi daripada metode PCAS. Selain itu, akurasi data uji sebesar 70% dapat dicapai oleh metode *active learning* hanya dengan menambahkan 20 data baru. Sedangkan, metode PCAS baru bisa mencapai akurasi data uji sebesar 70% saat menambahkan 150 data baru. Hasil penambahan data pada ukuran gugus data sebesar 375 dan 650 dapat dilihat pada Gambar 5 dan Gambar 6. Dari plot tersebut, tidak terlihat perbedaan yang signifikan antara penambahan data dengan metode PCAS dengan *active learning.*



Gambar 4: Plot penambahan data pada ukuran gugus data sebesar 100



Gambar 5: Plot penambahan data pada ukuran gugus data sebesar 375



Gambar 6: Plot penambahan data pada ukuran gugus data sebesar 650

Evaluasi rata-rata sensitivitas dan spesifisitas dapat dilihat pada Tabel 7. Terlihat rata-rata sensitivitas metode penambahan data dengan *active learning* lebih tinggi daripada PCAS. Sementara itu, rata-rata spesifisitas metode penambahan data dengan PCAS lebih tinggi daripada *active learning*. Uji t digunakan untuk menguji apakah terdapat perbeda nilai tengah akurasi antara penambahan data uji menggunakan PCAS dan metode *active learning*.

Tabel 7: Evaluasi akurasi, sensitivitas, dan spesifisitas setiap ukuran gugus data.

Ukuran Gugus Data	Metode	Rata-rata Akurasi (%)	Rata-rata Sensitivitas (%)	Rata-rata Spesifisitas (%)
100	PCAS	65.10	54.25	74.92
	Active Learning	69.00	66.77	71.02
375	PCAS	73.38	65.30	80.70
	Active Learning	73.45	68.98	77.49
650	PCAS	78.13	79.37	77.02
	Active Learning	78.38	81.02	76.00

Tabel 8: Hasil uji t untuk akurasi data uji pada setiap ukuran gugus data

Ukuran Gugus Data	Metode	Rata-rata Akurasi (%)	Simpangan Baku	*T-value*	*P-value*	Kesimpulan
100	PCAS	65.10	3.09	-6.17	0	*Active Learning* lebih baik
	Active Learning	69.00	1.56			
375	PCAS	73.38	1.54	-0.17	0.863	Tidak konklusif
	Active Learning	73.45	1.43			
650	PCAS	78.13	1.54	-0.17	0.863	Tidak konklusif
	Active Learning	78.38	1.43			

Hasil pengujian dengan uji t dapat dilihat pada Tabel 8. Penambahan data pada ukuran gugus data sebesar 100 memberikan keputusan penolakan hipotesis H_0 artinya terdapat perbedaan nilai tengah akurasi data uji antara metode PCAS dan *active learning*. Dalam hal ini metode *active learning* lebih baik daripada PCAS. Namun

demikian, penambahan data untuk ukuran gugus data sebesar 375 dan 650 memberikan keputusan menerima H_0 sehingga hasil uji tidak konklusif.

4. Simpulan

Berdasarkan hasil penelitian yang telah dilakukan, belum dapat dibuktikan adanya perbedaan performa dalam mendeteksi ujaran kebencian di Twitter dari model klasifikasi regresi logistik, regresi logistik dengan regularisasi LASSO, regresi logistik dengan regularisasi gulud, *multinomial naïve bayes*, dan *multinomial naïve bayes*. Kinerja model klasifikasi regresi logistik yang diperbaiki menggunakan metode penambahan data dengan *active learning* lebih baik dibandingkan dengan PCAS pada kondisi *learning curve* model yang masih berkembang dan ukuran gugus data sedikit.

Daftar Pustaka

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Hu, R. (2011). Active learning for text classification.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R. Spinger.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). Springer.

Medistiara, Y. (2017). Selama 2017 polri tangani 3.325 kasus ujaran kebencian. Retrieved from https://news.detik.com/berita/d-3790973/selama-2017-polri-tangani-3325-kasus-ujaran-kebencian

Schütze, H., Manning, C. D., & Raghavan, P. (2008). Introduction to information retrieval (Vol. 39). Cambridge University Press Cambridge.

Sudut Hukum. (2016). Tinjauan tentang ujaran kebencian (hate speech). Retrieved from https://suduthukum.com/2016/11/tinjauan-tentang-ujaran-kebencian-hate.html

Ying, X. (2019). An overview of overfitting and its solutions. Journal of Physics: Conference Series, 1168(2), 022022. IOP Publishing.