REPOSITOR, Vol. 2, No. 7, Juli 2020, Pp. 933-944
ISSN : 2714-7975
E-ISSN : 2716-1382

Klasifikasi Tweets Tindak Kejahatan Berbahasa Indonesia Menggunakan Naive Bayes

Setio Basuki[*1], Siti Maghfiroh[2], Yufis Azhar[3]
[1,2,3]Teknik Informatika/Universitas Muhammadiyah Malang
setiobasuki@umm.ac.id[*1], siti.maghfiroh@webmail.umm.ac.id[2], yufis@umm.ac.id[3]

Abstrak

Kasus tindak kejahatan konvensional seperti penganiayaan, penculikan, pencurian, dll masih jarang digunakan sebagai objek penelitian. Kasus kejahatan yang biasa diteliti hanya pada lingkup kejahatan cyber seperti pembajakan software, carding, penipuan online, dll. Maka dalam penelitian ini penulis mengangkat kasus kejahatan konvensional sebagai objek penelitian. Penulis mencoba mendapatkan informasi kejahatan dari media sosial, Twitter. Dari Twitter didapatkan data berupa cuitan para pengguna yang mengandung unsur kejahatan. Selanjutnya, akan dilakukan klasifikasi untuk menentukan mana di antara data tersebut yang benar-benar mengandung informasi kejahatan, dan bukan merupakan sebuah opini. Metode yang digunakan dalam pengklasifikasian data adalah algoritma Naive Bayes Classifier dengan 2 jenis dataset. Dataset pertama berisi fitur lexical atau bag of words dan dataset kedua berisi fitur sintaktik. Penulis menggunakan 2 dataset untuk membandingkan kinerja dari kedua fitur dalam proses klasifikasi data tweets. Rata-rata hasil akurasi model klasifikasi menggunakan fitur sintaktik adalah sebesar 88,1398% sedangkan pada fitur lexical atau bag of words sebesar 79,25%. Kemudian dari hasil klasifikasi, penulis mendapatkan lokasi di mana tindak kejahatan tersebut terjadi menggunakan metode Named Entity Recognition (NER). Dari proses NER tersebut, maka didapatkan hasil akurasi sebesar 65%.

Kata Kunci: *Klasifikasi, Naive Bayes, Kejahatan, Sintaktik, Bag of Words*

.

Abstract

Conventional crime cases such as torture, kidnapping, theft, etc. are still rarely used as research objects. Cases of crime commonly researched only on the scope of cyber crime such as software piracy, carding, online fraud, etc. So in this study the authors raised cases of conventional crime as the object of research. The author tries to get crime information from social media, Twitter. From Twitter obtained data in the form of cuitan users who contain elements of crime. Next, a classification will be made to determine which of these data actually contains criminal information, and is not an opinion. The method used in data classification is the Naive Bayes Classifier algorithm with 2 types of datasets. The first dataset contains lexical or bag of words features and the second dataset contains syntactic features. The author uses 2 datasets to compare the performance of both features in the process of classifying data tweets. The average result of classification model accuracy using syntactic feature is 88,1398% while in lexical or bag of words feature is 79,25%. Then from the classification results, the authors get the location where the crime occurred using Named Entity Recognition method (NER). From the process of NER, then obtained the result of accuracy of 65%.

Keywords: *Classification, Naive Bayes, Crime, Syntactic, Bag of Words*

1. Pendahuluan

Twitter merupakan suatu kumpulan kata yang berisikan tidak lebih dari 140 karakter. Pada media sosial ini masyarakat lebih mudah untuk memberikan kritik dan saran kepada media elektronik secara *realtime*. Kalimat yang dimuat dalam *Twitter* adalah bahasa alami manusia yang merupakan bahasa dengan struktur yang kompleks [1] .Penggunaan *Twitter* tidak lepas dari pendapat dan ekspresi dari para pengguna terhadap berbagai hal, seperti seni, isu sosial, isu politik, isu kejahatan dan sebagainya. Pada situs CNN Indonesia, Roy Simangunsong menyampaikan bahwa, Indonesia adalah salah satu basis pengguna *Twitter* terbesar di dunia. Berdasarkan data yang diperolehnya, paling tidak ada 4,1 juta *tweets* yang berasal dari Indonesia. 77 persen pengguna *Twitter* di Indonesia aktif setiap harinya. Dari 77 persen tersebut,

54 persen diantaranya melakukan 2 *tweets* setiap harinya [2].Dengan demikian kita dapat menggunakan *Twitter* untuk kepentingan tertentu seperti pengambilan data kejahatan yang ada di Indonesia.

Menurut Soerjono Kriminalitas atau sering disebut juga dengan kejahatan adalah suatu tindakan anti-sosial yang menimbulkan kerugian, ketidakpatutan dalam masyarakat sehingga dalam masyarakat terdapat kegelisahan dan untuk menentramkan masyarakat negara harus menjatuhkan pidana kepada barang siapa yang melakukan tindakan tersebut. Seperti yang dilansir oleh Badan Pusat Statistik, angka kejahatan (tindak pidana) pada tahun 2015 naik dari tahun sebelumnya yaitu 352 936 kasus kejahatan [3]. Tingginya angka kejahatan tersebut dapat berdampak pada isu pertahanan dan keamanan yang ada di Indonesia. Maka dari itu peneliti mencoba untuk mengklasifikasi data tindak kejahatan yang dapat memberikan informasi apa dan di mana kejahatan terjadi dengan mengambil data dari *Twitter*.

