KLASIFIKASI *CYBERBULLYING* BERBASIS *TEXT MINING* MENGGUNAKAN *BAYES OPTIMAL CLASSIFIER*

SKRIPSI

Disusun Oleh:

IVAN BAGOES PRAMADHANI

H76216038

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SUNAN AMPEL

SURABAYA

2020

LEMBAR PERSETUJUAN PEMBIMBING

Skripsi Oleh

Nama	:	IVAN BAGOES PRAMADHANI
NIM	:	H76216038
JUDUL	:	KLASIFIKASI *CYBERBULLYING* BERBASIS *TEXT MINING* MENGGUNAKAN *BAYES OPTIMAL CLASSIFIER*

Ini telah diperiksa dan disetujui untuk diujikan.

Surabaya, 9 Juli 2020

Dosen Pembimbing 1	Dosen Pembimbing 2
Achmad Teguh Wibowo, MT	Ahmad Yusuf, M.Kom
NIP. 198810262014031003	NIP. 199001202014031003

PENGESAHAN TIM PENGUJI SKRIPSI

Skripsi Ivan Bagoes Pramadhani ini telah dipertahankan
Di depan tim penguji skripsi
Di Surabaya, 13 Juli 2020

Mengesahkan,
Dewan Penguji

Dosen Penguji 1	Dosen Penguji 2
Achmad Teguh Wibowo, MT	Ahmad Yusuf, M.Kom
NIP. 198810262014031003	NIP. 199001202014031003
Dosen Penguji 3	Dosen Penguji 4
Muhammad Andik Izzuddin, MT	Khalid, M.Kom
NIP. 198403072014031001	NIP. 197906092014031002

Mengetahui,
Plt. Dekan Fakultas Sains dan Teknologi
UIN Sunan Ampel Surabaya

Dr. Evi Fatimatur Rusydyah, M.Ag
NIP. 197312272005012003

LEMBAR PERNYATAAN KEASLIAN

Saya yang bertanda tangan dibawah ini,

Nama	:	Ivan Bagoes Pramadhani
NIM	:	H76216038
Program Studi	:	Sistem Informasi
Angkatan	:	2016

Menyatakan bahwa saya tidak melakukan plagiat dalam penulisan skripsi saya yang berjudul: "KLASIFIKASI *CYBERBULLYING* BERBASIS *TEXT MINING* MENGGUNAKAN *BAYES OPTIMAL CLASSIFIER*". Apabila suatu saat nanti terbukti saya melakukan tindakan plagiat, maka saya bersedia menerima sanksi yang telah ditetapkan.

Demikian pernyataan keaslian ini saya buat dengan sebenar-benarnya.

Surabaya, 20 Juli 2020

Yang menyatakan,

Ivan Bagoes Pramadhani

NIM. H76216038

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL SURABAYA
PERPUSTAKAAN
Jl. Jend. A. Yani 117 Surabaya 60237 Telp. 031-8431972 Fax.031-8413300
E-Mail: perpus@uinsby.ac.id

LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI
KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika UIN Sunan Ampel Surabaya, yang bertanda tangan di bawah ini, saya:

Nama : IVAN BAGOES PRAMADHANI
NIM : H76216038
Fakultas/Jurusan : SAINS DAN TEKNOLOGI / SISTEM INFORMASI
E-mail address : ivanbagoespramadhani@gmail.com

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Perpustakaan UIN Sunan Ampel Surabaya, Hak Bebas Royalti Non-Eksklusif atas karya ilmiah :
■ Sekripsi ☐ Tesis ☐ Desertasi ☐ Lain-lain (……)
yang berjudul :
KLASIFIKASI *CYBERBULLYING* BERBASIS *TEXT MINING* MENGGUNAKAN *BAYES OPTIMAL CLASSIFIER*

beserta perangkat yang diperlukan (bila ada). Dengan Hak Bebas Royalti Non-Ekslusif ini Perpustakaan UIN Sunan Ampel Surabaya berhak menyimpan, mengalih-media/format-kan, mengelolanya dalam bentuk pangkalan data (database), mendistribusikannya, dan menampilkan/mempublikasikannya di Internet atau media lain secara ***fulltext*** untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan atau penerbit yang bersangkutan.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Perpustakaan UIN Sunan Ampel Surabaya, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini yang saya buat dengan sebenarnya.

Surabaya, 05 Agustus 2020

Penulis

(IVAN BAGOES PRAMADHANI)

ABSTRAK

KLASIFIKASI *CYBERBULLYING* BERBASIS *TEXT MINING* MENGGUNAKAN *BAYES OPTIMAL CLASSIFIER*

Oleh:

Ivan Bagoes Pramadhani

Pengaruh perkembangan teknologi yang semakin mudah digunakan, membuat banyak efek negatif yang bermunculan. Salah satunya adalah tindakan *cyberbullying* yang dilakukan diberbagai media *online*, seperti media sosial dan *online game*. Tidak jarang pemain *online game* melakukan tindakan tersebut, sehinggga dapat meresahkan individu maupun negara. Untuk menindak lanjuti permasalahan tersebut diterapkan sebuah sistem klasifikasi teks untuk mengetahui teks *cyberbullying*. Sehingga pada penelitian ini dilakukan beberapa proses seperti pengambilan data, pengolahan data dan analisis perhitungan performa dari metode klasifikasi *Bayes Optimal Classifier* (BOC) dan metode *Bernoulli Naïve Bayes* (BNB). Hasil penelitian dengan menggunakan tiga macam skenario proporsi data yakni pada sekenario pertama data sebesar 40% data latih dan 60% data uji, sekenario kedua sebesar 50% data latih dan 50% data uji, dan sekenario ketiga sebesar 60% data latih dan 40% data uji, didapatkan hasil akurasi metode BOC terbesar adalah 81,11% pada skenario ketiga dengan proporsi data sebesar 60% data latih dan 40% data uji. Sedangkan hasil pada skenario yang sama, nilai akurasi dari metode BNB sebesar 73,33%. Dari hasil analisis yang diperoleh, diketahui bahwa metode BOC mendapat nilai akurasi lebih tinggi daripada metode BNB.

Kata Kunci: *Bayes Optimal Classifier, Cyberbullying, Cyberathlete,* Klasifikasi Teks, *Naive Bayes*.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

ABSTRACT

CYBERBULLYING CLASSIFICATION BASED ON TEXT MINING USING BAYES OPTIMAL CLASSIFIER

By:

Ivan Bagoes Pramadhani

The influence of technological developments is increasingly easy to use, making many negative effects emerge. One of them is the action of cyberbullying conducted in various online media, such as social media and online games. It is not uncommon for online gamers to do these actions, so that individuals and countries can be unsettling. To follow up the issue is applied a text classification system to find out the text of cyberbullying. So in this research, there are several processes such as data retrieval, data processing and analysis of performance calculations from Bayes Optimal Classifier (BOC) Classification method and Bernoulli Naïve Bayes (BNB) method. The results of the study using three scenarios of data proportions are in the first scenario of data is 40% of training data and 60% test data, the second scenario is 50% of training data and 50% test data, and the third scenario is 60% of training data and 40% test data, obtained by the largest BOC method accuracy is 81.11% in third scenario with data proportion of 60% training data and 40% test data. While the results in the same scenario, the accuracy value of the BNB method amounted to 73.33%. From the results of the analysis obtained, it is known that the BOC method gets a higher accuracy than the BNB method.

Keywords: *Bayes Optimal Classifier, Cyberbullying, Cyberathlete, Naïve Bayes Classifier, Text Classification.*

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

DAFTAR ISI

HALAMAN JUDUL i
LEMBAR PERSETUJUAN PEMBIMBING iii
LEMBAR PERNYATAAN KEASLIAN v
MOTTO vi
KATA PENGANTAR vii
ABSTRAK ix
ABSTRACT x
DAFTAR ISI xi
DAFTAR TABEL xiii
DAFTAR GAMBAR xiv
1 BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Batasan Masalah 4
1.4 Tujuan Penelitian 4
1.5 Manfaat Penelitian 4
1.5.1 Akademik 4
1.5.2 Praktis 5
2 BAB II TINJAUAN PUSTAKA 6
2.1 Tinjauan Penelitian Terdahulu 6
2.2 Teori-teori Dasar 7
2.2.1 *Cyberbullying* 7
2.2.2 *Text Mining* 8
2.2.3 *Pre-processing* Teks 9
2.2.4 *Bayes Optimal Classifier* 11
2.2.5 Dota 2 12
2.2.6 Kamus SentiWordNet 13
2.2.7 *Bag of Words* 13
2.2.8 *Confusion Matrix* 13
2.3 Integrasi Keilmuan 15
3 BAB III METODOLOGI PENELITIAN 17

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

3.1 Jenis dan Sumber Data 17
3.2 Metodologi Penelitian 17
3.2.1 Tahapan Awal 18
3.2.2 Data *Processing* 19
3.2.3 Klasifikasi Teks 22
3.2.4 Analisis Hasil 22
4 BAB IV HASIL DAN PEMBAHASAN 23
4.1 Pengambilan Data 23
4.2 Data *Processing* 26
4.2.1 *Cleansing* 26
4.2.2 *Spelling Check* 26
4.2.3 *Labeling* 27
4.2.4 *Text Pre-processing* 28
4.3 Klasifikasi Teks 31
4.3.1 *Bag of Words* 32
4.3.2 *Bayes Optimal Classifier* 33
4.4 Analisis Hasil 36
4.4.1 Skenario pertama data *train* 40% dan data *test* 60% 36
4.4.2 Skenario kedua data *train* 50% dan data *test* 50% 38
4.4.3 Skenario ketiga data *train* 60% dan data *test* 40% 40
4.4.4 Hasil Klasifikasi 42
5 BAB V PENUTUP 44
5.1 Kesimpulan 44
5.2 Saran 45
DAFTAR PUSTAKA 46

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

DAFTAR TABEL

Tabel 3.1 Contoh Kata Dengan Penulisan Salah (*Typo*) 20
Tabel 4.1 Tahap *Cleansing* 26
Tabel 4.2 *Spelling Check* 27
Tabel 4.3 Perubahan *Positive Words* 27
Tabel 4.4 Perubahan *Negative Words* 27
Tabel 4.5 Contoh Proses *Filtering* 29
Tabel 4.6 Contoh Proses *Tokenizing* 29
Tabel 4.7 Contoh Proses *Stopword Removal* 30
Tabel 4.8 Contoh Proses *Stemming* 30
Tabel 4.9 Contoh Proses *Lemmatization* 31
Tabel 4.10 Dataset BoW 32
Tabel 4.11 Frekuensi kemunculan kata 32
Tabel 4.12 *Bag of Words* 33
Tabel 4.13 Contoh *Dataset* Siap Untuk Proses Klasifikasi 33
Tabel 4.14 Pemetaan Kemunculan Kata 34
Tabel 4.15 Nilai Probabilitas *Term* 35
Tabel 4.16 Hasil *Confusion Matrix* Skenario Pertama Metode BOC 36
Tabel 4.17 Hasil *Confusion Matrix* Skenario Pertama Metode BNB 37
Tabel 4.18 Perbandingan Hasil Metode BOC dan BNB Skenario Pertama 38
Tabel 4.19 Hasil *Confusion Matrix* Skenario Kedua Metode BOC 38
Tabel 4.20 Hasil *Confusion Matrix* Skenario Kedua Metode BNB 39
Tabel 4.21 Perbandingan Hasil Metode BOC dan BNB Skenario Kedua 40
Tabel 4.22 Hasil *Confusion Matrix* Skenario Ketiga Metode BOC 40
Tabel 4.23 Hasil *Confusion Matrix* Skenario Ketiga Metode BNB 41
Tabel 4.24 Perbandingan hasil Metode BOC dan BNB Skenario Ketiga 42
Tabel 4.25 Hasil dari Ketiga Skenario 42

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

DAFTAR GAMBAR

Gambar 2.1 *Confusion Matrix* 14
Gambar 3.1 *Crawling* Data *Cyberathlete* 17
Gambar 3.2 *History Chat Cyberathlete* 17
Gambar 3.3 Alur Penelitian Skripsi 18
Gambar 3.4 Contoh Data Hasil *Labeling* 21
Gambar 4.1 *Crawling* Data *Cyberathlete* Dota2 23
Gambar 4.2 Data *Cyberathlete* Dota2 24
Gambar 4.3 Data Variabel account_id 24
Gambar 4.4 *Crawling* Data *History Chat* 24
Gambar 4.5 C*rawling* Data *History* C*hat* 25
Gambar 4.6 *History* C*hat* Dalam Bentuk Excel 25
Gambar 4.7 Hasil Pengambilan Data *History* C*hat Cyberathlete* 26
Gambar 4.8 Proses *Labeling* 28
Gambar 4.9 *Wordcloud* dengan Label *Non-Cyberbullying* dan *Cyberbullying* 31
Gambar 4.10Diagram Nilai *Accuracy* Metode BOC dan BNB 42

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pengaruh perkembangan teknologi yang semakin mudah untuk digunakan, membuat banyak sekali efek negatif yang bermunculan (Kartiwi & Gunawan, 2019). Salah satu efek negatif tersebut adalah tindakan kejahatan yang paling merajalela dalam kegiatan *online*, terutama di kalangan anak muda yakni *cyberbullying* (Orel et al., 2017). *Cyberbullying* dikenal sebagai tindakan *bullying* yang dilakukan di dunia maya (Luqyana et al., 2018). *Bullying* adalah tindakan mengintimidasi yang membuat depresi, tertekan, dan rendahnya harga diri korbannya (Cénat et al., 2014). Umumnya, tindakan *cyberbullying* dilakukan oleh pelaku pada suatu media *online*. Namun tindakan ini sudah merambah pada suatu profesi seseorang, profesi tersebut adalah atlet.

