IMPLEMENTASI METODE *MULTINOMIAL NAIVE BAYES* UNTUK MENDETEKSI KICAUAN YANG MENGANDUNG UJARAN KEBENCIAN PADA DATA *TWITTER* BAHASA INDONESIA

SKRIPSI

Diajukan untuk Memenuhi Bagian Dari

Syarat Memperoleh Gelar Sarjana Komputer

Program Studi Ilmu Komputer

Umar Syahid Aulia Rahman

1405681

PROGRAM STUDI ILMU KOMPUTER

DEPARTEMEN PENDIDIKAN ILMU KOMPUTER

FAKULTAS PENDIDIKAN MATEMATIKA

DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS PENDIDIKAN INDONESIA

BANDUNG

2020

IMPLEMENTASI METODE *MULTINOMIAL NAIVE BAYES* UNTUK MENDETEKSI KICAUAN YANG MENGANDUNG UJARAN KEBENCIAN PADA DATA *TWITTER* BAHASA INDONESIA

oleh

Umar Syahid Aulia Rahman
1405681

Disetujui dan Disahkan oleh:
Pembimbing I

Dr. Yudi Wibisono, M.T.
NIP: 197507072003121003
Pembimbing II

Eddy Prasetyo Nugroho, M.T.
NIP: 197505152008011014
Mengetahui,
Ketua Departemen Pendidikan Ilmu Komputer

Dr. Lala Septem Riza, M.T.
NIP: 197809262008121001

KATA PENGANTAR

Segala puji bagi Allah SWT yang telah memberikan rahmat dan karunia-Nya kepada penulis, sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Shalawat dan salam senantiasa tercurah kepada Rasulullah SAW yang mengantarkan manusia dari zaman kegelapan ke zaman yang terang benderang ini. Penyusunan skripsi ini dimaksudkan untuk memenuhi sebagian syarat-syarat guna mencapai gelar Sarjana Ilmu Komputer di Universitas Pendidikan Indonesia.

Penulis menyadari bahwa penulisan ini tidak dapat terselesaikan tanpa dukungan dari berbagai pihak baik moril maupun materil. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih kepada semua pihak yang telah membantu dalam penyusunan skripsi ini terutama kepada:

1. Kedua Orang tua beserta adik dan kakak yang telah memberikan doa dan dukungan selama proses pembuatan skripsi.
2. Segenap keluarga dan teman yang telah menyemangati dan membantu penyelesaian skripsi ini.
3. Ibu Siti Fatimah, S.Pd., M.Si., Ph.D., selaku Dekan Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam UPI.
4. Bapak Dr. Lala Septem Riza, M.T., selaku Ketua Departemen Pendidikan Ilmu Komputer UPI.
5. Ibu Dr. Rani Megasari, M.T., selaku Ketua Prodi Ilmu Komputer UPI
6. Bapak Dr. Yudi Wibisono, M.T., selaku Pembimbing Skripsi I yang telah berkenan memberikan tambahan ilmu dan solusi pada setiap permasalahan atas kesulitan dalam penulisan skripsi ini.
7. Bapak Eddy Prasetyo Nugroho, M.T., selaku Pembimbing Skripsi II yang telah berkenan memberikan tambahan ilmu dan solusi pada setiap permasalahan atas kesulitan dalam penulisan skripsi ini.

8. Bapak Eki Nugraha, S.Pd., M.Kom., selaku Dosen Wali yang telah membimbing selama perkuliahan.
9. Seluruh Bapak/Ibu dosen Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam UPI yang telah memberikan pengetahuan yang sangat bermanfaat selama masa perkuliahan.
10. Seluruh teman-teman seangkatan, terutama kelas C2 Ilmu Komputer Angkatan 2014 yang selalu saling membantu dalam kegiatan kuliah dan saling menyemangati dalam pengerjaan skripsi.
11. Semua pihak yang tidak dapat disebutkan satu persatu yang telah membantu memberikan dukungan dalam pengerjaan skripsi.