Klasifikasi dokumen adalah bidang penelitian dalam perolehan informasi untuk menentukan atau mengkategorikan suatu dokumen ke dalam satu atau lebih kelompok berdasarkan isi dokumen. Klasifikasi dokumen bertujuan untuk mengelompokkan dokumen yang tidak terstruktur ke dalam kelompok atau kategori yang menggambarkan isi dari dokumen [4]. Sebelumnya telah dilakukan penelitian dengan memanfaatkan *Twitter* untuk melakukan pengklasifikasian data *tweet*s yang mengandung informasi tentang lalu lintas di kota Bandung menggunakan teknik *Naïve Bayes Classifiers* [5]. Serta terdapat penelitian sebelumnya yang menggunakan teknik *Naïve Bayes Classifiers* untuk menganalisis pola penyalahgunaan *facebook* sebagai alat kejahatan *trafficking* [6]. Algoritma *Naïve Bayes* Ini adalah teknik sederhana, dan harus digunakan sebelum mencoba metode yang lebih kompleks [7]. Hasil dari klasifikasi akan digunakan untuk mendapatkan lokasi kejadian tindak kejahatan. Dan dalam memperoleh lokasi kejahatan, digunakan metode NER (*Named Entity Recognition)* untuk mengekstrak informasi yang terstruktur dari teks yang tidak terstruktur. Terdapat beberapa perbedaan penelitian ini dengan penelitian sebelumnya antara lain, perbedaan terletak pada bentuk data, kelas target yang digunakan, serta fitur yang digunakan. Jika pada penelitian sebelumnya menggunakan data kemacetan, pada penelitian ini menggunakan data tindak kejahatan. Dengan perbedaan data yang diambil, maka untuk kelas target dan fitur yang ditentukan juga berbeda.

2. Metode Penelitian

2.1 Studi Pustaka

Mengumpulkan semua referensi dan memahami konsep pembangunan sistem tentang studi kasus yang diambil, yaitu mengenai algoritma *Naïve Bayes Classifier,* serta metode memperoleh lokasi dengan *Named Entity Recognition.* Studi pustaka diperoleh dari berbagai referensi seperti jurnal, *website* resmi, buku, laporan Tugas Akhir (TA), dsb.

2.2 Persiapan Data

Pada tahap ini dilakukan pengunduhan *data tweets* menggunakan *Twitter4j.* Saat pengunduhan *data tweets* diperlukan informasi tentang *Consumer Key, Consumer Secret, Token Key & Token Secret* yang telah disediakan oleh *Twitter* dari masing-masing akun pengguna. Jumlah *data tweets* yang digunakan pada penelitian ini adalah 1000 *tweets.*

2.3 Preprocessing Data

Preprocessing Data dilakukan untuk membersihkan data hasil *crawling* dan pengambilan informasi yang dibutuhkan untuk penelitian. Pada tahap ini, yang dilakukan antara lain *remove hastag, remove ReTweet (RT), remove punctuation, remove link, remove username* serta menghapus kata yang tidak sesuai dengan informasi yang digunakan.

2.4 Ekstraksi Fitur

Ekstraksi fitur dilakukan untuk mendapatkan karakteristik pembeda antar kelas untuk mempermudah proses klasifikasi. Fitur yang digunakan dalam penelitian ini adalah fitur sintaktik dan fitur *bag of words.*

2.4.1 Fitur Sintaktik

Fitur sintaktik merupakan perbedaan dari masing-masing kelas yang diekstraksi berdasarkan sintak dari *dataset*[8]. Dalam penyusunan pola terhadap kalimat, terdapat beberapa kata kunci kejahatan dan potensi negasi yang menjadi karakteristik masing-masing kelas. Kata

kunci berasal dari *kbbi.kemendikbud.go.id* untuk pengambilan kata dasar. Dan berasal dari *sinonimkata.com* untuk pengambilan kata yang mempunyai persamaan (sinonim) dari kata kunci kejahatan tersebut, dan untuk kata kunci pendukung kelas nonKejahatan kami dapatkan dari *dataset.* Tabel 1 menunjukkan daftar dari kata kunci kami paparkan pada perancangan sistem.

Tabel 1. Daftar Fitur Sintaktik

No	Fitur	Deskripsi
1	*Pembunuhan*	<*true,false*> bunuh, dibunuh, pembunuh, pembunuhan, berencana, genosida, pembantaian, rajapati
2	*Pencurian*	<*true,false*> Kata kunci kejahatan pencurian
3	*Unjuk Rasa Anarkis*	<*true,false*> curi, pencuri, pencurian, penggarongan, penyamunan, rampas, perampasan, rampok, perampokan
4	*Penculikan*	<*true,false*> culik, penculikan, penculik, diculik, melarikan
5	*Asusila*	<*true,false*> asusila, amoral, cabul, dursila, lacur, mesum, porno, tunasusila, pemerkosa, perkosa
6	*Pembakaran*	<*true*,*false*> bakar, pembakaran, dibakar, pembakar
7	*Narkotika dan Psikotropika*	<*true*,*false*> narkotika, psikotropika, sabu, ganja, pengedar, narkoba, heroin, opium
8	*Penganiayaan*	<*true*,*false*> aniaya, dianiaya, penganiaya, penganiayaan, kebengisan, kekejaman, tindas, penindasan, siksa, penyiksaan
9	*Tabrak Lari*	<*true,false*> tabrak
10	*Penggelapan*	<*true,false*> gelap, penggelapan, penipuan, penyelewengan, korupsi, manipulasi, kecurangan, digelapkan
11	*Kata Negasi*	<*true,false*> bukan, tidak, tak, gak, nggak
12	*Kata Potensi Negasi*	<*true*,*false*> sosialisasi, dugaan, diduga, stop, imbau Rekonstruksi, isu, *hoax*, semoga, khawatir, pengawasan, memerangi, gerakan, anti, waspada, hindari, usulan, apresiasi, singkatan, pencegahan, tabayyun, mengkaji, antisipasi, ancaman, cegah, galakkan, tips, hindari, bohong, cara, terhindar, mencegah, himbau, bantah
13	*Kata tanya potensi negasi*	<*true*,*false*> apa, siapa, di mana, kenapa, kapan, bagaimana, benarkah, mana, apakah
14	*Kata definisi potensi negasi*	<*true*,*false*> adalah, merupakan, dimaksud, bernama, disebut, dinamakan, bahwa
15	*Class*	<nominal> asusila, narkotikapsikotropika, pembakaran, pembunuhan, penculikan, pencurian, penganiayaan, penggelapan, tabrakLari, UnjukRasaAnarkis

