PERBANDINGAN METODE-METODE KLASIFIKASI UNTUK ANALISIS SENTIMEN BERDASARKAN PEMBENTUKAN TEKNIK *VECTOR SPACE MODEL*

Diajukan Sebagai Syarat untuk Menyelesaikan
Pendidikan Program Strata-1 pada
Jurusan Teknik Informatika Fakultas Ilmu Komputer
Universitas Sriwijya

Oleh :

M Agung Pribadi
09021181520036

JURUSAN TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2019

LEMBAR PENGESAHAN TUGAS AKHIR

PERBANDINGAN METODE-METODE KLASIFIKASI
UNTUK ANALISIS SENTIMEN BERDASARKAN
PEMBENTUKAN TEKNIK *VECTOR SPACE MODEL*

Oleh :

M AGUNG PRIBADI

NIM : 09021181520036

Indralaya, September 2019

Pembimbing I,

Novi Yusliani, M.T.
NIP. 198211082012122001

Pembimbing II,

Ir. M. Ihsan Jambak, M.Sc., M.M.
NIP. 196804052013081201

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T.
NIP. 199706012009121004

TANDA LULUS SIDANG TUGAS AKHIR

Pada hari Jumat, 13 September 2019 telah dilaksanakan ujian sidang tugas akhir oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : M Agung Pribadi

NIM : 09021181520036

Judul : Perbandingan Metode-Metode Klasifikasi untuk Analisis Sentimen Berdasarkan Pembentukan Teknik *Vector Space Model*

1. Pembimbing I

   Novi Yusliani, M.T.
   NIP. 198211082012122001 .....

2. Pembimbing II

   Ir. M. Ihsan Jambak, M.Sc., M.M.
   NIP. 196804052013081201 .....

3. Penguji I

   Alvi Syahrini Utami, M.Kom.
   NIP. 197812222006042003 .....

4. Penguji II

   Kanda Januar Miraswan, M.T.
   NIP. 199001092019031012 .....

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T.
NIP. 197706012009121004

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama	: M Agung Pribadi
NIM	: 09021181520036
Program Studi	: Teknik Informatika
Judul Skripsi	: Perbandingan Metode-Metode Klasifikasi untuk Analisis Sentimen berdasarkan Pembentukan Teknik *Vector Space Model*
Hasil *iThenticate/Turnitin*	: 19%

Menyatakan bahwa laporan proyek saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan proyek ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian pernyataan ini saya buat dengan sebenar-benarnya dan tidak ada unsur paksaan oleh siapapun.

Indralaya, 22 September 2019

M Agung Pribadi
NIM. 09021181520036

iv

You never fail until you stop trying.

Kupersembahkan karya tulis ini kepada :

- Orang tua terbaikku
- Kakak-kakak kesayanganku
- Keluarga terhebatku
- Rekan-rekan seperjuanganku
- Almamater kebanggaanku

COMPARISON OF CLASSIFICATION METHODS
FOR SENTIMENT ANALYSIS BASED ON
VECTOR SPACE MODEL TECHNIQUES

by:
M Agung Pribadi
09021181520036

ABSTRACT

The main task in sentiment analysis is to classify sentiments. There are several classification methods based on the type of classification approach, including the probability approach to the appearance of an attribute in a class, one of which is the Multinomial Naïve Bayes; approach to the distance of an attribute, one of which is Support Vector Machine; and binary decision approach to an attribute, one of them is Logistic Regression. Before classifying sentiments, a technique for transforming sentiments in the form of text into numerical techniques is called a vector space model technique. Vector space model techniques include Bag-Of-Words and Doc2Vec. This study will compare the application of the Multinomial Naïve Bayes classification method, Support Vector Machine and Logistic Regression to the Bag-Of-Words and Doc2Vec vector space model techniques. The most effective sentiment classification results in Bag Of Words vector space model technique is using the Multinomial Naïve Bayes classification method with an average accuracy of 85.140% and an average classification time of 0.154 seconds. In the vector space model of the Doc2Vec model, the most effective sentiment classification results use the Logistic Regression classification method with an average accuracy of 62.467% and an average classification time of 0.450 seconds.

Keywords: Classification Method, Vector Space Model Technique, Multinomial Naive Bayes, Logistic Regression, Support Vector Machine, Bag-Of-Words, Doc2Vec.