Tangerang, 17 Juni 2020
Penulis,

(Umar Syahid Aulia Rahman)

ABSTRAK

Pada Peneliltian ini kami membahas klasifikasi ujaran kebencian pada data kicauan (Twitter) dalam bahasa Indonesia dimana pada penelitian ini kami membangun *dataset* ujaran kebencian pada kicauan bahasa Indonesia dan melakukan pengklasifikasian dengan mengimplementasikan algoritma *Multinomial Naive Bayes* dengan menggunakan ekstraksi fitur *term frequency – inverse document frequency* (TF-IDF). Pada penelitian kami melakukan beberapa konfigurasi dalam modifikasi *data training* untuk mengatasi *imbalanced dataset* yaitu dengan menggunakan metode *random oversampling* dan *random undersampling*. Dari eksperimen tersebut kami melakukan evaluasi *menggunakan confusion matrix* dan didapatkan hasil implementasi metode *Multinomial Naive Bayes* dengan modifikasi *data training menggunakan random oversampling* dengan rasio *data testing* 10% *memiliki* hasil yang paling bagus dengan *fmeasure* sebesar 0.5307.

Kata Kunci—*Dataset Construction, Hate Speech Classification, Imbalanced Dataset, Multinomial Naive Bayes Classifier, Term Frequency Inverse Document Frequency*

ABSTRACT

In this reserach we discuss the classification of hate speech on Twitter data in Bahasa Indonesia where in this research we build hate speech datasets on Bahasa Indonesia and classify the classification by implementing the Multinomial Naive Bayes algorithm using the extraction of the term frequency – Inverse document frequency (TF-IDF) feature. In the research we do some configuration in the modification of training data to overcome the imbalanced dataset that is using the random oversampling and random undersampling methods. From the experiments we evaluated using confusion matrix and obtained the results of implementation of Multinomial Naive Bayes method with the modification of training data using random oversampling with testing data ratio 10% has the best results with f-measure of 0.5307.

Keywords —*construction dataset, Hate Speech Classification, Imbalanced Dataset, Multinomial Naive Bayes Classifier, Term Frequency Inverse Document Frequency*