2.4.2 Fitur Leksikal *(Bag of Words)*

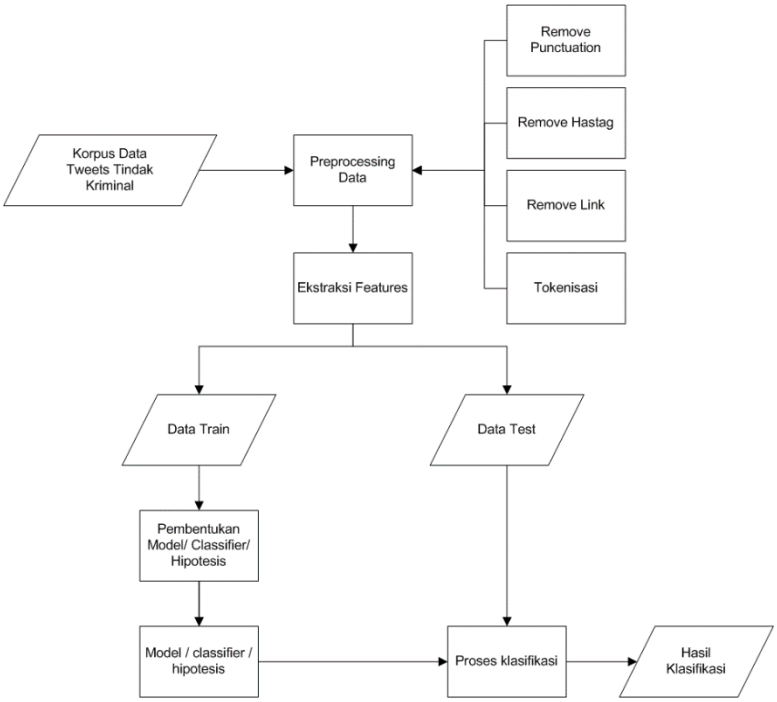
Umumnya fitur Leksikal diekstraksi berdasarkan konten kata dari *dataset.* Sedangkan *Bag of Words* merupakan bagian dari fitur Leksikal yang menggabungkan *Unigram* sebagai fiturnya [8]. Contoh dari fitur Leksikal seperti pada Tabel 2.

Tabel 2. Ekstraksi Fitur

Kalimat : “Kejahatan pencurian dan kejahatan penculikan sama tapi beda objek dituju”	
Fitur	Ekstraksi
Bag of Words	{(kejahatan,2), (pencurian,1), (dan,1), (penculikan,1), (sama,1), (tapi, 1), (beda, 1) (objek, 1), (dituju, 1)}

2.5 Klasifikasi menggunakan *Naive Bayes Classifier*

Model klasifikasi pada penelitian tugas akhir ini menggunakan model algoritma *Naive Bayes Classifiers.* Proses pelatihan diawali dengan memasukkan data hasil *preprocessing* dan hasil ekstraksi fitur, seperti pada Gambar 1.



Gambar 1. Flowchart Sistem Klasifikasi

Dalam melakukan klasifikasi, algoritma *Naive Bayes* menggunakan teorema Bayes yang menerapkan konsep probabilitas. *Naïve Bayes classifier* adalah *classifier* probabilistik sederhana berdasarkan penerapan teorema *Bayes* (dari statistik *Bayesian*) dengan asumsi independen (naif) yang kuat. Sebuah istilah yang lebih deskriptif untuk model probabilitas yang digarisbawahi adalah "model fitur independen" [9]. Prediksi *Bayes* didasarkan pada teorema *Bayes* dengan formula umum seperti pada Persamaan 1 [10].

$$\mathrm{P(Y|X)} = \frac{\mathrm{P(Y)} \prod_{i=1}^{q} \mathrm{P(X}_i \mid \mathrm{Y)}}{\mathrm{P(X)}} \quad (1)$$

2.6 Mendapatkan Lokasi

Arsitektur sistem dari *InaNLP* hampir sama dengan *Stanford NLP*. Di mana setiap modulnya, dapat diakses oleh pengguna *(user).* Modul-modul yang terdapat dalam InaNLP antara lain *Sentence Splitter, Tokenization, Word Normalization, Morphologically Analyzer, Pos Tagger, Phrase Tagger, Named Entity Tagger, Syntactic Parset, serta Semantic Analyzer.* Metode yang digunakan untuk modul bervariasi antara aturan berbasis *approaches* (*Rule based Approaches)* dan pendekatan berbasis statistik *(Statistical Based Approach)* [11].

2.7 Pengujian

Pengujian dalam penelitian ini akan dibagi menjadi 2, yaitu pengujian terhadap hasil klasifikasi tweets berdasarkan kelas target serta pengujian terhadap hasil mendapatkan lokasi. Berikut merupakan skenario pengujian.

Tabel 3. Skenario Pengujian

Tahap	Parameter Pengujian	Parameter Data
I	Data Latih	*Tweets* Kejahatan
II	Hasil Klasifikasi	*Tweets* Kejahatan
III	Hasil Lokasi	*Tweets* kejahatan & lokasi

Pada Tabel 3, tahap I dilakukan pengujian terhadap data latih menggunakan *cross validation* untuk mengetahui kinerja dari model klasifikasi yang telah dibentuk. Tahap II, dilakukan pengujian terhadap data uji dengan jumlah 200 data *tweets.* Dan pada tahap III, dilakukan pengujian terhadap data hasil klasifikasi untuk mendapatkan lokasi kejahatan yang terjadi. Untuk mendapatkan hasil pengujian, digunakan hasil *accuracy, precision, recall* dengan Persamaan 2, Persamaan 3, dan Persamaan 4.