Atlet merupakan suatu profesi seseorang yang menekuni suatu cabang olahraga tertentu dan berprestasi, di mana dibutuhkannya bakat, kekuatan, serta ketangkasan yang berguna dalam berkompetisi (Zufri, 2017). Di Indonesia, atlet merupakan salah satu profesi yang digemari oleh anak muda. Dari data yang didapatkan pada *Southeast Asian Games* (SEA Games) 2019 yang dilaksanakan di Filipina, Indonesia mengirim 21 atlet untuk mengikuti enam cabang olahraga *Electronic Sport* (eSport). SEA Games merupakan ajang olahraga yang mengikut sertakan negara-negara yang termasuk dalam benua Asia Tenggara. Pada SEA Games 2019, dipublikasikan pertama kali cabang olahraga baru yaitu cabang eSport. eSport ini merupakan suatu kegiatan bertanding untuk mengadu ketangkasan antar individu maupun kerjasama antar kelompok yang tidak terbatas dari kegiatan fisik yang menggunakan suatu alat yang digunakan secara elektronik (Karhulahti, 2017). eSport ini diikuti oleh para *Cyberathlete* yang merupakan sebutan baru untuk atlet dari cabang olahraga eSport.

Cyberathlete merupakan seseorang atau kelompok yang mengikuti suatu kompetisi permainan komputer maupun *mobile* di mana oleh masyarakat dianggap sebagai profesional yang memiliki bakat untuk mendapatkan uang atau mewakili negaranya dalam kompetisi multipemain (*Urban Dictionary: Cyberathlete*, 2004).

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Cyberathlete biasa berkompetisi di permainan komputer maupun *mobile games* atau yang biasa disebut dengan *Online Video Game* (*OVG*). *OVG* merupakan sebuah permainan yang dapat diakses oleh banyak pemain dengan cara menghubungkan beberapa mesin atau perangkat ke suatu jaringan (Adams & Rollings, 2010). Kompetisi eSport pada SEA Games 2019 telah dilakukan pertandingan *game online* seperti Mobile Legends, Arena of Valor (AOV), Heatstone, Starcraft 2, Tekken 7, dan Dota 2.

Di dalam permainannya, seorang pemain biasa maupun pemain profesional tidak luput untuk melakukan tindakan *cyberbullying*. Umumnya, tindakan ini dilakukan karena pemain tersebut menjumpai suatu kejadian yang menurutnya tidak menyenangkan, semisal menemui pemain lain yang melakukan tindakan "*trolling*", "*Away From Keyboard*", dan "*skill"* pemain lain yang minim. *Cyberbullying* memiliki banyak efek negatif, contohnya ada suatu berita berjudul "Seorang anak melakukan tindakan bunuh diri dikarenakan di*bully* oleh temannya di dalam game". Tidak hanya itu, masih banyak lagi efek negatif dari tindakan *cyberbullying*. Dengan permasalahan tersebut penelitian ini melakukan pengukuran performa dari metode klasifikasi yang digunakan untuk mengklasifikasikan tindakan *cyberbullying* pemain Dota2 dengan cara mengklasifikasikan kata-kata yang pernah dilontarkan oleh pemain tersebut. Kedepannya bahwa penelitian ini dapat diajukan kepada pihak Dota2 maupun pemerintahan untuk menambahkan peraturan baru dan sanksi bagi pelaku tindak *cyberbullying*.

Untuk mengetahui jenis dari suatu kata ataupun kalimat yang termasuk dalam *cyberbullying* maka digunakanlah algoritma *text mining* yang menggunakan metode klasifikasi. Metode klasifikasi ini memiliki macam-macam metode yang dapat digunakan untuk mengklasifikasikan hasil dari *text mining* diantaranya, yaitu *bernaulli naïve bayes (NBC), random forest classifier (RFC), decision tree (DT),* dan *support vector classifier (SVC).* Dari penelitian yang dilakukan oleh (Pranckevičius & Marcinkevičius, 2017) dengan judul "*Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification*" bahwa metode *DT* memiliki tingkat keakuratan paling rendah dibandingkan metode klasifikasi lainnya. Sedangkan untuk metode *BNB* memiliki tingkat akurasi paling tinggi, disusul

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

dengan metode *RFC*, kemudian metode *SFC* dengan selisih yang tidak jauh berbeda sebanyak 1 hingga 2%.

Metode *NBC* merupakan metode yang dapat memiliki nilai keakurasian yang tinggi, namun metode ini memiliki banyak kelemahan dan kelebihan. Pada penelitian yang dilakukan oleh (Sayfullina et al., 2015), metode klasifikasi yang digunakan adalah metode *Normalized Bernoulli Naive Bayes* (*nBNB*) yang merupakan metode dari *naive bayes*. Dalam penelitiannya, mengatakan bahwa metode *Bernoulli Naive Bayes* (*BNB*) memiliki nilai *False Positive Rate* (*FPR*) yang tinggi sehingga dengan menormalkan nilai dari *FPR*, metode BNB menjadi lebih baik dan dinamakan metode *nBNB*. Sedangkan penelitian oleh (Fadlil et al., 2017) menggunakan metode *Gaussian Naive Bayes* (*GNB*) atau bisa disebut dengan metode *Bayes Optimal Classifier* (*BOC*). Dalam penelitiannya metode klasifikasi *BOC* dapat memperbaiki nilai *error* dari nilai *FPR* dan *False Negative Rates* (*FNR*) yang tinggi. Sehingga dari beberapa penelitian terdahulu yang disebutkan di atas, apakah metode *BOC* ini dapat digunakan dalam klasifikasi hasil dari *text mining* yang kemudian hasilnya akan dibandingkan dengan metode *BNB*.

Dari data hasil *crawling* menggunakan *Application Programming Interface* (API) Dota2, didapatkan data berupa *history chat* tiap pemain yang selanjutnya dilakukan tahap *preprocessing* teks. Di dalam tahap *preprocessing* teks hanya dilakukan dua proses, yakni *text filtering* dan *stemming*. Kemudian dilakukan pelabelan kata yakni *cyberbullying* dan *non-cyberbullying* menggunakan kamus *sentiword* yang selanjutnya dilakukan klasifikasi menggunakan metode *Bayes Optimal Classifier*. Berawal dari permasalahan yang tertera pada latar belakang di atas, maka perlu dilakukannya penelitian ini yang berjudul **"KLASIFIKASI *CYBERBULLYING* BERBASIS *TEXT MINING* MENGGUNAKAN *BAYES OPTIMAL CLASSIFIER*"**. Sehingga, penelitian ini dapat menghasilkan informasi yang akan ditampilkan berupa nilai akurasi dari masing-masing proporsi metode.

1.2 Rumusan Masalah

Dengan mengacu permasalahan pada latar belakang di atas, dirumuskan rumusan masalah seperti berikut:

1. Bagaimana cara mengimplementasikan algoritma *Bayes Optimal Classifier* untuk mengklasifikasi *cyberbullying* yang dilakukan oleh *cyberathlete*?

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

2. Bagaimana performa *Bayes Optimal Classifier* dalam klasifikasi *cyberbullying* oleh *cyberathlete* jika dibandingkan dengan metode *Bernoulli Naive Bayes*?

1.3 Batasan Masalah

Batasan masalah yang membatasi penelitian pada penelitian ini yaitu:

1. Jumlah data sebanyak 300 *cyberathlete* Dota2.
2. Data yang digunakan adalah data *history chat* tiap pemain hasil *crawling* dari game Dota2.
3. Penelitian ini hanya melakukan *text mining* terhadap *chat player* yang berbahasa Inggris.
4. Metode yang digunakan untuk klasifikasi adalah *Bayes Optimal Classifier*.
5. Metode pembanding yang digunakan adalah metode *Bernoulli Naïve Bayes.*
6. Implementasi bahasa pemrograman yang digunakan untuk klasifikasi menggunakan bahasa pemrograman python.

1.4 Tujuan Penelitian

Sesuai dengan rumusan masalah yang ada, tujuan dari penelitian ini adalah sebagai berikut:

1. Untuk mengetahui bagaimana mengklasifikasikan *cyberbullying* menggunakan algoritma *Bayes Optimal Classifier*.
2. Untuk mengetahui hasil performa metode klasifikasi *cyberbullying* yang dilakukan oleh *cyberathlete.*

1.5 Manfaat Penelitian

Manfaat yang didapatkan dari penelitian ini yang berdampak luas baik dari segi akademik maupun dalam segi aplikatif.

1.5.1 Akademik

1. Sebagai salah satu bahan rujukan dalam bidang *text mining* yang menganalisis *Cyberbullying*.
2. Sebagai salah satu bahan rujukan dalam penelitian mengenai klasifikasi teks.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

1.5.2 Praktis

Membantu pihak Dota2 maupun pihak pemerintahan untuk menambahkan peraturan baru dan sanksi dalam tindakan *cyberbullying* yang dilakukan oleh pemain dan *cyberathlete*nya.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Penelitian Terdahulu

Sebelum melakukan penelitian, penulis mencari referensi terdahulu yang serupa untuk mendapatkan pengetahuan dalam menganalisis data dan mengolahnya sehingga yang nantinya akan membantu penelitian ini. Berdasarkan penelitian dengan judul "*Cyberbullying in E-Learning Platform : An Exploratory Study Through Text Mining*" menyatakan bahwa penggunaan *text mining* dalam *Platform E-Learning* dapat mendeteksi tindakan *cyberbullying* di dalamnya (Kartiwi & Gunawan, 2019). Dalam penelitian ini didapatkan solusi praktis untuk meningkatkan kesadaran terhadap *cyberbullying* dan dampak negatif secara psikologis kepada para korbannya.

Penelitian dengan judul "*Cyberbullying Detection in Social Networks: A Multi-Stage Approach*", pada penelitiannya dengan menggunakan metode *Multi-Stage Approach* untuk mendeteksi tindakan *cyberbullying* di sosial media. (Del Bosque & Garza, 2018). Selain itu, metode *Timeseries, Big Data,* dan *deep learning* dapat dimasukkan ke dalam metode *approach* untuk membuatnya lebih terukur, kuat, dan mampu menangani sejumlah besar data (Del Bosque & Garza, 2018).

Penelitian ketiga dengan judul "*Analysis of the Indonesian Cyberbullying through Data Mining: The Effective Identification of Cyberbullying through Characteristics of Messages*", dari penelitian tersebut didapatkan bahwa penggunaan *supervised learning* seperti *naïve* bayes, *decision tree,* dan *neural network* dapat dihubungkan secara sinkron untuk menganalisis data (Margono, 2019). Proses analisis menggunakan model *performance* untuk mencapai tingkat akurasi yang tinggi dalam prediksi data. Hasil dari analisis model mampu mengidentifikasi 80,37% *cyberbullying messages* dan 19,63% non-*cyberbullying messages.*

Penelitian selanjutnya yang berjudul "Teknik Data *Mining* Menggunakan Metode *Bayes Classifier* Untuk Optimalisasi Pencarian Pada Aplikasi Perpustakaan", dilakukan klasifikasi data menggunakan metode *bayes classifier.*

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Hasil akhir menunjukan bahwa metode *bayes classifier* dapat digunakan untuk proses mengklasifikasikan data hasil dari data *mining* (Ginting & Trinanda, 2015)

Penelitian terakhir dengan judul "*On the Rates of Convergence from Surrogate Risk Minimizers to the Bayes Optimal Classifier*". Menyatakan bahwa, penggunaan metode *Bayes Optimal Classifier* dalam perbandingan tingkat konvergensi dapat meminimalisir data posteriori dari berbagai macam algoritma klasifikasi (Zhang et al., 2018).

Dengan melihat dari beberapa penelitian yang pernah dilakukan dapat ditarik kesimpulan, bahwa penelitian terdahulu sudah melakukan penelitian dengan topik *Text Mining* untuk mengetahui tindakan *Cyberbullying* yang dilakukan pengguna internet di dunia maya. Oleh karena itu penelitian pada skripsi ini bertemakan *text mining*, dengan menggunakan metode klasifikasi teks yakni metode *Bayes Optimal Classifier.*

2.2 Teori-teori Dasar

2.2.1 *Cyberbullying*

Cyberbullying merupakan tindakan *bullying* yang dilakukan di dunia maya (Luqyana et al., 2018). Karakteristik dari pesan *cyberbullying* dapat diketahui dari isi pesannya, di mana karakteristik tersebut berupa kata positif atau negatif (Sanders et al., 2009). Bentuk-bentuk dari *cyberbullying* menurut (Willard, 2006) seperti berikut:

1. *Flaming*: mengirim pesan *online* berisi kata-kata kotor dan vulgar dalam mengekspresikan kemarahan.
2. *Harassment*: mengirim pesan berisi ejekan dan kata-kata kotor bekali-kali secara *online*.
3. *Denigration*: memberikan komentar pedas yang kejam dan tidak adil.
4. *Impersonation*: menggunakan identitas seseorang untuk melakukan pelecehan.
5. *Outing and Trickery*: membagikan informasi yang memalukan korban dan kemudian membuat korban menunjukan informasi personalnya.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

6. *Cyberstalking*: mengirim pesan secara *online* berkali-kali untuk mengintimidasi korban sehingga membuat korban takut untuk kesalamatannya.

Selain ke enam bentuk dari *cyberbullying* di atas, karakteristik yang dapat dilihat dari *cyberbullying* berikut seperti: Ketidak jelasan identitas pelaku (O'Brien & Moules, 2010), waktu pengiriman pesan ancaman (Privitera & Campbell, 2009), dan makna sesungguhnya dari isi pesan (Mishna et al., 2009) dapat mengidentifikasi bahwa pesan tersebut merupakan *cyberbullying*.

2.2.2 *Text Mining*

Text mining adalah suatu reka cipta baru yang sebelumnya belum diketahui oleh komputer yang secara otomatis mengekstraksi informasi dari data berbentuk teks yang berkualitas tinggi dan berasal dari sumber yang berbeda-beda (Hearst, 2003). Teknik *classifying* adalah suatu teknik yang dapat digunakan dalam pemrosesan data hasi dari *text mining* di mana mengelompokan data berdasarkan kesamaan jenis dalam beberapa kelas. *Text mining* memiliki tiga proses tahapan, proses itu adalah *text-preprocessing, text transformation,* dan *pattern discovery* (Even, 2002).