Indralaya, September 2019

Pembimbing I,

Novi Yusliani, M.T.
NIP. 198211082012122001

Pembimbing II,

Ir. M. Ihsan Jambak, M.Sc., M.M.
NIP. 196804052013081201

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T.
NIP. 197706012009121004

PERBANDINGAN METODE-METODE KLASIFIKASI
UNTUK ANALISIS SENTIMEN BERDASARKAN
PEMBENTUKAN TEKNIK *VECTOR SPACE MODEL*

oleh:
M Agung Pribadi
09021181520036

ABSTRAK

Tugas utama dalam analisis sentimen adalah melakukan klasifikasi terhadap sentimen. Terdapat beberapa metode klasifikasi berdasarkan jenis pendekatan pengklasifikasiannya, diantaranya pendekatan probabilitas kemunculan suatu atribut di dalam suatu kelas, salah satunya *Multinomial Naïve Bayes;* pendekatan pada jarak suatu atribut, salah satunya *Support Vector Machine;* dan pendekatan *binary decision* pada suatu atribut, salah satunya *Logistic Regression.* Sebelum melakukan klasifikasi sentimen, diperlukan teknik mentransformasi sentimen yang berupa teks menjadi numerik yang dinamakan teknik *vector space model*. Teknik *vector space model* diantaranya *Bag-Of-Words* dan *Doc2Vec.* Penelitian ini akan membandingkan penerapan metode klasifikasi *Multinomial Naïve Bayes, Support Vector Machine* dan *Logistic Regression* pada teknik *vector space model Bag-Of-Words* dan *Doc2Vec.* Hasil klasifikasi sentimen yang paling efektif pada teknik *vector space model Bag Of Words* yaitu menggunakan metode klasifikasi *Multinomial Naïve Bayes* dengan rata-rata akurasi sebesar 85,140% dan rata-rata waktu pengklasifikasian sebesar 0,154 detik. Pada teknik *vector space model Doc2Vec,* hasil klasifikasi sentimen yang paling efektif menggunakan metode klasifikasi *Logistic Regression* dengan rata-rata akurasi sebesar 62,467% dan rata-rata waktu pengklasifikasian sebesar 0,450 detik.

Kata Kunci: Metode Klasifikasi, Teknik *Vector Space Model, Multinomial Naïve Bayes, Logistic Regression, Support Vector Machine, Bag-Of-Words, Doc2Vec.*

Indralaya, September 2019

Pembimbing I,

Novi Yusliani, M.T.
NIP. 198211082012122001

Pembimbing II,

Ir. M. Ihsan Jambak, M.Sc., M.M.
NIP. 196804052013081201

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T.
NIP. 197706012009121004

KATA PENGANTAR

Puji syukur kepada Allah SWT atas berkat dan rahmat-Nya yang telah diberikan kepada Penulis sehingga dapat menyelesaikan Tugas Akhir ini dengan baik. Tugas akhir ini disusun untuk memenuhi salah satu syarat guna menyelesaikan pendidikan program Strata-1 pada Fakultas Ilmu Komputer Program Studi Teknik Informatika di Universitas Sriwijaya.

Dalam menyelesaikan Tugas Akhir ini banyak pihak yang telah memberikan bantuan dan dukungan baik secara langsung maupun secara tidak langsung. Penulis ingin menyampaikan rasa terima kasih kepada:

1. Orang tuaku Ujang Yusmana dan Rosnila; kakak-kakakku Viesca Septianata, Meta Gapila dan Melda Miranda; serta seluruh keluarga besarku yang selalu mendoakan serta memberikan dukungan baik moril maupun materiil.
2. Bapak Jaidan Jauhari, S.Pd., M.T. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
3. Bapak Rifkie Primartha, S.T., M.T. selaku Ketua Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
4. Bapak Muhammad Ali Buchari, M.T. selaku dosen pembimbing akademik.
5. Ibu Novi Yusliani, M.T. selaku dosen pembimbing I dan Bapak Ir. M. Ihsan Jambak, M.Sc., M.M. selaku dosen pembimbing II yang telah membimbing, mengarahkan dan memberikan motivasi dalam proses perkuliahan dan pengerjaan Tugas Akhir.
6. Ibu Alvi Syahrini Utami, M.Kom. selaku dosen penguji I dan Bapak Kanda Januar Miraswan, M.T. selaku dosen penguji II yang telah memberikan masukan dan dorongan dalam proses pengerjaan Tugas Akhir.
7. Seluruh dosen Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
8. Kak Ricy Firnando dan Mbak Winda serta seluruh staf tata usaha yang telah membantu dalam kelancaran proses administrasi dan akademik selama masa perkuliahan.