$$\boldsymbol{Accuracy}: \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

$$\boldsymbol{Precision}: \frac{TP}{TP + FP} \quad (3)$$

$$\boldsymbol{Recall}: \frac{TP}{TP + FN} \quad (4)$$

3. Hasil Penelitian dan Pembahasan

3.1 *Crawling Data*

Tabel 3. Data Hasil Crawling

{"contributors": null, "truncated": false, "text": "@nununsativa @Jakarta_Kece @Takviri @TanpaDelusi @WinnerLawyer itu mah kta si rizik..eh dia lgi kena kasus asusila kan", "is_quote_status": false, "in_reply_to_status_id": 828213384373956608, "in_reply_to_user_id": 860947477, "id": 828225501802033153, "favorite_count": 0, "entities": {"symbols": [], "user_mentions": [{"indices": [0, 12], "screen_name": "nununsativa", "id": 860947477, "name": "#AhokerSumberHoax", "id_str": "860947477"}, {"indices": [13, 26], "screen_name": "Jakarta_Kece", "id": 3328346382, "name": "Jakarta Kece", "id_str": "3328346382"}, {"indices": [27, 35], "screen_name": "Takviri", "id": 2547595849, "name": "Takviri", "id_str": "2547595849"}, {"indices": [36, 48], "screen_name": "TanpaDelusi", "id": 2266351464, "name": "De' Lusi", "id_str": "2266351464"}, {"indices": [49, 62], "screen_name": "WinnerLawyer", "id": 61388496, "name": "WJLaw", "id_str": "61388496"}], "hashtags": [], "urls": []}, "retweeted": false, "coordinates": null, "source": "Twitter for Android", "in_reply_to_screen_name": "nununsativa", "id_str": "828225501802033153", "retweet_count": 0, "metadata": {"iso_language_code": "in", "result_type": "recent"}, "favorited": false, "user": {"follow_request_sent": null, "has_extended_profile": false, "profile_use_background_image": true, "id": 811565469589127168, "verified": false, "translator_type": "none", "profile_text_color": "333333", "profile_image_url_https": "https://abs.twimg.com/sticky/default_profile_images/default_profile_6_normal.png", "profile_sidebar_fill_color": "DDEEF6", "entities": {"description": {"urls": []}}, "followers_count": 5, "protected": false, "location": "", "default_profile_image": true, "id_str": "811565469589127168", "lang": "id", "utc_offset": null, "statuses_count": 295, "description": "", "friends_count": 132, "profile_background_image_url_https": null, "profile_link_color": "1DA1F2", "profile_image_url": "http://abs.twimg.com/sticky/default_profile_images/default_profile_6_normal.png", "following": null, "geo_enabled": false, "profile_background_color": "F5F8FA", "profile_background_image_url": null, "screen_name": "giwanto4", "is_translation_enabled": false, "profile_background_tile": false, "favourites_count": 15, "name": "giwanto", "notifications": null, "url": null, "created_at": "Wed Dec 21 13:34:20 +0000 2016", "contributors_enabled": false, "time_zone": null, "profile_sidebar_border_color": "C0DEED", "default_profile": true, "is_translator": false, "listed_count": 0}, "geo": null, "in_reply_to_user_id_str": "860947477", "lang": "in", "created_at": "Sun Feb 05 12:55:21 +0000 2017", "in_reply_to_status_id_str": "828213384373956608", "place": null}

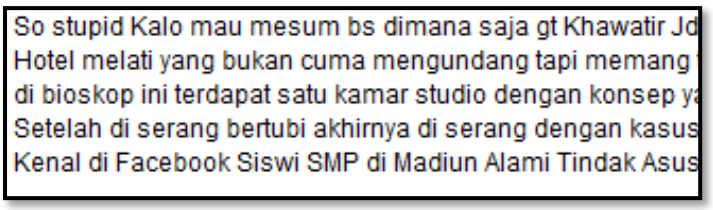
Tabel 3 merupakan contoh kalimat data hasil *crawling* dari *Twitter*. Dari hasil tersebut, pada saat *preprocessing* hanya akan diambil data yang mengandung *cuitan* dari pengguna.

3.2 *Preprocessing Data*

Langkah pertama yang dilakukan oleh pengguna adalah melakukan pembersihan data. Di mana pada proses ini pengguna akan membersihkan hasil *crawling data* sesuai dengan rancangan yang telah dibuat. Gambar 2 berikut merupakan hasil dari *preprocessing data.*

Gambar 2. Hasil Preprocessing Data

Pada hasil *preprocessing data* tahap 1, terdapat *space* kosong yang dapat mempengaruhi proses selanjutnya. Sehingga untuk tahap selanjutnya dilakukan penghapusan *space* tersebut dengan *method removeDuplicates().* Sehingga hasil yang didapatkan seperti pada Gambar 3.



Gambar 3. Hasil Preprocessing Data Setelah Space Kosong dihapus

Namun pada hasil pembersihan data terdapat beberapa tanda baca dan kata-kata tidak jelas yang tidak terhapus dan nantinya akan mempengaruhi proses klasifikasi. Gambar 4 berikut contoh dari kekurangan proses pembersihan data.

MetroTV Polda Kalteng Terbitkan SP Kasus Asusila Bupati Katingan
Kasus 'Asusila' Terpa Dunia Pendidikan di Soppeng
\n\udd\ude\nWaduhhh mana mungkin \nZina jelas pada zona asusila
mawar tuh korban korban asusila selalu mawar Mawar doyan di lecehin

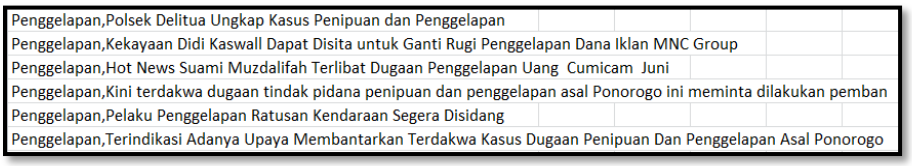
Gambar 4. Kekurangan proses pembersihan data

Gambar di atas menunjukkan kekurangan pada proses pembersihan data. Terlihat pada tanda baca petik (‘), tanda baca garing miring (\) serta kata-kata tidak jelas seperti “udd\ude\n” yang tidak hilang walaupun sudah dibersihkan. Dari hasil ini, peneliti mencoba untuk menghapus secara manual tanda baca serta kata tersebut. Sehingga setelah proses pembersihan data atau *preprocessing* selesai, masih ada satu proses yang dilakukan secara manual yaitu melakukan cek hasil *preprocessing* serta menghapus tanda baca dan kata-kata yang dapat mempengaruhi hasil klasifikasi.