Fokus utama dalam *text mining* adalah dalam *Natural Language Processing*, di mana aspek ini juga dilakukan dalam *empirical computer linguistic*. Ada beberapa masalah dalam memproses komputasi linguistik di mana hal ini dibahas pada *information retrieval.*

a. *Information Retrieval*

Pencarian Informasi atau *Information Retrieval* adalah metode yang digunakan untuk mencari teks dari sekumpulan data teks yang tidak terstruktur untuk mendapatkan informasi yang dibutuhkan (Manning et al., 2008). *Information Retrieval* memiliki beberapa konsep dalam menganalisis teks diantaranya:

1. *Vector*, dengan cara memetakan teks dan query ke model ruang vector.
2. *Similarity*, yaitu dengan menghitung persamaan teks dan query. Sehingga sebagian besar perhitungan frekuensi dari query kata akan menghasilkan nilai sama dengan 0.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

3. *Indexing*, merupakan pemrosesan teks yang telah terindeksi sebelumnya. Proses ini dilakukan ketika pencocokan dengan indeks pencarian, sehingga menghindari memproses teks yang sebenarnya selama pencarian. Penerapan konsep ini telah digunakan dalam mesin pencari web seperti google dan sejenisnya.
4. *Filtering*, terdapat dua tingkatan di mana pada tingkat pertama adalah *Text Filtering* dan tingkat selanjutnya adalah *Passage Retrieval*. *Text Filtering* digunakan untuk mengklasifikasi teks yang akan digunakan atau tidak digunakan. *Passage Retrieval* digunakan untuk menentukan bagian dari teks yang sesuai dengan *query* yang di inputkan.

b. *Natural Language Processing*

Natural Language Processing merupakan sebuah proses komputasi otomatis maupun semi otomatis bahasa manusia. Proses ini digunakan untuk mengetahui sejauh mana komputasi dengan menggunakan bahasa manusia dapat berinteraksi dengan komputer. Proses NLP ini sering terjadi ambiguitas karena perkataan manusia sering kali ambigu dan tidak terstruktur karena penggunaan bahasa gaul dan dialek daerah. Oleh karena itu algoritma teks perlu untuk dapat masalah tersebut.

c. Data *Mining*

Data *mining* adalah istilah yang digunakan untuk pengambilan data di dalam suatu *database* dengan menggunakan teknik statistik, matematika, dan *machine learning* untuk menjadikan suatu informasi menjadi bermanfaat (Aggarwal & Zhai, 2012). Sedangkan data mining menurut Gartner Group, data *mining* adalah teknik dari gabungan beberapa ilmu seperti *machine learning*, *pattern discovery*, statistik, dan visualisasi dari pengambilan informasi pada suatu *database* (Daniel, 2005).

Data *mining* merupakan proses awal dari penerapan *machine learning*, di mana proses ini digunakan untuk menyiapkan *dataset* sebelum diproses untuk mempelajari isi dari *dataset*.

2.2.3 *Pre-processing* Teks

Pre-processing Teks merupakan tahapan terpenting dalam *text mining*. Oleh karena itu tahapan ini dilakukan untuk mengolah suatu teks di dalam dokumen agar dapat digunakan dalam proses klasifikasi. Pada dasarnya suatu dokumen yang akan

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

digunakan memiliki teks yang tidak jelas dan tidak berstruktur, sehingga perlu adanya pemrosesan data sebelum dilakukan proses klasifikasi secara langsung. Alasan lain, tidak semua kata yang akan diproses di dalam suatu dokumen dapat mencerminkan isi dari dokumen tersebut. Umumnya, tahapan *pre-processing* teks memiliki beberapa tahapan proses seperti berikut:

1. *Case Folding*

   Case Folding adalah tahapan yang harus dilakukan dalam pre-proses teks. *Case Folding* merupakan proses yang dilakukan untuk merubah teks pada suatu dokumen, yang awalnya menggunakan huruf kapital (*uppercase*) menjadi huruf kecil (*lowercase*).

2. *Cleansing*

   Proses kedua adalah *Cleansing* yang merupakan proses membersihkan dataset dari simbol-simbol yang biasa digunakan dalam tanda baca seperti titik, koma, tanda seru, at, hashtag, dan lain-lain.

3. *Tokenizing*

   Tokenizing adalah proses selanjutnya yang dapat dilakukan dalam pre-*processing* teks. *Tokenizing* merupakan proses memisahkan teks pada suatu kalimat maupun paragraf menjadi potongan atau bagian-bagian yang disebut *token* yang nantinya akan di analisis (Kalra & Aggarwal, 2018). Proses ini bertujuan untuk mempermudah dalam pembobotan kata.

4. *Text Filtering*

   Proses selanjutnya yaitu *text filtering* yang merupakan tahapan mengambil kata-kata yang tidak mempunyai makna (data *outlier*) dari hasil *tokenizing* menggunakan algoritma *stoplist* (Kalra & Aggarwal, 2018).

5. *Stemming*

   Stemming adalah tahapan dalam pre-proses teks yang harus dilakukan. *Stemming* adalah proses menghilangkan *inflection* dalam suatu kata ke bentuk dasarnya (Kalra & Aggarwal, 2018). Proses *stemming* dilakukan ketika telah mendapat hasil dari proses *filtering*.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

6. *Lemmatization*

Lemmatization merupakan proses menjadikan kata atau kalimat kebentuk kata dasar dan bentuk *lemma* nya. Proses ini biasa dilakukan dalam memproses teks berbahasa asing. Proses ini merupakan proses penyempurnaan dari pemrosesan kata pada tahap *stemming*.

2.2.4 *Bayes Optimal Classifier*

Bayes Optimal Classifier adalah metode untuk meminimalkan kemungkinan kesalahan klasifikasi terhadap estimasi kepadatan (Tong & Koller, 2000). *Bayes Optimal Classifier* juga dikenal sebagai *Bayes Optimal Learner*, *Bayes Classifier*, *Bayes Optimal Decision Boundary*, serta *Bayes Optimal Discriminant Function* (Brownlee, 2019). *Bayes Optimal Classifier* menganut teori dari bayes yang biasa disebut *Bayes Theorem*. *Bayes Theorem* memiliki dasar rumus sebagai berikut:

$$\boldsymbol{P(A|B) = \frac{(P(B|A) * P(A))}{P(B)}} \tag{1}$$

Di mana rumus tersebut dapat dibaca bahwa *probability* A sebagai B ditentukan dari *probability* B saat A, *probability* A, dan *probability* B. Pada pengaplikasiannya rumus tersebut akan berubah menjadi:

$$P(C_i|D) = \frac{P(D|C_i) * P(C_i)}{P(D)} \tag{2}$$

Bayes Optimal Classifier merupakan model penyederhanaan dan pengoptimalan dari metode *Bayes*, persamaan fungsinya seperti berikut:

$$\arg\max_{vj} \sum_{h_i \in H} (P(vj|h_i)\, P(h_i|D)) \tag{3}$$

Di mana *vj* merupakan sebuah contoh label yang perlu untuk diklasifikasikan, H merupakan set hipotesis untuk mengklasifikasikan *vj*, dan h_i salah satu hipotesis. Sedangkan D adalah jumlah data pada sebuah dokumen (Brownlee, 2019). Sehingga, rumus dari *Bayes Optimal Classifier* digunakan seperti berikut:

$$B_{MAP} = (P(vj|h_1) * P(h_1|D) + P(vj|h_2) * P(h_2|D) + \cdots + P(vj|h_n) * P(h_n|D)) \tag{4}$$

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Di mana untuk:

B_{MAP} = *Bayes Maximum a Priori*

$P(vj|h_i)$ = *Probability* kemunculan hipotesis pada suatu kelas

$P(h_i|D)$ = *Probability* kemunculan hipotesis pada suatu dokumen

vj = kelas klasifikasi

Langkah-langkah dalam klasifikasi teks menggunakan metode *Bayes Optimal Classifier* adalah:

1. Menyiapkan *Dataset*.
2. Menghitung probabilitas kemunculan hipotesis pada data *training*.

$$P(h_i|D) = \frac{jumlah\ dokumen\ dengan\ label\ h}{total\ dokumen} \quad (5)$$

3. Menghitung probabilitas *term* pada data *training*.
4. Menghitung probabilitas *class* dari data *training*.

$$P(vj|h_i) = \frac{py + 1}{n + term} \quad (6)$$

Di mana,

py = probabilitas kata y

n = total jumlah kata pada label h

term = jumlah total kata

py + 1 = probabilitas kata y ditambah 1 supaya hasil tidak 0 (nol)

5. Menghitung nilai B_{MAP}.
6. Jika nilai B_{MAP} h_1 lebih besar daripada h_2 maka kategori class h_1, sebaliknya.

2.2.5 Dota 2

Dota merupakan suatu permainan *online video game* bergaya *multiplayer online battle arena* (*MOBA*), yang merupakan versi lanjutan dari *Defense of the Ancient mod* pada *Warcraft 3 : Reign of Chaos dan Warcraft 3 : The Frozen Throne.* Steam mengadopsi game Dota untuk dapat dimainkan secara gratis dengan mengakses Steam *store.* Dota dikembangkan oleh Valve Corporation, terbit pada Juli tahun 2013 dan dapat dimainkan secara gratis pada *operation system* Windows, OS X, dan Linux (*Dota 2 Wiki*, 2013).

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Dota 2 memiliki *Application Programming Interface* (API) yang dapat diakses dan digunakan secara gratis (*Opendota API*, 2018). Di mana API ini dapat digunakan untuk membangun suatu program ataupun hanya digunakan untuk mencari data-data pemain Dota 2.

2.2.6 Kamus SentiWordNet

Kamus SentiWordNet merupakan kamus yang berisi kata-kata yang termasuk *negative* dan *positive*. Melihat sejarahnya, kamus *sentiword* dalam sentimen analisis dibuat pertama kali oleh Liu, pada awalnya hanya tersedia dalam bahasa Inggris, namun saat ini sudah tersedia dalam berbagai bahasa (Cahyani, 2019). Kamus *sentiword* ini merupakan kumpulan kata positif dan negatif. Dalam penelitian ini Kamus *Sentiword* yang digunakan berbahasa inggris untuk pelabelan kata *cyberbullying* yang didapatkan pada *website* SentiWordNet dengan sedikit perubahan untuk mencocokan dalam penggunaan *dataset*.

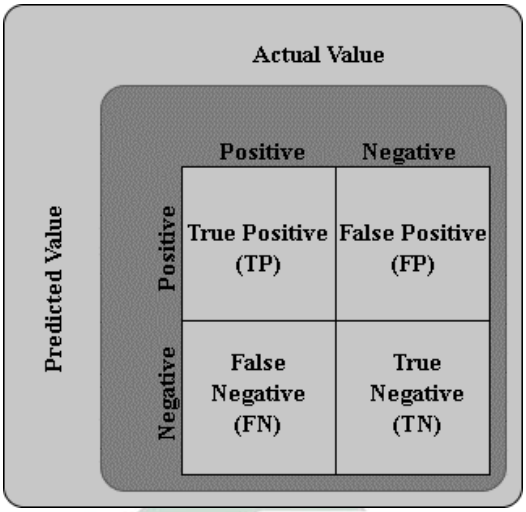
2.2.7 *Bag of Words*

Bag of Words adalah suatu model yang terbentuk dari proses pengumpulan kalimat dari dataset yang telah diolah dalam teks pre-proses sehingga membentuk menjadi satu kantung (*bag*). *Bag of Words* berguna dalam melihat frekuensi kemunculan kata (Raj & Rajaraajeswari, 2016).

2.2.8 *Confusion Matrix*

Confusion Matrix adalah suatu cara untuk menampilkan data prediksi dan kondisi aktual (kondisi sebenarnya), *confusion matrix* juga digunakan untuk menghitung *accuracy*, *precision*, dan *recall*. Berikut pada Gambar 2.1 merupakan *confusion matrix*.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id



Gambar 2.1 *Confusion Matrix*

Melihat Gambar 2.1 dari *confusion matrix* di atas dapat dibuat rumus untuk menghitung *accuracy*, *precision*, dan *recall*.

a. *Accuracy*

Accuracy atau akurasi adalah suatu rasio perhitungan dari prediksi benar dari keselurahan jumlah data. Untuk menghitung akurasi menggunakan rumus berikut:

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{(True\ Positive + False\ Positive + True\ Negative + Flase\ Negative)} \quad (7)$$

b. *Precision*

Precision adalah rasio perhitungan dari suatu prediksi benar (positif maupun negatif) dibandingkan dengan jumlah hasil yang diprediksi. Rumusnya berikut ini:

$$Precision\ Positive = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \quad (8)$$

$$Precision\ Negative = \frac{True\ Negative}{(True\ Negative + False\ Negative)} \quad (9)$$

c. *Recall*

Recall merupakan perhitungan rasio dari suatu prediksi benar positif/negatif dibandingkan dengan jumlah keseluruhan data yang benar positif/negatif. *Recall positive* disebut juga sebagai *sensitivity*, sedangkan *recall negative* disebut *specificity*. Rumusnya sebagai berikut:

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

$$Sensitivity = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \quad (10)$$

$$Specificity = \frac{True\ Negative}{(True\ Negative + False\ Positive)} \quad (11)$$

d. *F1 Score*

F1 Score atau *F Measure* adalah hasil perhitungan dari nilai rata-rata perbandingan dari *precision* dan *recall*. Rumus dari *F1 Score* seperti berikut:

$$F1\ Score = \frac{2\ ((Recall)(Precision))}{Recall + Precision} \quad (12)$$

2.3 Integrasi Keilmuan

Untuk mendapat kejelasan mengenai tema skripsi terhadap pandangan Islam dengan mengintegrasikan ayat Al-Qur'an di dalamnya, penulis menggunakan beberapa ayat Al-Qur'an berikut yang didapat dari wawancara terhadap ustadz Rochman, beliau adalah ustadz lulusan Pondok Tebu Ireng Jombang dan juga sebagai ketua jama'ah yasin tahlil di Dukuh Kawal Kali Kendal RT 02. Pada saat wawancara membahas tentang pandangan dalam bertutur kata yang baik dan mengikuti perkembangan teknologi menurut perspektif Islam.