DAFTAR ISI

KATA PENGANTAR i

ABSTRAK iii

ABSTRACT iv

DAFTAR ISI v

DAFTAR TABEL viii

DAFTAR GAMBAR ix

BAB I PENDAHULUAN 1

1.1 Latar Belakang 1

1.2 Rumusan Masalah 4

1.3 Tujuan Penelitian 5

1.4 Batasan Masalah 5

1.5 Sistematika Penulisan 5

BAB II KAJIAN PUSTAKA 8

2.1 Ujaran Kebencian 8

2.1.1 Jenis Ujaran Kebencian 10

2.2 Klasifikasi Teks 10

2.2.1 Representasi Teks 12

2.3 *Dataset* 14

2.3.1 Imbalanced Dataset 16

2.4 *Machine Learning* 18

2.4.1 Naïve Bayes Classifier 20

2.5 Twitter 25

2.5.1 Praproses Twitter 26

2.6 *Confusion Matrix* 27

BAB III METODOLOGI 29

3. 1 Desain Penelitian 29

3.2 Alat dan Bahan Penelitian 31

3.3 *Metode Penelitian* 32

BAB IV PEMBAHASAN 34

4.1 *Dataset* 34

4.1.1 Proses Pengumpulan *Dataset* 34

4.1.2 Pelabelan Data 37

4.1.3 *Library* Pembangunan *Dataset* 37

4.2 Praproses 38

4.2.1 *Library* Praproses 40

4.2.3 Implementasi Kode Program Praproses 40

4.3 Tokenisasi dan *Stemming* 41

4.3.1 *Library* Tokenisasi dan *Stemming* 44

4.3.3 Implementasi Kode Program Tokenisasi dan *Stemming* 44

4.4 Penanganan Imbalanced Dataset 45

4.4.1 Library Penanganan Imbalanced Dataset 46

4.4.2 Implementasi Kode Program Penanganan *Imabalanced Dataset* 47

4.5 Ekstraksi Fitur 49

4.5.1 Pembobotan TF-IDF 49

4.5.2 Vectorize 53

4.5.3 *Library* Ekstraksi Fitur 55

4.5.4 Implementasi Kode Program Ekstraksi Fitur 55

4.6 Proses Learning dan Testing 57

4.6.1 Learning M*ultinomial Naive Bayes* (MNB) 57

4.6.2 Testing Multinomial Naive Bayes (MNB) 60

4.6.3 Library Proses Learning dan Testing 62

4.6.4 Implementasi Kode Program *Learning* dan *Testing* 62
4.7 Evaluasi Confusion Matrix 64
4.7.1 Library Evaluasi Confusion Matrix 65
4.7.2 Implementasi Kode Program *Confusion Matrix* 66
4.8 Konfigurasi Eksperimen 66
4.9 Analisis Hasil 67
BAB V KESIMPULAN DAN SARAN 73
5.1 Kesimpulan 73
5.2 Saran 73
DAFTAR PUSTAKA 75
LAMPIRAN 78

DAFTAR TABEL

Tabel 2.1 Contoh kasus TF-IDF.....14
Tabel 2.2 *Confusion matrix*.....28
Tabel 4.1 Kata kunci untuk proses *crawling*.....36
Tabel 4.2 *Library* pembangunan *dataset*.....38
Tabel 4.3 *Library* praproses.....41
Tabel 4.4 *Library* tokenisasi dan *stemming*.....45
Tabel 4.5 *Library* penanganan *imbalanced dataset*.....48
Tabel 4.6 *Library* ekstraksi fitur.....56
Tabel 4.7 *Library* proses *learning* dan *testing*.....63
Tabel 4.8 *Confusion matrix*.....65
Tabel 4.9 *Library confusion matrix*.....67
Tabel 4.10 Konfigurasi dan hasil eksperimen.....68
Tabel 4.11 Konfigurasi dan hasil eksperimen menggunakan *dataset* (Ibrohim & Budi 2019) 68

DAFTAR GAMBAR

Gambar 2.1 Alur proses klasifikasi(Devi & Saharia, 2020).....11
Gambar 2.2 Alur proses klasifikasi.....11
Gambar 2.3 Alur proses pembangunan dataset(Ibrohim & Budi, 2019).....15
Gambar 2.4 Alur proses pembangunan dataset(Alfina, Mulia, Fanany, & Ekanata, 2017).....16
Gambar 3.1 Desain penelitian.....31
Gambar 4.1 Alur proses pembangunan dataset pada penelitian ini.....35
Gambar 4.2 Kode program pengambilan data kicauan bahasa Indonesia 36
Gambar 4.3 *Dataset* hasil *crawling*.....37
Gambar 4.4 *Flowchart* praproses.....40
Gambar 4.5 *Flowchart* proses tokenisasi.....43
Gambar4.6 *Flowchart* proses *stemming*.....44
Gambar4.7 *Data training* setelah *oversampling*.....46
Gambar4.8 *Data training* setelah *undersampling*.....47
Gambar4.9 *Flowchart* proses penghitungan *term frequency*.....51
Gambar4.10 *Flowchart* proses penghitungan *document frequency*.....52
Gambar4.11 *Flowchart* proses penghitungan *inverse document frequency*.....53
Gambar4.12 *Flowchart* proses penghitungan TF-IDF.....54
Gambar4.13 *Flowchart* proses *Vectorize*.....55
Gambar4.14 *Flowchart* proses penghitungan total TF-IDF.....59
Gambar4.15 *Flowchart* proses penghitungan kata unik.....60
Gambar4.16 *Flowchart* proses *learning*.....61
Gambar4.17 *Flowchart* proses *testing*.....62
Gambar4.18 *Flowchart* proses penghitungan *confusion matrix*.....66
Gambar4.19 Tabel rata rata f-measure penelitian(Alfina et al., 2017).....71
Gambar4.20 Tabel hasil penelitian(Ibrohim & Budi, 2018) 73