3.3 Ekstraksi Fitur

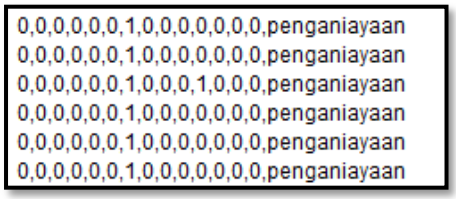
3.3.1 Fitur Sintaktik

Setelah dilakukan pembersihan data, pengguna diwajibkan untuk melakukan labelisasi data *tweets* ke dalam beberapa kelas target kejahatan yang telah ditentukan. Labelisasi data *tweets* dilakukan secara manual oleh pengguna. Setelah dilakukan labelisasi, data *tweets* tetap disimpan dalam format *.csv. Seperti yang ditunjukkan pada Gambar 5.


Penggelapan,Polsek Delitua Ungkap Kasus Penipuan dan Penggelapan
Penggelapan,Kekayaan Didi Kaswall Dapat Disita untuk Ganti Rugi Penggelapan Dana Iklan MNC Group
Penggelapan,Hot News Suami Muzdalifah Terlibat Dugaan Penggelapan Uang Cumicam Juni
Penggelapan,Kini terdakwa dugaan tindak pidana penipuan dan penggelapan asal Ponorogo ini meminta dilakukan pemban
Penggelapan,Pelaku Penggelapan Ratusan Kendaraan Segera Disidang
Penggelapan,Terindikasi Adanya Upaya Membantarkan Terdakwa Kasus Dugaan Penipuan Dan Penggelapan Asal Ponorogo

Gambar 5. Data yang telah dilabeli target kelas

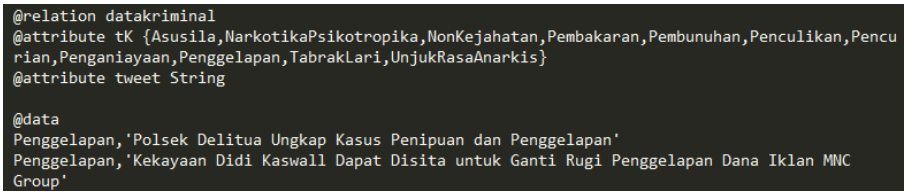
Dalam melakukan ekstraksi fitur Sintaktik, peneliti menggunakan *ArrayList()* untuk membaca fitur dari *file* *.txt. Terdapat 14 fitur yang digunakan untuk ekstraksi. Gambar 6 berikut merupakan hasil dari ekstraksi fitur Sintaktik.


0,0,0,0,0,0,1,0,0,0,0,0,0,0,penganiayaan
0,0,0,0,0,0,1,0,0,0,0,0,0,0,penganiayaan
0,0,0,0,0,0,1,0,0,0,1,0,0,0,penganiayaan
0,0,0,0,0,0,1,0,0,0,0,0,0,0,penganiayaan
0,0,0,0,0,0,1,0,0,0,0,0,0,0,penganiayaan
0,0,0,0,0,0,1,0,0,0,0,0,0,0,penganiayaan

Gambar 6. Hasil Ekstraksi Fitur Sintaktik

3.3.2 Fitur Leksikal *(Bag of Words)*

Untuk mendapatkan hasil dari ekstraksi fitur *bag of words,* format data awal yang digunakan berbeda dengan ekstraksi fitur *Sintaktik.* Dimana pada *bag of words* ini, data awal setelah dibersihkan dan dilabeli akan diubah ke dalam bentuk sesuai *.arff . Gambar 7 berikut contoh dari format bentuk dasar dari *bag of words* sebelum dilakukan ekstraksi.

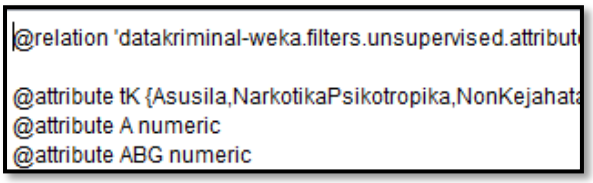


```
@relation datakriminal
@attribute tK {Asusila,NarkotikaPsikotropika,NonKejahatan,Pembakaran,Pembunuhan,Penculikan,Pencu
rian,Penganiayaan,Penggelapan,TabrakLari,UnjukRasaAnarkis}
@attribute tweet String

@data
Penggelapan,'Polsek Delitua Ungkap Kasus Penipuan dan Penggelapan'
Penggelapan,'Kekayaan Didi Kaswall Dapat Disita untuk Ganti Rugi Penggelapan Dana Iklan MNC
Group'
```

Gambar 7. Data Awal Bag of Words

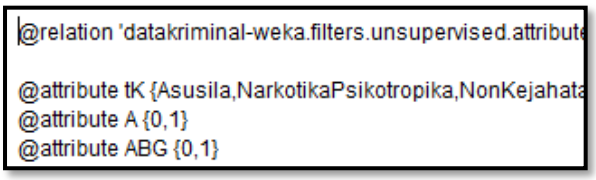
Selanjutnya proses ekstraksi fitur *Bag of Words* yaitu melakukan filter terhadap *dataset.* Tipe data awal *String* akan dibentuk menjadi bentuk *Numeric.* Proses ini menggunakan *filter StringToWordVector.* Gambar 8 berikut hasil *StringToWordVector.*


@relation 'datakriminal-weka.filters.unsupervised.attribut

@attribute tK {Asusila,NarkotikaPsikotropika,NonKejahata
@attribute A numeric
@attribute ABG numeric

Gambar 8. Hasil StringToWordVector

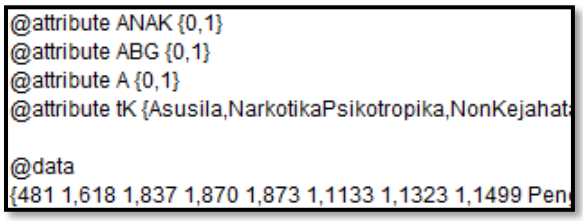
Karena algoritma *Naive Bayes Classifier* tidak mendukung tipe data *numeric,* maka dilakukan *filter NumericToNominal.* Gambar 9 berikut hasil filter.