Allah SWT berfirman dalam Al-Qur'an surat Al-Baqarah ayat 83, di mana sebagai manusia sudah seharusnya bertutur kata yang baik sesamanya. Berikut surat Al-Baqarah ayat 83 yang berbunyi:

وَإِذْ أَخَذْنَا مِيثَاقَ بَنِي إِسْرَائِيلَ لَا تَعْبُدُونَ إِلَّا اللَّهَ وَبِالْوَالِدَيْنِ إِحْسَانًا وَذِي الْقُرْبَىٰ وَالْيَتَامَىٰ وَالْمَسَاكِينِ وَقُولُوا لِلنَّاسِ حُسْنًا وَأَقِيمُوا الصَّلَاةَ وَآتُوا الزَّكَاةَ ثُمَّ تَوَلَّيْتُمْ إِلَّا قَلِيلًا مِّنكُمْ وَأَنتُم مُّعْرِضُونَ (٨٣)

Artinya:"Dan ingatlah ketika Kami mengambil janji dari Bani Israil, 'Janganlah kamu menyembah selain Allah, dan berbuat baiklah kepada kedua orang tua, kerabat, anak-anak yatim dan orang-orang miskin. Dan bertutur katalah yang baik kepada manusia, tegakkanlah sholat dan tunaikan lah zakat.' Tetapi kemudian kamu berpaling (mengingkari), kecuali sebagian kecil dari kamu, dan kamu (masih menjadi) pembangkang"

Dalam Islam sendiri, Al-Qur'an tidak pernah mengekang umatnya untuk maju dan modern, justru Islam sendiri sangat mendukung kemajuan umatnya untuk

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

melakukan penelitian dan semacamnya termasuk juga dalam bidang teknologi. Seperti pada surat Ali-Imran ayat 190-191 berikut:

إِنَّ فِي خَلْقِ السَّمَاوَاتِ وَالْأَرْضِ وَاخْتِلَافِ اللَّيْلِ وَالنَّهَارِ لَآيَاتٍ لِّأُولِي الْأَلْبَابِ (١٩٠) الَّذِينَ يَذْكُرُونَ اللَّهَ قِيَامًا وَقُعُودًا وَعَلَىٰ جُنُوبِهِمْ وَيَتَفَكَّرُونَ فِي خَلْقِ السَّمَاوَاتِ وَالْأَرْضِ رَبَّنَا مَا خَلَقْتَ هَٰذَا بَاطِلًا سُبْحَانَكَ فَقِنَا عَذَابَ النَّارِ (١٩١)

Artinya: "Sesungguhnya dalam penciptaan langit dan bumi, dan pergantian malam dan siang terdapat tanda-tanda (kebesaran Allah) bagi orang yang berakal", "(yaitu) orang-orang yang mengingat Allah sambil berdiri, duduk, atau dalam keadaan berbaring, dan mereka memikirkan tentang penciptaan langit dan bumi (seraya berkata), 'Ya Tuhan kami, tidaklah Engkau menciptakan semua ini sia-sia, Maha Suci Engkau, lindungi lah kami dari azab neraka'."

Ayat Ali Imran di atas berisi tentang semua yang ada di langit dan bumi penuh dengan misteri yang berarti dapat kita cari tahu kebenarannya. Diharapkan dengan melakukan penelitian tersebut, kita dapat membantu dalam mencari kemudahan hidup baik di dunia maupun di akhirat dalam bidang apapun termasuk teknologi.v Jadi dari kedua surat di atas dapat kita simpulkan bahwa penelitian ini dengan ayat Al-Qur'an saling melengkapi di mana sebagai manusia harus bertutur kata yang baik sesamanya dalam mengikuti perkembangan jaman yang lebih baik. Sehingga penelitian ini dapat bermanfaat bagi seluruh umat manusia.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

BAB III
METODOLOGI PENELITIAN

3.1 Jenis dan Sumber Data

Data yang dipakai dalam proses penelitian ini merupakan *dataset* riwayat percakapan (*chat history*) permainan dari *cyberathlete* Dota2. Data didapatkan dari hasil *crawling* menggunakan *Application Programming Interface* (API) yang disediakan oleh pihak Dota2 yang dapat diakses pada situsnya di "https://www.opendota.com/api-keys". Proses *crawling* data dilakukan sebanyak dua kali tahapan di mana tahap pertama yakni, *crawling* data *cyberathlete* yang akan digunakan untuk *crawling* data *history chat* nya. Berikut ini pada Gambar 3.1 merupakan hasil dari tahap pertama yaitu *crawling* data *cyberathlete* dan pada Gambar 3.2 adalah tahap kedua dari pengambilan data yakni pengambilan *history chat* dari data *cyberathlete*.

account_id	steamid	avatar	avatarme	avatarful	profileur	personar	last_logi	full_histc	cheese
88470	76561197	https://st	https://st	https://st	https://st	Tzy		2020-03-0	0
1296625	76561197	https://st	https://st	https://st	https://st	Newbsha	2018-12-2	2020-03-2	0
3916428	76561197	https://st	https://st	https://st	https://st	gigablast		2020-03-1	0
3940262	76561197	https://st	https://st	https://st	https://st	March		2020-03-1	0
4281729	76561197	https://st	https://st	https://st	https://st	syndereN		2020-03-1	0
5150808	76561197	https://st	https://st	https://st	https://st	mindmin	2020-03-2	2020-03-0	0
6103402	76561197	https://st	https://st	https://st	https://st	3 digit ch	2019-01-0	2020-01-3	0

Gambar 3.1 *Crawling* Data *Cyberathlete*

_key	u	scold	him	aim	me	dog	thank	who	give	ramp	ment	break	stfu
my_word_counts	18	1	1	3	1	2	2	4	3	2	1	1	2
all_word_counts	597	4	56	17	302	114	41	54	43	13	4	2	27

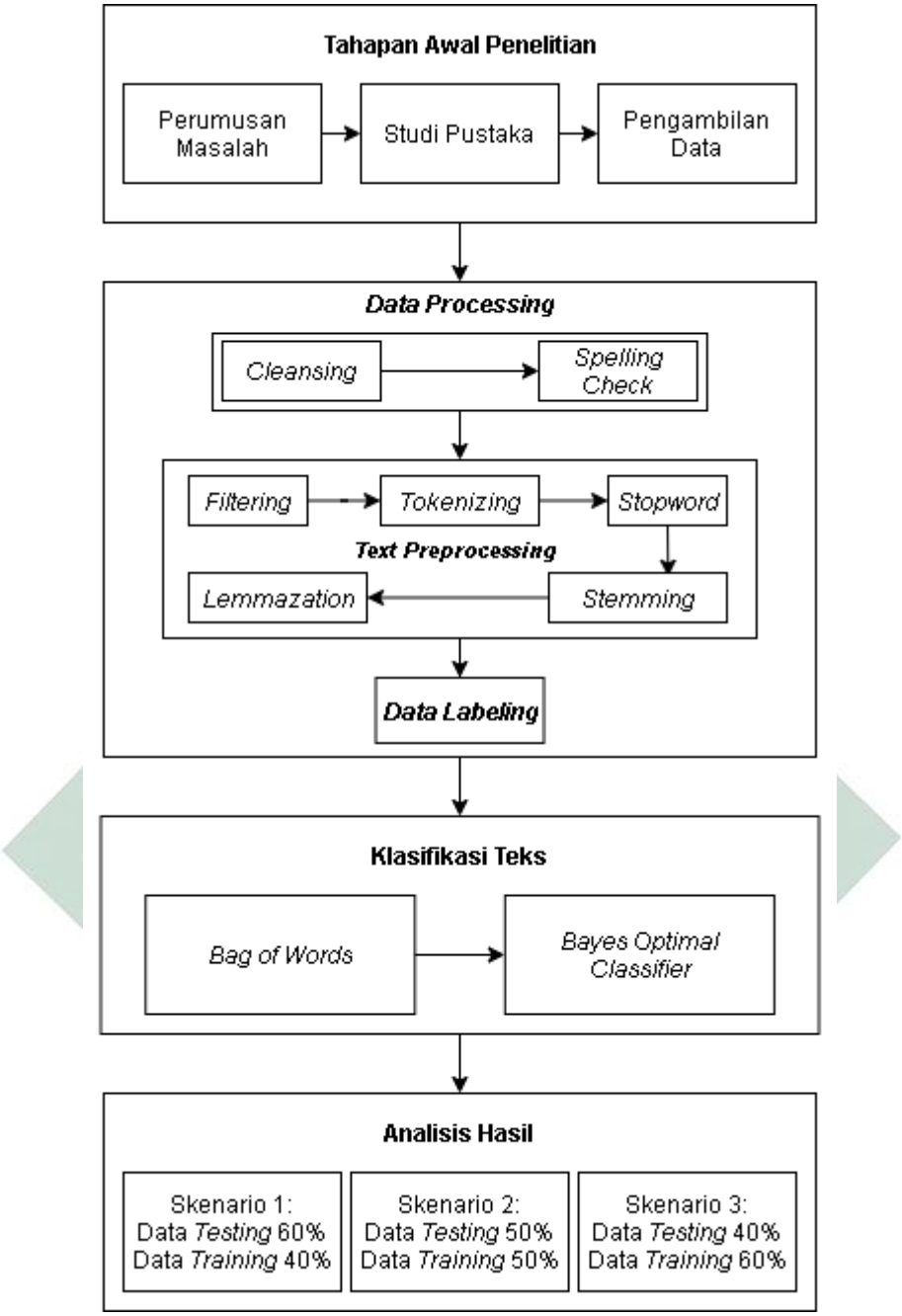
Gambar 3.2 *History Chat Cyberathlete*

Dari kedua proses pengambilan data tersebut, data kemudian diolah dengan mengambil variabel *_key* pada variabel *all_word_counts*.

3.2 Metodologi Penelitian

Alur penelitian ini memiliki empat tahapan utama yakni tahap awal penelitian, tahap pemrosesan data, tahap klasifikasi teks dan tahap analisis hasil. Tahapan tersebut dapat dilihat dalam diagram alur pada gambar berikut.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id



Gambar 3.3 Alur Penelitian Skripsi

Terdapat empat proses utama dalam penelitian ini, yaitu proses pada tahap awal, data *processing*, klasifikasi teks, dan analisis hasil. Berikut adalah penjelasan dari proses-proses tersebut.

3.2.1 Tahapan Awal

Pada tahapan awal atau tahap perumusan terbagi menjadi tiga proses yang akan dilakukan sebagai berikut.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

1. Perumusan Masalah

   Pada proses pertama yakni proses perumusan masalah yang merupakan proses yang melatar belakangi penelitian yang berjudul "KLASIFIKASI *CYBERBULLYING* BERBASIS *TEXT MINING* MENGGUNAKAN *BAYES OPTIMAL CLASSIFIER"*, sehingga ditemukannya rumusan masalah yang sudah tertera pada bab pertama. Masalah yang diangkat pada penelitian ini yaitu mendeteksi *cyberbullying* pemain profesional. Dengan menggunakan metode *bayes optimal classifier* dan berbasis *text mining* dapat mendeteksi *cyberbullying* yang dilakukan oleh seseorang.

2. Studi Pustaka

   Tahap selanjutnya yaitu proses literatur untuk menganalisis kebutuhan apa saja yang dibutuhkan untuk membangun mengimplementasikan suatu algoritma ke dalam sebuah sistem, dan teori-teori yang akan digunakan. Didapatkan hasil untuk pengembangan sistem ini menggunakan bahasa pemrograman python.

3. Pengumpulan Data

   Sebelum melakukan penelitian, terlebih dahulu mengumpulkan data yang akan diolah pada penelitian. Pengumpulan data ini dilakukan dengan cara *crawling* dengan API Dota2 yang telah ada. Penggunaan API Dota2 ini menggunakan IDE Jupyter Lab.

3.2.2 Data *Processing*

Setelah proses pada tahapan pertama dilakukan, selanjutnya yaitu mengolah data yang telah didapatkan. Pada tahapan data *processing* ini dilakukan beberapa proses seperti *cleansing* dan *spelling check* yang akan dijelaskan berikut ini.

1. *Cleansing*

   Proses pertama dalam pengolahan data yaitu pembersihan data. Proses ini dilakukan dengan menghilangkan semua simbol-simbol dan tanda baca seperti koma, titik, garing, tanda seru, at, hashtag, dollar, underscore, dan lain-lain.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

2. *Spelling Check*

Setelah data dilakukan pembersihan, data selanjutnya diproses untuk membenahi kata-kata yang tertulis *typo* atau salah penulisannya sebagaimana diperlihatkan pada Tabel 3.1 berikut.

Tabel 3.1 Contoh Kata Dengan Penulisan Salah (*Typo*)

Text	***Typo***
Arguing	Agruin
Amazing	Amazeing
Blessing	Blessign
Commend	Comnemd

3. *Text Preprocessing*

Pada tahap ini terdapat lima sub-proses yang merupakan kunci dalam pemrosesan teks. Lima sub proses ini dijelaskan sebagai berikut:

a. *Filtering*

Filtering digunakan dalam menghilangkan kata-kata dalam kalimat yang dianggap tidak memiliki makna seperti kata penulisan website dan kata yang berulang. Contohnya seperti wwwyoutube, twitchtv, httpsfacebook, wkwkjajahahalala, asdfkajsjak dan sejenisnya.

b. *Tokenizing*

Selanjutnya kalimat-kalimat tersebut dilakukan proses *tokenizing* yang berarti merubah kalimat dan teks menjadi sebuah token-token.

c. *Stopword Removal*

Setelah teks berubah menjadi *token*, selanjutnya yaitu menghilangkan kata yang tidak memiliki makna seperti *i*, *me*, *mine*, *can*, *could*, *would*, dan lain-lain. *Stoplist* yang digunakan telah disediakan oleh nltk, namun *stoplist* yang disediakan harus ditambahkan daftar kata tambahan secara manual.

d. *Stemming* dan *Lemmatization*

Proses *stemming* dan *lemmatization* ini merupakan proses merubah kata menjadi ke bentuk *lemma* dan ke bentuk dasarnya. Proses s*temming* sendiri yaitu menghilangkan *suffix* dan *prefix* pada suatu teks. Di mana proses ini mengambil kata dasar dari teks tersebut, contohnya *amazing*

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

menjadi *amaz*, dengan tambahan dari proses *lemmatization amaz* diubah ke kata yang seharusnya yaitu *amaze*. Algoritma *stemming* yang digunakan pada proses ini yaitu algoritma PorterStemmer, sedangkan untuk lemmatizationnya digunakan algoritma dari WordNetLemmatizer.

4. *Data Labeling*

Proses selanjutnya yaitu proses *labeling*, proses ini dilakukan untuk melabelkan kalimat dalam *dataset* yang menjadikan teks atau kalimat tersebut terlabelkan yang menunjukan bahwa kalimat tersebut termasuk dalam *cyberbullying* atau *non-cyberbullying* dengan menggunakan daftar kata-kata *positive* dan *negative* yang telah dimodifikasi disesuaikan dengan tema dari *dataset*. Menurut (Del Bosque & Garza, 2018) dalam penelitiannya, proses *labeling* ini dapat menggunakan acuan berupa daftar kata *aggresive words* dan proses manual pengisian dari daftar kata tersebut. Sehingga proses ini dapat dinamakan sebagai *rulebase classification.* Contoh dari hasil proses *labeling* seperti pada Gambar 3.4 berikut.

P-Score	N-Score	Text	Label
0,0390	0,0338	more points poor v	Non-Cyberbullying
0,0264	0,0287	en play friend will	Cyberbullying
0,0238	0,0268	whered go can sor	Cyberbullying
0,0444	0,0376	ssue paused and sr	Non-Cyberbullying
0,0250	0,0306	warlord and youre	Cyberbullying
0,0178	0,0227	chela est funkciya	Cyberbullying
0,0201	0,0267	hmdare deja de ju	Cyberbullying

Gambar 3.4 Contoh Data Hasil *Labeling*

Gambar 3.4 di atas menunjukan contoh data hasil *labeling*, pada gambar tersebut untuk melabelkan data menggunakan perhitungan sederhana di mana rumus untuk menghitung P-Score dan N-Score tersebut adalah sebagai berikut.

$$\frac{banyak\ kata\ positif\ atau\ negatif}{banyak\ kata\ dalam\ kalimat} \tag{13}$$

Setelah proses perhitungan selesai, jika nilai dari P-Score lebih tinggi di bandingkan dengan nilai N-Score maka data tersebut dilabelkan dengan label *Non-Cyberbullying*, sebaliknya jika nilai N-Score lebih besar daripada nilai P-Score maka data dilabelkan dengan label *Cyberbullying*.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

3.2.3 Klasifikasi Teks

Tahap selanjutnya setelah pengolahan data yaitu tahap klasifikasi teks. Pada tahap ini terbagi menjadi dua proses yakni:

1. *Bag of Words*

   Proses menjadikan sekumpulan teks menjadi satu wadah atau kantung (*bag*). Proses ini dilakukan untuk mempermudah dalam melakukan proses penerapan metode.

2. *Bayes Optimal Classifier*

   Pada proses ini dilakukan langkah-langkah bagaimana metode BOC dapat mengklasifikasikan *dataset* sehingga dapat mengetahui cara kerja dari metode BOC.

3.2.4 Analisis Hasil

Tahap ini merupakan proses pengujian metode dan menganalisis data yang telah diklasifikasi menggunakan metode *Bayes Optimal Classifier*. Pada tahap ini dilakukan perbandingan performa dari kedua metode, yaitu metode *Bernoulli Naive Bayes* dengan metode *Bayes Optimal Classifier*. Untuk mendapatkan hasil yang sesuai, dilakukan pembuatan skenario proporsi *dataset*. Skenario ini terbagi menjadi tiga, di mana skenario tersebut seperti berikut:

1. Data *training* 40% dan data *testing* 60%
2. Data *training* 50% dan data *testing* 50%
3. Data *training* 60% dan data *testing* 40%

Dari ketiga skenario tersebut mendapatkan hasil yang ditampilkan berupa *confusion matrix*, dari nilai *confusion matrix* ini di hitunglah nilai *accuracy*, *precision* (positif dan negatif), *recall* (*sensitivity* dan *specificity*), dan *F1 score*. Dari hasil tersebut, dapat dianalisis bahwa hasil dari kedua metode tersebut apakah memiliki perbedaan hasil akurasi yang berbeda ataupun tidak berbeda.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

BAB IV
HASIL DAN PEMBAHASAN

Bab ini menampilkan pembahasan dari hasil proses dalam mendapatkan dan mengolah data yang telah didapatkan dan proses klasifikasi dari proses awal hingga proses akhir penelitian.

4.1 Pengambilan Data

Proses pengambilan data dilaksanakan sebanyak dua kali proses dengan cara *crawling* data menggunakan API Dota2. Proses pertama didapatkan data yang telah di ambil dari fungsi di API Dota2 yakni "GET /proPlayers" yang menghasilkan 21 variabel. Fungsi ini digunakan untuk mengambil data *cyberathlete* Dota2, penelitian ini menggunakan 300 data *cyberathlete*. Dari 21 variabel ini, diantaranya: *account_id, steamid, avatar, avatarmedium, avatarfull, profileurl, personaname, last_login, full_history_time, cheese, fh_unvailable, loccountrycode, name, country_code, fantasy_role, team_id, team_name, team_tag, is_locked, is_pro, locked_until* hanya diambil satu variabel yang akan digunakan untuk proses selanjutnya yakni *account_id*. Berikut ini gambar yang dihasilkan dari kedua proses yang dilakukan.


```
[
  - {
        "account_id": 0,
        "steamid": "string",
        "avatar": "string",
        "avatarmedium": "string",
        "avatarfull": "string",
        "profileurl": "string",
        "personaname": "string",
        "last_login": null,
        "full_history_time": null,
        "cheese": 0,
        "fh_unavailable": true,
        "loccountrycode": "string",
        "name": "string",
        "country_code": "string",
        "fantasy_role": 0,
        "team_id": 0,
        "team_name": "string",
        "team_tag": "string",
        "is_locked": true,
        "is_pro": true,
        "locked_until": 0
    }
]
```


Gambar 4.1 *Crawling* Data *Cyberathlete* Dota2

Gambar di atas merupakan 21 variabel yang didapat dari proses pertama pengambilan data *cyberathlete*.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

account_id	steamid	avatar	avatarme	avatarful	profileur	personar	last_logi	full_histo	cheese
88470	76561197	https://st	https://st	https://st	https://st	Tzy		2020-03-0	0
1296625	76561197	https://st	https://st	https://st	https://st	Newbsha	2018-12-2	2020-03-2	0
3916428	76561197	https://st	https://st	https://st	https://st	gigablast		2020-03-1	0
3940262	76561197	https://st	https://st	https://st	https://st	March		2020-03-1	0
4281729	76561197	https://st	https://st	https://st	https://st	syndereN		2020-03-1	0
5150808	76561197	https://st	https://st	https://st	https://st	mindmin	2020-03-2	2020-03-0	0
6103402	76561197	https://st	https://st	https://st	https://st	3 digit ch	2019-01-0	2020-01-3	0

Gambar 4.2 Data *Cyberathlete* Dota2

Pada Gambar 4.2 adalah hasil dari proses pertama dalam pengambilan data. Digunakan sebanyak 300 data *cyberathlete*, dengan 21 variabel tiap satu data *cyberathlete*. Data yang didapat dalam bentuk *file JavaScript Oriented Number* (*JSON*). Selanjutnya, dari data tersebut diambil hanya variabel *account_id*, sehingga seperti pada Gambar 4.3 berikut.

account_id
3916428
6103402
6199147
6922000
8591574

Gambar 4.3 Data Variabel account_id

Gambar 4.3 di atas merupakan hasil pengolahan data pada proses awal pengambilan data. Data tersebut sudah dijadikan dalam bentuk file excel untuk mempermudah proses selanjutnya. Proses berikutnya adalah mengambil *chat history* dari 300 data *cyberathlete*, dengan menggunakan fungsi API Dota2 “GET /players/{account_id}/wordcloud”. Dari proses ini didapatkan dua variabel yakni variabel *my_word_counts* dan *all_word_counts* seperti yang terlihat pada gambar berikut.

```
{
    "my_word_counts": { },
    "all_word_counts": { }
}
```

Gambar 4.4 *Crawling* Data *History Chat*

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Gambar 4.4 di atas adalah hasil *crawling* data *history chat*. Variabel *my_word_counts* merupakan sekumpulan kata yang digunakan oleh pemain di dalam pertandingan, sedangkan *all_word_counts* adalah sekumpulan data yang digunakan di dalam *game* Dota2 baik di dalam *lobby*, maupun di dalam permainan. Berikut merupakan gambar dari hasil *crawling* data pada Gambar 4.5.


```
{
  "all_word_counts": {
    "the": 1968,
    "fuck": 625,
    "is": 2081,
    "going": 168,
    "on": 881,
    "up": 333,
    "here": 225,
    "can": 896,
    "we": 985,
    "go": 1096,
    "i": 3970,
    "think": 349,
    "hes": 537
```


Gambar 4.5 C*rawling* Data *History* C*hat*

Pada Gambar 4.5 di atas merupakan hasil dari *crawling* data *chat history* dari satu *cyberathlete*, yang masih dalam bentuk *file JSON*. Sehingga, dari hasil data di atas dirubah menjadi *file CSV* agar mempermudah proses selanjutnya. Berikut ini adalah gambar data yang telah di*convert* ke bentuk file excel yang tertera pada gambar berikut.

_key	the	fuck	is	going	on	up	here	can	we	go
all_word_counts	1968	625	2081	168	881	333	225	896	985	109

Gambar 4.6 *History* C*hat* Dalam Bentuk Excel

Gambar 4.6 di atas merupakan *chat history* dan telah di*convert* ke bentuk *file* excel yang menghasilkan satu variabel yakni variabel _key. Dari proses dalam pengambilan data, selanjutnya adalah memproses data tersebut sehingga membentuk kesatuan data yang dapat digunakan dalam proses berikutnya. Hasil dari proses penggabungan data dari kedua proses tersebut tertera pada Gambar 4.7 berikut.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

1	_key glgl glhf algeria algeria rdy gg wp ggwp lycka till grabar gl boys shoutou
2	_key sup bratans min sec thank gg garter fuck you mom famaly suck my dick
3	_key sven good min free farm disruptor rapes him gg game wd mid oic tusk
4	_key ya papi ewe rdy door sorry se orina oke confi brb sec zrro sorry gg lag g
5	_key the fuck going on up here can we go think hes done you guys are fuck
6	_key gg zavis fenix gege sec sorry lag really\ gl final hfhfhfhfhf rdy glhf miss
7	_key wtf bro nice swap report green please him sorry more pause and done
8	_key ts problems soryg gg lag sec sorry gl hf glhf binds glh ggf glgl min rdy cr
9	_key lag nvm afk super gg wp glhf ggwp best of luck feelas fellas big fan har
10	_key gg jakiro mvp doyou speak english dick go end not playing with this gr

Gambar 4.7 Hasil Pengambilan Data *History* C*hat Cyberathlete*

Pada Gambar 4.7 di atas merupakan gambar hasil dari proses menggabungkan 300 data *cyberathlete*, sehingga dapat mempermudah dalam proses berikutnya yakni proses data *processing*.

4.2 Data *Processing*

Proses selanjutnya adalah data *processing*, pada proses ini terbagi menjadi beberapa tahap dalam mengolah *dataset*. Proses-proses tersebut akan dijelaskan seperti berikut ini.

4.2.1 *Cleansing*

Proses *cleansing* dilakukan untuk menghilangkan tanda baca seperti koma, titik, tanda seru, *underscore*, dan sebagainya. Proses *cleansing* ditampilkan pada Tabel 4.1 berikut.

Tabel 4.1 Tahap *Cleansing*

Sebelum *Cleansing*	Sesudah *Cleansing*
_key the fuck is going on i am soery *lagging i* raelly\ *cant use my* sklli *got* prbolems *with my internet\ caring*	*key the fuck is going on i am* soery *lagging i* raelly *cant use my* sklli *got* prbolems *with my internet caring*

4.2.2 *Spelling Check*

Proses *spelling check* ini dilakukan untuk membenarkan penulisan kata yang salah (*typo*) seperti pada tabel berikut.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Tabel 4.2 *Spelling Check*

Sebelum	Sesudah
key the fuck is going on i am soery *lagging i* raelly *cant use my* sklli *got* prbolems *with my internet caring*	*key the fuck is going on i am sorry lagging i really cant use my skill got problems with my internet caring*

4.2.3 *Labeling*

Proses *labeling* ini dilakukan dengan menggunakan list kata positif dan negatif yang di unduh dari SentiWordNet. Namun dalam tema *dataset* kali ini yang menggunakan *chat history* dari suatu pemain Dota2 dalam mengklasifikasikan kata *cyberbullying* maka list kata positif dan negatif tersebut sedikit diubah, yakni dengan cara mengurangi dan menambahkan beberapa data yang termasuk dalam kata *cyberbullying*. Berikut beberapa kata yang ditambah dan dikurangi dalam *list* kata positif dan negatif pada tabel-tabel berikut.

Tabel 4.3 Perubahan *Positive Words*

Kata yang Dihilangkan	Kata yang Ditambahkan
accurate	*gege*
agile	*goodgame*
agility	*havefun*
cure	*goodluck*
divine	*wellplayed*

Kata yang dihilangkan merupakan kata-kata suatu *skill* atau kemampuan dari *character* yang ada pada *game* Dota, dan kata yang ditambahkan merupakan kata yang sering digunakan pemain dalam mengapresiasikan perasaan bermainnya.

Tabel 4.4 Perubahan *Negative Words*

Kata yang Dihilangkan	Kata yang Ditambahkan
agony	*fuck*
banish	*shithead*
cancer	*motherfucker*
demon	*cunt*
fear	*nigga*

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Kata yang dihilangkan pada *negative words* adalah kata-kata yang berhubungan dengan yang ada pada *game* tersebut seperti *skill*, ras, nama karakter. Dan kata yang ditambahkan adalah kata yang termasuk dalam kata rasis seperti "nigga" yang merupakan nigerian yang berhubungan pada suatu ras dan beberapa kata yang termasuk kata-kata kotor (*profanity*).