@relation 'datakriminal-weka.filters.unsupervised.attribut

@attribute tK {Asusila,NarkotikaPsikotropika,NonKejahata
@attribute A {0,1}
@attribute ABG {0,1}

Gambar 9. Hasil NumericToNominal

Hasil dari filter *StringToWordVector* dan *NumericToNominal* membuat susunan atribut target kelas berada di atas, sehingga pada saat proses klasifikasi atribut paling bawahlah yang menjadi patokan. Harus dilakukan *re-order* posisi atribut dengan menggunakan *filter re-order.* Gambar 10 berikut hasil *filter re-order.*

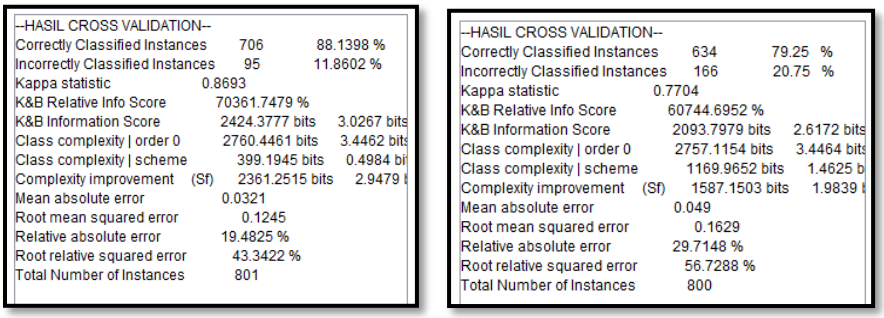


Gambar 10. Hasil Filter Re-Order

3.4 Klasifikasi

3.4.1 Pembentukan Model dari Fitur Sintaktik dan Fitur *Bag Of Words*

Model dibentuk dari *data train* yang telah diekstraksi. Model dibentuk berdasarkan algoritma *Naive Bayes* dan selanjutnya akan dilakukan pengujian terhadap kinerja model dengan menggunakan *K-Cross Validation* 10 kali iterasi. Hasil pengujiannya seperti pada Gambar 11.

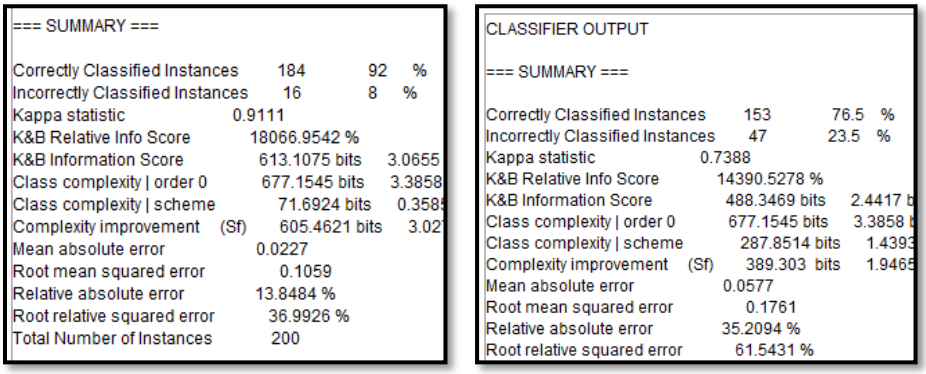


Gambar 11. Hasil Pengujian Model Fitur Sintaktik (Kiri) dan Fitur Bag Of Words (Kanan)

Dari hasil pengujian yang ditunjukkan oleh Gambar 12 dan Gambar 13, didapatkan hasil yang cukup optimal dari *data train* sejumlah 800 data mendapatkan *accuracy* dari fitur Sintaktik adalah 88,1398% dan fitur *bag of words* adalah 79,25%.

3.4.2 Pengujian Klasifikasi

Tes klasifikasi dilakukan dengan memasukkan 200 data uji *(data test)* yang selanjutnya akan diuji dengan model yang sudah dibentuk. Setelah itu dilakukan tes klasifikasi dan mempunyai hasil seperti pada Gambar 12.



Gambar 22. Hasil Pengujian Data Test Fitur Sintaktik (Kiri) dan Fitur Bag Of Words (Kanan)

Untuk mendapatkan *accuracy, precision, recall* harus diketahui *confusion matrix* dari hasil klasifikasi. Tabel 4 dan Tabel 5 berikut *Confusion Matrix* dari hasil klasifikasi.

Tabel 4. Confusion Matrix Tes Klasifikasi Fitur Sintaktik

```
=== CONFUSION MATRIX ===
  a  b  c  d  e  f  g  h  i  j  k   <-- classified as
 25  0  0  0  0  1  0  0  0  0  0 |  a = asusila
  0 15  0  0  0  0  0  0  0  0  0 |  b = narkotikapsikotropika
  0  0  8  0  0  0  0  0  0  0  1 |  c = pembakaran
  0  0  0 21  0  0  0  0  0  0  0 |  d = pembunuhan
  0  1  0  0 22  0  1  0  0  0  0 |  e = penculikan
  0  0  0  0  0 17  0  0  0  0  1 |  f = pencurian
  0  0  0  1  0  1 19  0  0  0  0 |  g = penganiayaan
  0  0  0  0  0  0  0 16  0  0  0 |  h = penggelapan
  0  0  0  0  0  1  0  0 22  0  0 |  i = tabraklari
  0  0  0  0  0  0  0  0  0 19  0 |  j = unjukrasaanarkis
  0  1  1  3  0  1  0  0  1  1  0 |  k = nonkejahatan
```