Setelah perubahan pada daftar kata positif dan negatif, maka dataset akan diproses dan diberikan label dengan cara mengecek dataset apakah isi dari dataset termasuk dalam kata yang ada pada daftar kata positif maupun kata negatif. Dengan menggunakan rumus sederhana dihasilkan data yang tertera pada gambar berikut.

P-Score	N-Score	Text	Label
0,0390	0,0338	key glgl glhf algeria a	Non-Cyberbullying
0,0264	0,0287	key sup bratans min	Cyberbullying
0,0238	0,0268	key sven good min fr	Cyberbullying
0,0444	0,0376	key ya papi ewe rdy	Non-Cyberbullying
0,0250	0,0306	key the fuck going or	Cyberbullying

Gambar 4.8 Proses *Labeling*

Seperti pada Gambar 4.8 di atas, jika P-Score lebih besar daripada N-Score maka dilabelkan *Non-Cyberbullying*, sebaliknya jika N-Score lebih besar daripada P-Score maka dilabelkan *Cyberbullying*. Setelah proses pelabelan selesai maka dataset siap diolah ke proses selanjutnya.

4.2.4 *Text Pre-processing*

Proses teks *processing* ini merupakan kunci dari analisis teks, di mana proses ini memiliki beberapa proses yang dijelaskan sebagai berikut.

1. *Filtering*

   Proses *filtering* dilakukan untuk menghilangkan nama *website* dan juga kata yang berulang seperti httpyoutube, httpsfacebook, twitchtv, youtubecom, jajaja, hahaha, xaxaxa, xdxdxxd, asfgasjkfhaskd, dan sebagainya. Proses ini juga merupakan salah satu proses dari teks *processing* yang harus dilakukan. Berikut pada Tabel 4.5 merupakan contoh dari proses *filtering.*

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Tabel 4.5 Contoh Proses *Filtering*

Sebelum	Sesudah
key come to my youtube channel this is the link <u>httpsyoutubecomaska</u> *im streaming right now*	*key come to my youtube channel this is the link im streaming right now*
key subscribe to my twitch chanel <u>twitchtvaska</u> *or my youtube channel* <u>youtubecomaska</u> <u>*gogogogo*</u> *you will feel lost if not subscribing*	*key subscribe to my twitch chanel or my youtube chanel you will feel lost if not subscribing*

2. *Tokenizing*

Proses selanjutnya adalah proses *tokenizing*, yang merupakan proses membuat suatu kata menjadi *token* atau unit kecil. Contoh kalimat pada proses *tokenizing* seperti pada tabel berikut.

Tabel 4.6 Contoh Proses *Tokenizing*

Sebelum	Sesudah
key the fuck is going on i am sorry lagging i really cant use my skill got problems with my internet caring	*['key','the','fuck','is','going','on','i' ,'am','sorry','lagging','i','really','cant' ,'use','my','skill','got','problems','with' ,'my','internet','caring']*
key you are so cool like a wolves maybe i will like those sometimes	*['key','you','are','so','cool','like' ,'a','wolves','maybe','i' ,'will','like','those','sometimes']*

3. *Stopword Removal*

Proses ini merupakan proses menghapus daftar kata yang terdapat pada *stoplist*. Di mana kata-kata ini termasuk dalam kata yang tidak memiliki makna, seperti *i*, *my*, *mine*, *am*, dan lain-lain. Proses ini juga yang menentukan kalimat atau teks apa saja yang layak akan digunakan dalam klasifikasi teks, berikut pada Tabel 4.7 merupakan contoh dari *Stopword removal*.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Tabel 4.7 Contoh Proses *Stopword Removal*

Sebelum	Sesudah
['key','the','fuck','is','going','on','i' ,'am','sorry','lagging','i','really' ,'use','my','skill','got','problems','with' ,'my','internet','caring']	*['fuck','going','sorry' ,'lagging','really','use' ,'skill','got','problems', 'internet','caring']*
['key','you','are','so','cool','like' ,'a','wolves','maybe','i' ,'will','like','those','sometimes']	*['cool','like','wolves' ,'maybe','like','sometimes']*

Stopword yang disediakan oleh NLTK tidak seberapa lengkap, oleh karena itu penambahan *stopword* disesuaikan dengan tema dari penelitian. Sehingga dalam penelitian ini ditambahkan beberapa kata yang tidak ada pada *stopword* seperti *key*, *cant*, *couldnt*, *wouldnt*, dan sebagainya.

4. *Stemming*

Setelah proses *stopword removal* telah selesai maka proses selanjutnya yaitu *stemming*. *Stemming* digunakan untuk menghilangkan *suffix* dan *prefix* suatu kata, atau dalam artian luas *stemming* digunakan untuk mengetahui kata dasar dari suatu kata. Berikut contoh dari *stemming* yang ditunjukan pada Tabel 4.8.

Tabel 4.8 Contoh Proses *Stemming*

Sebelum	Sesudah
['fuck','going','sorry','lagging' ,'really','use','skill','got' ,'problems','internet','caring']	*['fuck','go','sorry','lagg' ,'really','use','skill','got' ,'problem','internet','car']*
['cool','like','wolves','maybe' ,'like','sometimes']	*['cool','like','wolv','mayb' ,'like','sometim']*

5. *Lemmatization*

Proses ini biasanya dilakukan bersamaan dengan proses *stemming*. *Lemmatization* merupakan proses penyempurnaan kata hasil *stemming*. Contoh dari proses ini ditampilkan pada Tabel 4.8 berikut.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

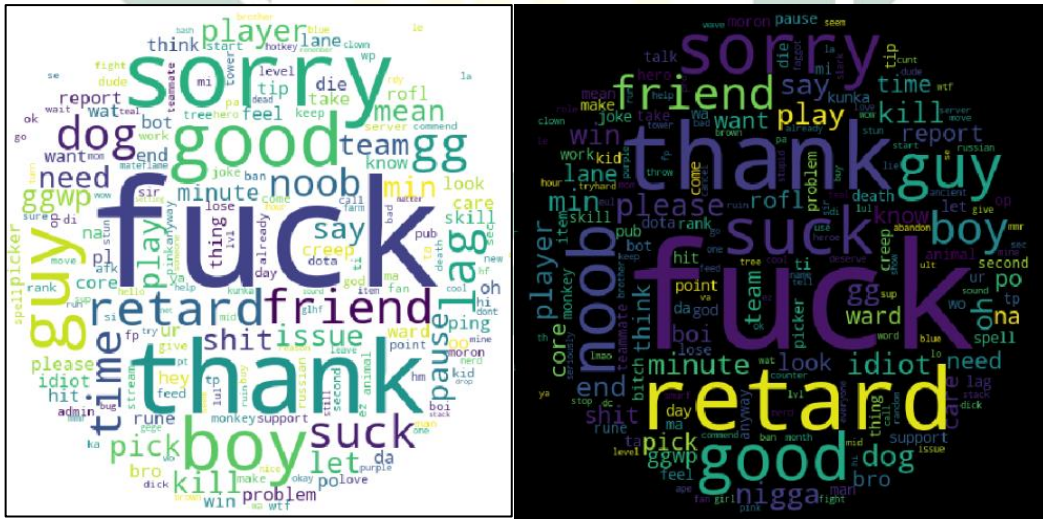
Tabel 4.9 Contoh Proses *Lemmatization*

Sebelum	Sesudah
['fuck','go','sorry','lagg' ,'really','use','skill','got' ,'problem','internet','car']	*['fuck','go','sorry','lag' ,'really','use','skill','got' ,'problem','internet','care']*
['cool','like','wolv','mayb' ,'like','sometim']	*['cool','like','wolf','maybe' ,'like','sometime']*

Setelah proses-proses tersebut telah dilakukan, maka *dataset* tersebut siap untuk diterapkan dengan algoritma atau metode klasifikasi teks *Bayes Optimal Classifier*.

4.3 Klasifikasi Teks

Hasil dari pengolahan data sebelum dimasukan ke model *bag of words*, data tersebut dipaparkan menjadi sebuah *wordcloud*. *Wordcloud* adalah sekumpulan kata yang ada pada dokumen, langkah ini bertujuan untuk memperlihatkan *unique words* pada *dataset*. Hasil dari *wordcloud* ditampilkan pada gambar berikut.


sorry
good
fuck
thank
retard
friend
boy
suck
guy
dog
noob
lag
shitissue
pause
time
player
kill
report
play
say
team
gg
mean
need
minute
min
please
idiot
pick
bro
problem
nigga
ward
second
spell
core
think
lane
feel
rofl
tip
time
kill
report
guy
boy
suck
thank
fuck
retard
good
friend
sorry
play
say
minute
look
idiot
need
dog
bro
nigga
pick
noob
min
please
lane
win
work
kid
core
player
think
point
hit
team
boi
lose
end
shit
day
ggwp
feel
support
issue
second
spell
po
ward
gg
ona
oh
death
know
let
give
creep
abandon
rofl
problem
come
joke
want
make
kunka
clown
russian
start
tip
cunt
die
say
time
wtf
server
move
stun
lag
u
ur
po
wo
tp
sound
blue
word
counter
ban
month
msd
issue
stop
gc
level
ma
ta
rune
bitch
anyway
comment
ape
girl
fan
pink
fight
mark
nigga
bro
support
thing
call
dick
picker
teammate
animal
monkey
lmao
lose
point
va
da god
boi
hit
tree
cool
item
fan
help
kunka
sound
cool
hi
dont
mine
moron
kid
drop
nerd
boi
black
ka
one
wtf
sc
pink
girl
fan
fight
ape
mark
blue
la
lag
dick
issue
support
thing
call

4.3.1 *Bag of Words*

Langkah ini bertujuan untuk mengumpulkan dataset menjadi satu ke sebuah kantung (*bag*). Kantung ini berisi matriks dari *unique words* atau berisi *term* dari proses pengolahan data. Proses ini dipaparkan seperti berikut:

Tabel 4.10 Dataset BoW

Data 1	**Data 2**	**Data 3**	**Data 4**
Nice	*Noob*	*Great*	*Wellplay*
Bro	*Come*	*Skill*	*Nice*
Great	*Low*	*Bro*	*Skill*
Skill	*Skill*	*Gege*	*Bro*
Suppose	*Retard*	-	*Gege*
Noob	*Cunt*	-	-

Tabel 4.10 di atas merupakan contoh data dari dataset, berisikan 4 data yang selanjutnya akan dihitung jumlah frekuensi dari *unique words dataset* di atas.

Tabel 4.11 Frekuensi kemunculan kata

Word	**Frequency**
Skill	4
Bro	3
Great	2
Nice	2
Gege	2
Noob	2
Come	1
Low	1
Retard	1
Cunt	1
Suppose	1
Wellplay	1

Dari Tabel 4.11 di atas, merupakan tabel dari frekuensi kemunculan kata dapat dilihat bahwa kata yang sering muncul adalah kata "*Skill*" sebanyak 4 kali,

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

kata "*Bro*" sebanyak 3 kali, dan kemunculan kata-kata lainnya. Selanjutnya adalah proses inti dari *BoW*, hasil dari proses tersebut tertera pada tabel berikut.

Tabel 4.12 *Bag of Words*

	Skill	*Bro*	*Great*	*Nice*	*Gege*	*Noob*	*Come*	*Low*	*Retard*	*Cunt*	*Suppose*	*Wellplay*
Data 1	1	1	1	1	0	1	0	0	0	0	1	0
Data 2	1	0	0	0	0	1	1	1	1	1	0	0
Data 3	1	1	1	0	1	0	0	0	0	0	0	0
Data 4	1	1	0	1	1	0	0	0	0	0	0	1

Tabel 4.12 merupakan hasil dari proses *BoW*, dari proses ini dibentuk menjadi sebuah matriks yang berisi nilai dari kemunculan kata dari tiap data. Untuk memahami matriks di atas pada Data 1, kata "S*kill*" muncul sebanyak 1 kali, dan pada kata "G*ege*" karena pada Data 1 tidak muncul maka di isi dengan nilai 0. Selanjutnya pada Data 2, kata "S*kill*" muncul sebanyak sekali, sedangkan kata "*Bro*" tidak muncul maka di isi nilai 0. Dengan cara tersebut nilai dari matriks di isi dengan nilai 1 atau 0 tergantung dari jumlah kata yang muncul pada setiap data.

4.3.2 *Bayes Optimal Classifier*

Proses selanjutnya yaitu proses klasifikasi dengan menggunakan metode *BOC* dengan cara menghitung probabilitas label dari *cyberbullying*. Dengan menggunakan *dataset* yang sudah layak untuk dihitung probabilitasnya (sudah melewati tahap pemrosesan data). Pada Tabel 4.10 berikut data siap untuk melanjutkan ke proses klasifikasi menggunakan metode *BOC*.

Tabel 4.13 Contoh *Dataset* Siap Untuk Proses Klasifikasi

No	**Teks**	**Label**
N1	*nice* | *bro* | *great* | *skill* | *suppos* | *noob*	*Non-cyberbullying*
N2	*noob* | *come* | *low* | *skill* | *retard* | *follow* | *mid*	*Cyberbullying*
N3	*great* | *skill* | *wellplay* | *although* | *toxic*	*Non-cyberbullying*

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

N4	*toxic* | *low* | *skill* | *anyth* | *go* | *home* | *bitch*	*Cyberbullying*
N5	*big* | *shoot* | *great* | *skill* | *gege* | *wellplay*	*Non-cyberbullying*
N6	*fuck* | *feed* | *retard*	*Cyberbullying*
N7	*amaze* | *annihil* | *gege* | *wellplay*	*Non-cyberbullying*
N8	*goodluck* | *havefun* | *bro*	*Non-cyberbullying*
N9	*great* | *bulleye* | *high* | *skill* | *bro*	?

Melihat Tabel 4.13 di atas dapat ditarik kesimpulan bahwa teks yang berlabelkan *Non-cyberbullying* terdapat 16 kali kata yang muncul, sedangkan pada data berlabelkan *Cyberbullying* terdapat 14 kali kata yang muncul sehingga jumlah keseluruhan kemunculan kata sebanyak 30 kata. Jika dipetakan maka terlihat seperti pada tabel berikut.

Tabel 4.14 Pemetaan Kemunculan Kata

Non-cyberbullying	nice (2), bro (2), great (3), skill (3), suppos (1), noob (1), wellplay (3), although (1), toxic (1), big (1), shoot (1), gege (2), amaze (1), annihil (1), goodluck (1), havefun (1)
Cyberbullying	noob (1), come (1), low (2), skill (2), retard (2), follow (1), mid (1), toxic (1), anyth (1), go (1), home (1), bitch (1), fuck (1), feed (1)

Setelah mengetahui seberapa banyak kemunculan kata seperti pada Tabel 4.14 di atas, langkah selanjutnya adalah menghitung probabilitas dengan rumus perhitungan probabilitas, di mana pada rumus ini harus menentukan probabilitas nilai $P(vj \mid h_i)$ dan probabilitas $P(h_i|D)$. Dari Tabel 4.10 diketahui bahwa terdapat 8 teks data sehingga nilai $P(vj \mid h_i)$ adalah.

$P(Non|h_i) \quad = {}^{5}/_{8}$

$P(Bully|h_i) \quad = {}^{3}/_{8}$

Dan untuk nilai dari $P(h_i|D)$ dihitung menggunakan rumus probabilitas kata sehingga seperti berikut.

$P(skill|D) \quad = \frac{3+1}{16+30} = {}^{4}/_{46}$

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

$P(skill|D)$ $= \frac{2+1}{14+30} = {}^{3}/_{44}$

Dengan menggunakan perhitungan seperti di atas maka nilai dari $P(h_i|D)$ ditampilkan pada tabel berikut.

Tabel 4.15 Nilai Probabilitas *Term*

Label	$P(vj\|h_i)$	$P(h_i\|D)$								
		skill	great	wellplayed	bro	noob	low	retard	toxic	gege
Non	${}^{5}/_{8}$	4/46	4/46	4/46	3/46	2/46	1/46	1/46	2/46	3/46
Bully	${}^{3}/_{8}$	3/44	1/44	1/44	1/44	2/44	3/44	3/44	2/44	1/44

Label	$P(vj\|h_i)$	$P(h_i\|D)$								
		nice	suppos	come	follow	mid	althou	anyth	go	home
Non	${}^{5}/_{8}$	2/46	2/46	1/46	1/46	1/46	2/46	1/46	1/46	1/46
Bully	${}^{3}/_{8}$	1/44	1/44	2/44	2/44	2/44	1/44	2/44	2/44	2/44

Label	$P(vj\|h_i)$	$P(h_i\|D)$								
		bitch	big	shoot	fuck	feed	amaze	annihil	goodluck	havefun
Non	${}^{5}/_{8}$	1/46	2/46	2/46	1/46	1/46	2/46	2/46	2/46	2/46
Bully	${}^{3}/_{8}$	2/44	1/44	1/44	2/44	2/44	1/44	1/44	1/44	1/44

Dengan nilai dari $P(vj|h_i)$ dan $P(h_i|D)$ didapatkan maka selanjutnya yaitu menghitung dengan menggunakan Rumus 4 sehingga hasil akan tampak seperti berikut.

$P(Non)$ $=P(great|D) * P(Non|h_i) + P(bulleye|D) * P(Non|h_i) + P(high|D) * P(Non|h_i) + P(skill|D) * P(Non|h_i) + P(bro|D) * P(Non|h_i)$

$P(Non)$ = 4/46 * 5/8 + 1/46 *5/8 + 1/46 * 5/8 + 4/46 * 5/8 + 3/46 * 5/8
= 0,17663

$P(Bully)$ $=P(great|D) * P(Bully|h_i) + P(bulleye|D) * P(Bully|h_i) + P(high|D) * P(Bully|h_i) + P(skill|D) * P(Bully|h_i) + P(bro|D) * P(Bully|h_i)$

$P(Bully)$ = 1/44 * 3/8 + 1/44 * 3/8 + 1/44 * 3/8 + 3/44 * 3/8 + 1/44 * 3/8
= 0,059659

Dari perhitungan menggunakan metode BOC di atas, didapatkan nilai $P(Non)$ lebih besar daripada nilai $P(Bully)$ sehingga teks N9 merupakan teks yang berlabel Non-cyberbullying.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

4.4 Analisis Hasil

Pada tahap analisis hasil disiapkan skenario proporsi jumlah data *test* dan data *train*, dengan itu dilakukan proses melatih data dengan menggunakan metode pembanding lain yakni metode *Bernoulli Naive Bayes*. Sehingga dapat ditarik kesimpulan apakah metode *Bayes Optimal Classifier* memiliki perbandingan hasil yang cukup berbeda dengan metode pembandingnya. Oleh karena itu pada tahap ini di siapkan 3 skenario yakni:

1. Data *training* 40% dan data *testing* 60%
2. Data *training* 50% dan data *testing* 50%
3. Data *training* 60% dan data *testing* 40%

4.4.1 Skenario pertama data *train* 40% dan data *test* 60%

Pada skenario pertama menggunakan data sebanyak 300 data di mana data *test* sebesar 60% atau sejumlah 180 data dan data *train* sebesar 40% atau sejumlah 120 data. Dengan menggunakan metode *Bayes Optimal Classifier* dihasilkan data yang tertera pada tabel *confusion matrix* berikut.

Tabel 4.16 Hasil *Confusion Matrix* Skenario Pertama Metode BOC

		Nilai Sebenarnya	
		Positif	Negatif
Nilai Prediksi	Positif	86	0
	Negatif	27	7

Selanjutnya dengan melihat hasil dari tabel *confusion matrix* di atas dapat dihitung nilai dari *accuracy*, *precision* (positif dan negatif), *recall* (*sensitivity* dan *specificity*), dan *F1 score*.

Accuracy $= \frac{(86+7)}{(86+0+27+7)}$ x 100%

= 77,5%

Precision Positive $= \frac{86}{(86+0)}$ x 100%

= 100,00%

Precision Negative $= \frac{7}{(7+27)}$ x 100%

= 20,59%

Sensitivity $= \frac{86}{(86+27)}$ x 100%

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

= 76,11%

Specificity $= \frac{7}{(27+0)}$ x 100%

= 25,93%

F1 Score $= \frac{2((76,11)(100))}{(76,11+100)}$

= 86,43%

Dengan menggunakan proporsi data yang sama tetapi menggunakan metode *Bernoulli Naive Bayes*, dihasilkan data dari proses perhitungan yang tertera pada tabel *confusion matrix* berikut.

Tabel 4.17 Hasil *Confusion Matrix* Skenario Pertama Metode BNB

		Nilai Sebenarnya	
		Positif	Negatif
Nilai Prediksi	Positif	83	3
	Negatif	33	1

Kemudian dari nilai yang tertera pada Tabel 4.17 di atas, dihitunglah nilai dari *accuracy*, *precision* (positif dan negatif), *recall* (*sensitivity* dan *specificity*), dan *F1 score* seperti berikut cara perhitungannya.

Accuracy $= \frac{(83+1)}{(83+3+33+1)}$ x 100%

= 70%

Precision Positive $= \frac{83}{(83+3)}$ x 100%

= 96,51%

Precision Negative $= \frac{1}{(1+33)}$ x 100%

= 2,94%

Sensitivity $= \frac{83}{(83+33)}$ x 100%

= 71,55%

Specificity $= \frac{1}{(33+3)}$ x 100%

= 2,78%

F1 Score $= \frac{2((71,55)(96,51))}{(71,55+96,51)}$

= 83%

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Dari kedua hasil nilai *accuracy*, *precision*, *recall* dan *F1 score* menggunakan metode *Bayes Optimal Classifier* dan metode *Bernoulli Naive Bayes*, hasil tersebut dipaparkan pada tabel berikut.

Tabel 4.18 Perbandingan Hasil Metode BOC dan BNB Skenario Pertama

Metode	*Accuracy*	*Precision Positive*	*Precision Negative*	*Sensitivity*	*Specificity*	*F1 Score*
BOC	77,5%	100%	20,59%	76,11%	25,93%	86,43%
BNB	70%	98,81%	2,94%	71,55%	2,78%	83%

4.4.2 Skenario kedua data *train* 50% dan data *test* 50%

Selanjutnya pada skenario kedua, data yang digunakan sebesar 300 data. *Dataset* dibagi sama rata 1:1 yang berarti 50% data train dan 50% data test yang berarti data *test* sejumlah 150 data dan data *train* sejumlah 150 data. Data kemudian diolah menggunakan metode *Bayes Optimal Classifier* untuk memprediksikan hasil dari klasifikasi teks. Selanjutnya agar lebih mudah untuk ditampilkan, data dipaparkan dalam bentuk *confusion matrix*.

Tabel 4.19 Hasil *Confusion Matrix* Skenario Kedua Metode BOC

		Nilai Sebenarnya	
		Positif	Negatif
Nilai Prediksi	Positif	108	0
	Negatif	31	11

Dari hasil perhitungan menggunakan algoritma BOC mendapat hasil seperti pada Tabel 4.19 di atas. Langkah selanjutnya yaitu menghitung nilai *accuracy*, *precision* (positif dan negatif), *recall* (*sensitivity* dan *specificity*), dan *F1 score*.

Accuracy $= \frac{(108+10)}{(108+0+31+11)}$ x 100%

= 79,33%

Precision Positive $= \frac{108}{(108+0)}$ x 100%

= 100,00%

Precision Negative $= \frac{11}{(11+31)}$ x 100%

= 26,19%

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Sensitivity $= \frac{108}{(108+31)}$ x 100%

= 77,7%

Specificity $= \frac{11}{(31+0)}$ x 100%

= 35,49%

F1 Score $= \frac{2((77,7)(100))}{(77,7+100)}$

= 87,45%

Masih dengan proporsi data yang sama, dihitung proses klasifikasi dengan menggunakan metode BNB. Hasil yang didapatkan dari proses tersebut dipaparkan dalam tabel berikut.

Tabel 4.20 Hasil *Confusion Matrix* Skenario Kedua Metode BNB

		Nilai Sebenarnya	
		Positif	Negatif
Nilai Prediksi	Positif	108	0
	Negatif	42	0

Dari Tabel 4.19 di atas adalah hasil dari proses perhitungan yang ditampilkan pada *confusion matrix* metode BOC. Selanjutnya adalah proses menghitung nilai dari *accuracy, precision(positive and negative), recall(sensitivity and specificity),* dan *F1 score.*

Accuracy $= \frac{(108+0)}{(108+0+42+0)}$ x 100%

= 72%

Precision Positive $= \frac{108}{(108+0)}$ x 100%

= 100%

Precision Negative $= \frac{0}{(0+42)}$ x 100%

= 0%

Sensitivity $= \frac{108}{(108+42)}$ x 100%

= 72%

Specificity $= \frac{0}{(42+0)}$ x 100%

= 0%

F1 Score $= \frac{2((74,15)(99,09))}{(74,15+99,09)}$ = 83,72%

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Dari kedua hasil nilai *accuracy*, *precision*, dan *recall* menggunakan metode *Bayes Optimal Classifier* dan metode *Bernoulli Naive Bayes*, data yang diperoleh kemudian dipaparkan seperti pada Tabel 4.21 berikut.

Tabel 4.21 Perbandingan Hasil Metode BOC dan BNB Skenario Kedua

Metode	*Accuracy*	*Precision Positive*	*Precision Negative*	*Sensitivity*	*Specificity*	*F1 Score*
BOC	79,33%	100%	26,19%	77,7%	35,49%	87,45%
BNB	72%	100%	0%	72%	0%	83,72%

4.4.3 Skenario ketiga data *train* 60% dan data *test* 40%

Skenario terakhir dengan menggunakan 300 jumlah data, data dibagi sehingga data tersebut menjadi 60% data train, dan 40% data test berikut prosesnya.

Tabel 4.22 Hasil *Confusion Matrix* Skenario Ketiga Metode BOC

		Nilai Sebenarnya	
		Positif	Negatif
Nilai Prediksi	Positif	132	0
	Negatif	34	14

Tabel 4.22 di atas, didapatkan nilai TP, TN, FP, dan FN maka selanjutnya adalah menghitung nilai *accuracy*, *precision(positive and negative)*, *recall(sensitivity and specificity)*, dan *F1 score*.

Accuracy $= \frac{(132+9)}{(132+0+39+9)}$ x 100%

= 78,33%

Precision Positive $= \frac{132}{(132+0)}$ x 100%

= 100%

Precision Negative $= \frac{9}{(9+39)}$ x 100%

= 18,75%

Sensitivity $= \frac{132}{(132+39)}$ x 100%

= 77,19%

Specificity $= \frac{9}{(39+0)}$ x 100%

= 23,08 %

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

F1 Score $= \frac{2((77,19)(100))}{(77,19+100)}$

$= 87,13\%$

Masih dengan menggunakan proporsi data yang sama tetapi menggunakan metode *Bernoulli Naive Bayes*, dihasilkan data dari proses perhitungan yang tertera pada tabel *confusion matrix* berikut.

Tabel 4.23 Hasil *Confusion Matrix* Skenario Ketiga Metode BNB

		Nilai Sebenarnya	
		Positif	Negatif
Nilai Prediksi	Positif	132	0
	Negatif	48	0

Pada proses tersebut menghasilkan nilai dari TP, FP, TN, dan FN. Kemudian data tersebut dilakukan proses penghitungan nilai *accuracy, precision(positive and negative), recall(sensitivity and specificity),* dan *F1 score*.