Tabel 5. Confusion Matrix Tes Klasifikasi Fitur Bag of Words

```
=== CONFUSION MATRIX ===
  a  b  c  d  e  f  g  h  i  j  k   <-- classified as
 18  1  0  0  0  2  0  2  2  0  1 |  a = Asusila
  1 14  0  0  0  0  0  0  0  0  0 |  b = NarkotikaPsikotropika
  0  1  3  0  1  0  1  0  1  1  0 |  c = NonKejahatan
  1  0  1  1  0  0  0  2  2  1  1 |  d = Pembakaran
  0  0  0  0 13  0  1  5  2  0  0 |  e = Pembunuhan
  1  0  0  0  0 22  0  1  0  0  0 |  f = Penculikan
  0  2  2  0  0  1 13  0  0  0  0 |  g = Pencurian
  1  1  0  0  2  0  0 16  0  0  1 |  h = Penganiayaan
  0  0  0  0  0  0  0  0 16  0  0 |  i = Penggelapan
  0  1  0  0  0  0  0  0  0 22  0 |  j = TabrakLari
  0  0  0  0  0  0  1  0  2  1 15 |  k = UnjukRasaAnarkis
```

Dari hasil *Confusion Matrix* di atas, maka dapat diketahui hasil dari *accuracy, precision, recall* dari masing-masing target kelas. Kemudian hasil dari perhitungan seluruh target kelas akan didapatkan hasil rata-rata. Tabel 6 dan Tabel 7 berikut hasil *precision, recall* dari *detail Accuracy by Class*.

Tabel 6. Detail Accuracy By Class fitur Sintaktik

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.962    0          1          0.962     0.98       1         asusila
                 1        0.011      0.882      1         0.938      0.995     narkotikapsikotropika
                 0.889    0.005      0.889      0.889     0.889      0.997     pembakaran
                 1        0.022      0.84       1         0.913      0.99      pembunuhan
                 0.917    0          1          0.917     0.957      1         penculikan
                 0.944    0.022      0.81       0.944     0.872      0.996     pencurian
                 0.905    0.006      0.95       0.905     0.927      0.999     penganiayaan
                 1        0          1          1         1          1         penggelapan
                 0.957    0.006      0.957      0.957     0.957      0.997     tabraklari
                 1        0.006      0.95       1         0.974      1         unjukrasaanarkis
                 0        0.01       0          0         0          0.74      nonkejahatan
Weight avg.      0.92     0.008      0.897      0.92      0.907      0.987
```

Tabel 7. Detail Accuracy By Class fitur Sintaktik

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.692    0.023      0.818     0.692     0.75       0.974     Asusila
                 0.933    0.032      0.7       0.933     0.8        0.997     NarkotikaPsikotropika
```

0.375	0.016	0.5	0.375	0.429	0.825	NonKejahatan
0.111	0	1	0.111	0.2	0.965	Pembakaran
0.619	0.017	0.813	0.619	0.703	0.978	Pembunuhan
0.917	0.017	0.88	0.917	0.898	0.988	Penculikan
0.722	0.016	0.813	0.722	0.765	0.984	Pencurian
0.762	0.056	0.615	0.762	0.681	0.941	Penganiayaan
1	0.049	0.64	1	0.78	0.999	Penggelapan
0.957	0.017	0.88	0.957	0.917	0.999	TabrakLari
0.789	0.017	0.833	0.789	0.811	0.951	UnjukRasaAnarkis
Weighted Avg.	*0.765*	*0.025*	*0.784*	*0.765*	*0.75*	*0.972*

3.5 Mendapatkan Lokasi

Dalam mendapatkan lokasi, peneliti menggunakan InaNLP dengan *Named Entity.* Data yang digunakan adalah data uji hasil klasifikasi yang berjumlah 200 data. Dan pada saat dilakukan pengujian, hasil data yang terprediksi secara benar adalah 130 data. Persamaan 5 berikut hasil akurasi yang didapatkan.

$$\text{Akurasi} = (\text{jumlah data hasil prediksi benar} : \text{jumlah keseluruhan data}) \times 100\% \quad (1)$$

$$= (130 : 200) \times 100\% = 0{,}65 \times 100\% = 65\%$$

Gambar 13 berikut hasil pengujian dari mendapatkan lokasi.

Israel	ata dan utilitas curian dari Cellebrite perusahaan forensik mobile asal Israel dipaparkan peretas - Lokasi	Israel	TRUE
Hutan	nfo Hutan Alhamdulillah Sapi Curian Berhasil Ditemukan di Hutan Okezone Okezoneu - Lokasi	Okezoneu	FALSE
Tidak Ditemukan	angan mobil curian atau aset gelap Dibakar supaya nggak bisa dilacak - Lokasi	Tidak Ditemukan	TRUE
Tidak Ditemukan	angan mobil curian atau aset gelap Dibakar supaya nggak bisa dilacak - Lokasi	Tidak Ditemukan	TRUE
	KLASIFIKASI BENAR		130
	KLASIFIKASI SALAH		70
	AKURASI		65%

Gambar 13. Hasil mendapatkan lokasi dengan Ina-NLP

4. Kesimpulan

Berdasarkan hasil pembahasan pada penelitian sistem klasifikasi, maka dapat ditarik kesimpulan sebagai berikut:

1. Data *tweets* didapatkan dengan mengimplementasikan library *Twitter4j* dengan menambahkan *consumer key, consumer secret, token key* serta *token secret* yang telah disediakan oleh pihak *Twitter.* Dalam mendapatkan data *tweets,* kata kunci yang digunakan dalam pencarian telah terlampir di bab III. Data yang didapatkan sangat kotor, sehingga perlu dilakukan pembersihan data yang akurat untuk mendapatkan teks yang hanya berisi cuitan pengguna *Twitter.*
2. Tahap *preprocessing* yang digunakan untuk pembersihan data hasil *crawling* antara lain menghapus *username,* menghapus *hastag,* menghapus *link,* menghapus *ReTweet,* menghapus *punctuation* (tanda baca) serta menghapus kata yang tidak dibutuhkan untuk proses selanjutnya. Dalam tahap *preprocessing,* terdapat kekurangan yaitu kesalahan pada hasil *preprocessing* saat dilakukan pembersihan data, ada beberapa kata ataupun tanda baca yang tidak dapat dibersihkan secara maksimal. Dari kesalahan tersebut dapat mempengaruhi hasil dari ekstraksi fitur *bag of words.*
3. Dalam implementasi algoritma *NBC* terhadap sistem klasifikasi, penulis menggunakan *tools* WEKA sebagai jembatan antara bahasa pemrograman yang digunakan dan program untuk melakukan klasifikasi. Dengan menggunakan *library weka-stable,* seluruh kegiatan dalam klasifikasi dapat dilakukan dengan baik. Dalam proses klasifikasi, terdapat 2 fitur yang digunakan yaitu fitur Sintaktik dan fitur *bag of words.*
4. *Named Entity* berbasis aturan *(rule)* dari *inaNLP* merupakan teknik yang digunakan untuk mendapatkan lokasi dari data *tweets* hasil klasifikasi. Dalam proses mendapatkan lokasi, terdapat 2 *library* yang digunakan yaitu iPostagger.jar & inaNLP.jar.
5. Dari pembahasan dan impelementasi yang dipaparkan, didapatkan hasil pengujian dari sistem klasifikasi sebagai beriku:
   - Dari hasil pengujian model untuk mengetahui kinerja algoritma *Naive Bayes* dengan fitur Sintaktik menggunakan *K-Cross Validation* 10 kali iterasi menunjukkan hasil *accuracy*

sebesar 88,1398%. Dan untuk *accuracy* dengan fitur *Bag of Words* menunjukkan hasil sebesar 79,5%. Dari kedua hasil tersebut, dapat dikatakan model layak untuk digunakan ke dalam sistem klasifikasi karena mempunyai hasil *accuray* diatas 70%.

- Dari hasil pengujian klasifikasi menggunakan 200 *data test* atau data uji dengan ekstraksi fitur Sintaktik dapat disimpulkan bahwa mengimplementasikan fitur berdasarkan susunan kalimat menunjukkan hasil yang optimal dengan *accuracy, precision, recall* adalah 92%, 89.7%, 92%. Implementasi fitur Sintaktik juga dipengaruhi oleh ketepatan dalam memilih kata kunci yang digunakan.
- Dari hasil pengujian klasifikasi menggunakan 200 *data test* atau data uji dengan ekstraksi fitur *Bag of Words* dapat disimpulkan bahwa mengimplementasikan fitur dengan menggunakan seluruh kata yang ada pada *dataset* menunjukkan hasil cukup optimal dengan hasil *accuracy, precision, recall* adalah 76.5%, 78.4%, 76.5%. Implementasi fitur *Bag of Words* dipengaruhi oleh variasi kata dalam *dataset* yang digunakan.

6. Hasil pengujian mendapatkan lokasi menggunakan 200 *data test* dari hasil klasifikasi menunjukkan nilai akurasi sebesar 65%.

Dari hasil yang diperoleh terdapat beberapa kekurangan, sehingga perlu adanya pengembangan. Saran pengembangan dari peneliti adalah sebagai berikut:

1. Pengembangan terhadap proses pembersihan data atau *preprocessing data.* Dimana pada proses pembersihan data ini pengembang dapat menghapus kata yang tidak penting secara otomatis serta menghapus tanda baca garis miring (\) & petik koma (“”).
2. Pengembang dapat mencoba melakukan proses *stemming* dan *stopword removal* untuk mengetahui apakah proses tersebut dapat mempengaruhi hasil dari klasifikasi atau tidak.
3. Pengembang dapat memvisualisasikan *Google Maps* dari hasil mendapatkan lokasi.

Daftar Notasi

P = Probabilitas	TN = True Negative	TP = True Positive
X = Evidence	FN = False Negative	
Y = Kelas	FP= False Positive	

Referensi

[1] D. A. N. K. Com, “Menggunakan Twitter Studi Kasus Detik . COM,” p. 2015, 2015.
[2] S. D. Prihadi, “Berapa Jumlah Pengguna Facebook dan Twitter di Indonesia?,” 2015. [Online]. Available: https://www.cnnindonesia.com/teknologi/20150327061134-185-42245/berapa-jumlah-pengguna-facebook-dan-twitter-di-indonesia/. [Accessed: 03-Oct-2016].
[3] D. B. Prasetyo, “Sistem Informasi Geografis Berbasis Google Maps API Untuk Pemetaan Profil Kriminalitas Tipe Konvensional Di Wilayah Hukum Polresta Yogyakarta,” 2014.
[4] A. Darmawan, “Aplikasi Pengklasifikasian Dokumen Info Pada Twitter Menggunakan Algoritma Naive Bayes,” Malang, 2014.
[5] S. Rodiyansyah and E. Winarko, “Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification,” *Indones. J. Comput. Cybern. Syst.*, vol. 6, no. 1, pp. 91–100, 2012.
[6] L. Jayanti, S. R. Sentinuwo, O. A. Lantang, and A. Jacobus, “Analisa Pola Penyalahgunaan Facebook Sebagai Alat Kejahatan Trafficking Menggunakan Data Mining,” vol. 8, no. 1, pp. 2301–8364, 2016.
[7] K. Makhtidi, “Sistem SMS Spam Detector Untuk SMS Berbahasa Indonesia Pada Smartphone Android,” 2012
[8] B. Loni, “Enhanced Question Classification with Optimal Combination of Features: A New Approach on Automated Question Answering Systems,” *Pattern Recognit.*, 2012.
[9] Y. Ganisaputra and R. Tan, “Naïve Bayes Classifier,” pp. 173–188.
[10] E. Prasetyo, *Data Mining-Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: ANDI, 2012.
[11] A. Purwarianti, A. Andhika, A. F. Wicaksono, I. Afif, and F. Ferdian, “InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification,” *4th IGNITE Conf. 2016 Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2016*, no. August, 2016.

ISSN: 2714-7975; E-ISSN: 2716-1382