Accuracy $= \frac{(131+2)}{(131+1+46+2)}$ x 100%

$= 73,89\%$

Precision Positive $= \frac{131}{(131+1)}$ x 100%

$= 99,24\%$

Precision Negative $= \frac{2}{(2+46)}$ x 100%

$= 4,17\%$

Sensitivity $= \frac{131}{(131+46)}$ x 100%

$= 74,01\%$

Specificity $= \frac{2}{(46+1)}$ x 100%

$= 4,26\%$

F1 Score $= \frac{2((74,01)(99,24))}{(74,01+99,24)}$

$= 84,79\%$

Dari kedua hasil nilai *accuracy, precision(positive and negative), recall(sensitivity and specificity),* dan *F1 score* menggunakan metode *Bayes Optimal Classifier* dan metode *Bernoulli Naive Bayes*, data yang diperoleh kemudian dipaparkan seperti pada tabel berikut.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Tabel 4.24 Perbandingan hasil Metode BOC dan BNB Skenario Ketiga

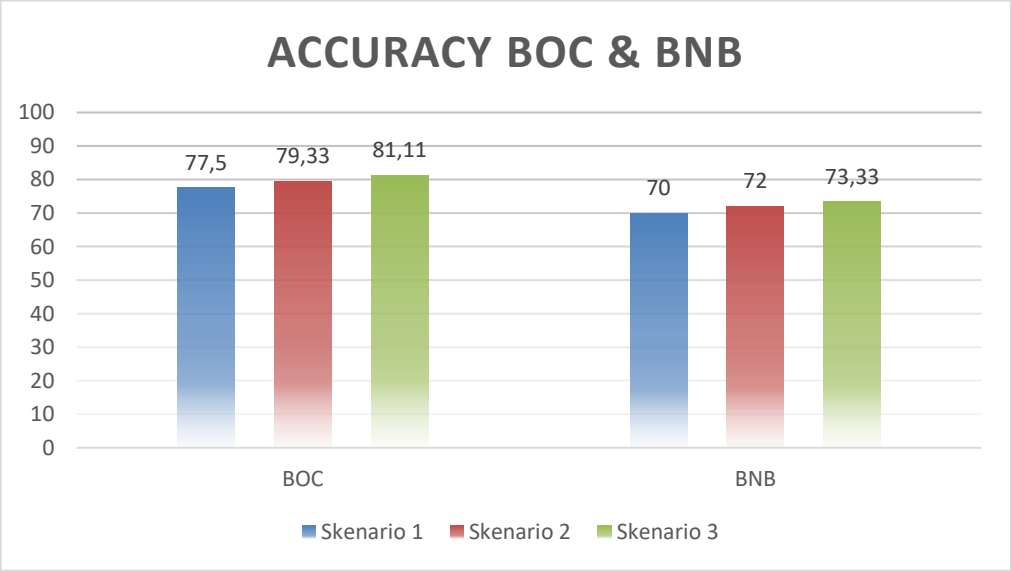
Metode	*Accuracy*	*Precision Positive*	*Precision Negative*	*Sensitivity*	*Specificity*	*F1 Score*
BOC	81,11%	100%	29,17%	79,52%	41,18%	88,59%
BNB	73,33%	100%	0%	73,33%	0%	84,61%

4.4.4 Hasil Klasifikasi

Berdasarkan hasil perhitungan melalui proses klasifikasi menggunakan metode BOC dan BNB sebanyak 300 data *sample*, maka dapat disimpulkan bahwa *accuracy*, *precision(positive and negative)*, *recall(sensitivity and specificity)*, dan *F1 score* pada ketiga skenario dapat dilihat pada tabel perbandingan berikut.

Tabel 4.25 Hasil dari Ketiga Skenario

Nilai	Skenario ke-1 (40% *Train* & 60% *Test*)		Skenario ke-2 (50% *Train* & 50% *Test*)		Skenario ke-3 (60% *Train* & 40% *Test*)	
	BOC	BNB	BOC	BNB	BOC	BNB
Accuracy	77,5%	70%	79,33%	72%	81,11%	73,33%
Precision Positive	100%	98,81%	100%	100%	100%	100%
Precision Negative	20,59%	2,94%	26,19%	0%	29,17%	0%
Sensitivity	76,11%	71,55%	77,7%	72%	79,52%	73,33%
Specificity	25,93%	2,78%	35,49%	0%	41,18%	0%
F1 Score	86,43%	83%	87,45%	83,72%	88,59%	84,61%



Gambar 4.10Diagram Nilai *Accuracy* Metode BOC dan BNB

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Melihat data dari Tabel 4.25 dan Gambar 4.11, bahwa nilai akurasi tertinggi dari ketiga skenario dengan metode BOC adalah pada skenario ketiga dengan nilai *accuracy* sebesar 81,1% di mana jumlah proporsi data adalah 60% data *training* dan 40% data *testing*. Sedangkan nilai akurasi dari metode BNB pada skenario yang sama sebesar 73,33%. Jika melihat pada nilai dari *F1 score* tertinggi metode BOC adalah pada skenario ketiga sebesar 88,59%, dan nilai *F1 Score* metode BNB adalah sebesar 84,61%. Rentang selisih nilai akurasi metode BOC dan BNB sebesar 7,78%. Sehingga dapat disimpulkan bahwa metode BOC lebih tepat digunakan dalam *dataset* pada penelitian ini dengan nilai akurasi lebih besar 7,78% daripada metode BNB.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

BAB V
PENUTUP

5.1 Kesimpulan

Penggunaan dan pemilihan metode disesuaikan dengan *dataset* yang digunakan, pada penelitian ini *dataset* bersifat data biner. Sehingga metode yang memungkinkan untuk dapat mengklasifikasi data dengan sifat tersebut salah satunya adalah Algoritma *Bayes*. Algoritma ini memiliki beberapa macam metode, seperti metode *BNB* dan *BOC*. Metode *BNB* bekerja secara *naive* (*simple* dan sederhana). Sedangkan metode *BOC* bekerja dengan memaksimalkan nilai dari probabilitas *posterior* sehingga hasil yang didapatkan dari proses tersebut lebih tinggi dibandingkan dengan metode *BNB*.

Berdasarkan hasil dari proses klasifikasi dan proses perhitungan yang dilakukan pada bab sebelumnya, dapat disimpulkan:

1. Untuk dapat melakukan implementasi metode *Bayes Optimal Classifier* pada *dataset history chat cyberathlete* dalam proses klasifikasi *cyberbullying*, pertama-tama adalah mengambil data mana yang layak untuk dilakukan proses klasifikasi *cyberbullying*. Langkah selanjutnya setelah dataset terkumpul adalah mengolah *dataset* tersebut, dengan proses pada *pre-processing text* seperti *cleansing, filtering, tokenizing, stemming* dan *lemmatization*. *Dataset* yang sudah dibersihkan dan diolah, kemudian *dataset* akan dimasukan ke sebuah kantung (*bag of words*) untuk menjadikan data tersebut ke bentuk *vector* sehingga *dataset* dapat dilakukan klasifikasi menggunakan metode *Bayes Optimal Classifier*.
2. Hasil evaluasi performa dari penggunaan metode *Bayes Optimal Classifier*, di mana *dataset* dibagi menjadi tiga macam skenario proporsi data yang kemudian hasilnya akan dibandingkan dengan metode *Bernoulli Naive Bayes*. Dari pembedaan proses ini metode *BOC* mendapatkan hasil akurasi tertinggi pada skenario pertama sebesar 81,11%, sedangkan pada metode *Bernoulli Naive Bayes* mendapat nilai akurasi sebesar 73,33%. Sehingga dari nilai akurasi yg didapatkan tersebut dapat ditarik kesimpulan bahwa selisih nilai akurasi antara metode *BOC* dan *BNB* sebesar 7,78%. Dari ketiga skenario

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

tersebut dengan menghitung rata-rata perbedaan nilai akurasi antara metode *BOC* dan *BNB*, mendapat nilai sebesar 7,54 %. Sehingga dari proses ini dapat disimpulkan bahwa metode *BOC* mendapatkan nilai akurasi lebih tinggi sebesar 7,54% daripada metode *BNB*, kesimpulan lainnya bahwa pembedaan perlakuan dapat mempengaruhi nilai *accuracy*, *precision*, dan *recall* pada metode yang digunakan.

5.2 Saran

Berdasarkan kesimpulan pada penelitian ini, berikut merupakan saran – saran yang dapat dilakukan untuk penelitian selanjutnya:

1. Pada proses pengolahan data, dilakukan proses tertentu tergantung dengan *dataset* yang digunakan. Seperti pada penelitian ini dataset terdapat banyak sekali data outlier, sehingga perlu banyak proses secara manual untuk memproses dataset tersebut ke dataset yang layak digunakan. Sebaiknya proses ini dilakukan dengan cara baru secara otomatis sehingga lebih memudahkan untuk proses penelitian.
2. Hasil akurasi dari metode yang digunakan akan selalu berbeda tergantung dengan *dataset* dan metode yang digunakan. Oleh karena itu pada metode yang digunakan dalam penelitian ini, hasilnya belum tentu sama dengan hasil dari metode lain yang digunakan.
3. Diharapkan pada penelitian selanjutnya dapat menerapkan metode lainnya seperti *Support Vector Machine*, *Logistic Regression*, *Random Forest*, dan metode klasifikasi teks lainnya.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

DAFTAR PUSTAKA

Adams, & Rollings. (2010). Development of the Game Addiction for Adults (GAIA). *Journal of Cognitive PsychoterapyJournal of Cognitive Psychoterapy*.

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Brownlee, J. (2019). *A Gentle Introduction to the Bayes Optimal Classifier*. https://machinelearningmastery.com/bayes-optimal-classifier/

Cahyani, R. (2019). *Analisis sentimen pada media sosial twitter terhadap tokoh publik peserta pilpres 2019*.

Cénat, J. ., Hébert, M., Blais, M., Lavoie, F., Guerrier, M., & Derivois, D. (2014). Cyberbullying, psychological distress and self-esteem among youth in Quebec schools. *Journal of Affective Disorders*.

Daniel, T. (2005). *An Introduction of Data Mining*.

Del Bosque, L. P., & Garza, S. E. (2018). *Cyberbullying Detection in Social Networks: A Multi-Stage Approach*. *148*(3), 285–296.

Dota 2 Wiki. (2013). Wikipedia. https://id.wikipedia.org/wiki/Dota_2

Even, Y. (2002). Introduction to Text Mining. Automated Learning Group National Center For Supercomputing Applicarions. *University of Illionis*.

Fadlil, A., Riadi, I., & Aji, S. (2017). DDoS Attacks Classification using Numeric Attribute-based Gaussian Naive Bayes. *International Journal of Advanced Computer Science and Applications*, *8*(8). https://doi.org/10.14569/ijacsa.2017.080806

Ginting, S. L. B., & Trinanda, R. P. (2015). Teknik Data Mining Menggunkan Metode Bayes Classifier Untuk Optimalisasi Pencarian Pada Aplikasi Perpusatakan. *Universitas Pasundan*, *d*(Pencarian Informasi), 1–14.

Hearst, M. (2003). What is text mining. *SIMS, UC Berkeley*, *5*.

Kalra, V., & Aggarwal, R. (2018). Importance of Text Data Preprocessing & Implementation in RapidMiner. *Proceedings of the First International Conference on Information Technology and Knowledge Management*, *14*, 71–75. https://doi.org/10.15439/2017km46

Karhulahti, V. M. (2017). Perkembangan eSport di Mata Internasional dan

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Pengakuan serta Pengaruhnya Terhadap Masyarakat di Indonesia. *Physical Culture and Sport, Studies and Research*, *74*(1), 43–53. https://doi.org/10.1515/pcssr-2017-0010

Kartiwi, M., & Gunawan, T. S. (2019). Cyberbullying in E-Learning Platform : An Exploratory Study Through Text Mining. *Journal of Information System and Digital Technologies*, *1*(2), 40–47.

Luqyana, W. A., Cholissodin, I., & Perdana, R. S. (2018). Analisis Sentimen Cyberbullying Pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *2*(11), 4704–4713.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to Information Retrieval*.

Margono, H. (2019). *Analysis of the Indonesian Cyberbullying through Data Mining: The Effective Identification of Cyberbullying through Characteristics of Messages*.

Mishna, F., Saini, M., & Solomon, S. (2009). Ongoing and online: Children and youth's perceptions of cyber bullying. *Children and Youth Services Review*.

O'Brien, N., & Moules, T. (2010). The impact of cyber-bullying on young people's mental health. *Chelmsford: Anglia Ruskin University*.

Opendota API. (2018). Opendota API Documentation. https://docs.opendota.com/

Orel, A., Campbell, M., Wozencroft, K., Leong, E., & Kimpton, M. (2017). Exploring University Stundents' Coping Strategy Intentions for Cyberbullying. *Journal of Interpersonal Violence, doi:10.1177/0886260515586363*.

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing*, *5*(2). https://doi.org/10.22364/bjmc.2017.5.2.05

Privitera, C., & Campbell, M. (2009). *Cyberbullying: the new face of workplace bullying?*

Raj, S., & Rajaraajeswari, S. (2016). A Framework for Text Analytics using the Bag of Words (BoW) Model for Prediction. *International Journal of Advanced*

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id

Networking & Applications (IJANA).

Sanders, J., Smith, P., & Cillessen, A. (2009). Cyberbullies: Their motives, characteristics, and types of bullying. *Fourteenth European Conference on Developmental Psychology*.

Sayfullina, L., Eirola, E., Komashinsky, D., Palumbo, P., & Miche, Y. (2015). *Efficient detection of zero-day Android Malware using Normalized Bernoulli Naive Bayes*.

Tong, S., & Koller, D. (2000). Restricted bayes optimal classifiers. *Proceedings of the National Conference on Artificial Intelligence*, 658–664. papers2://publication/uuid/2B83AF8F-FBC6-46BF-AFB4-DE6DB2AB990C

Urban Dictionary: Cyberathlete. (2004). Urbandictionary.Com. https://www.urbandictionary.com/define.php?term=Cyberathlete

Willard, N. (2006). Cyberbullying and cyverthreats. *Eugene, OR: Center for Safe and Responsible Internet Use*.

Zhang, J., Liu, T., & Tao, D. (2018). *On the Rates of Convergence from Surrogate Risk Minimizers to the Bayes Optimal Classifier*. 1–19. http://arxiv.org/abs/1802.03688

Zufri, M. (2017). Peningkatan Kualitas Atlet Pada KONI Kota Semarang dikaji dengan Analisis Value For Money (Studi Kasus Pada Program Pembinaan Atlet). *Ilmiah Untag*, *6*(1), 69–83.

digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id digilib.uinsby.ac.id