9. Manawa Salwa Fadilla yang selalu menemani, membantu serta memberikan semangat, motivasi dan kasih sayang kepada penulis dalam menyelesaikan perkuliahan dan tugas akhir skripsi.
10. Kerel, Ardi, Tirta, Udin, Albar, Gian, Blend, Halim, Genda, Ikrom, Rasyid, Ayep, Faiz, Ega, Taca, Vira, dan Arfah yang penulis sendiri tidak tahu apa yang telah mereka bantu selama proses pengerjaan Tugas Akhir, serta teman-teman IF Reguler dan Bilingual 2015 yang selalu berjuang bersama-sama walaupun kalau sedang ujian tetap berjuang sendiri-sendiri.
11. Rahmad, Singgih, Irfan, Ucok, Deri dan teman-teman lainnya yang menemani penulis apabila penulis pulang kampung.
12. Stack Overflow, Medium, Github, Towards Data Science dan media pembelajaran lainnya yang telah membantu penulis dalam menambah wawasan pemrograman.

Penulis menyadari dalam penyusunan Tugas Akhir ini masih terdapat banyak kekurangan disebabkan keterbatasan pengetahuan dan pengalaman, oleh karena itu kritik dan saran yang membangun sangat diharapkan untuk kemajuan penelitian selanjutnya.

Akhir kata semoga Tugas Akhir ini dapat berguna dan bermanfaat bagi kita semua.

Indralaya, September 2019

M Agung Pribadi

DAFTAR ISI

Halaman

HALAMAN JUDUL i
LEMBAR PENGESAHAN SKRIPSI ii
TANDA LULUS UJIAN SKRIPSI iii
HALAMAN PERNYATAAN iv
MOTTO DAN PERSEMBAHAN v
ABSTRACT vi
ABSTRAK vii
KATA PENGANTAR viii
DAFTAR ISI x
DAFTAR TABEL xiv
DAFTAR GAMBAR xix
DAFTAR LAMPIRAN xxii

BAB I PENDAHULUAN I-1
1.1. Pendahuluan I-1
1.2. Latar Belakang Masalah I-1
1.3. Rumusan Masalah I-4
1.4. Tujuan Penelitian I-4
1.5. Manfaat Penelitian I-5
1.6. Batasan Masalah I-5
1.7. Sistematika Penulisan I-6
1.8. Kesimpulan I-7

BAB II KAJIAN LITERATUR II-1
2.1. Pendahuluan II-1
2.2. Landasan Teori II-1
2.2.1. *Logistic Regression* II-1
2.2.2. *Multinomial Naive Bayes* II-4
2.2.3. *Support Vector Machine* II-6

2.2.4. *Vector Space Model* (VSM) II-10
2.2.5. *Area Under the Receiver Operating Characteristics* (AUROC) II-16
2.2.6. *K-Fold Cross Validation* II-18
2.3. Penelitian Lain yang Relevan II-19
2.4. Kesimpulan II-21

BAB III METODOLOGI PENELITIAN III-1
3.1. Pendahuluan III-1
3.2. Metode Pengumpulan Data III-1
3.3. Tahapan Penelitian III-1
3.3.1. Menetapkan Kerangka Kerja III-2
3.3.2. Menetapkan Kriteria Pengujian III-5
3.3.3. Menetapkan Format Data Pengujian III-5
3.3.4. Menentukan Alat yang Digunakan dalam Pelaksanaan Penelitian III-6
3.3.5. Melakukan Pengujian Penelitian III-7
3.3.6. Melakukan Analisis Hasil Pengujian dan Membuat Kesimpulan III-7
3.4. Metode Pengembangan Perangkat Lunak III-8
3.4.1. Fase Insepsi III-9
3.4.2. Fase Elaborasi III-9
3.4.3. Fase Konstruksi III-9
3.4.4. Fase Transisi III-10
3.5. Manajemen Proyek Penelitian III-10
3.6. Kesimpulan III-22

BAB IV PENGEMBANGAN PERANGKAT LUNAK IV-1
4.1. Pendahuluan IV-1
4.2. Fase Insepsi IV-1
4.2.1. Pemodelan Bisnis IV-2
4.2.2. Kebutuhan Sistem IV-4
4.2.3. Analisis dan Desain IV-18

4.3. Fase Elaborasi IV-19
4.3.1. Pemodelan Bisnis IV-19
4.3.2. Kebutuhan Sistem IV-26
4.3.3. Analisis Desain IV-29
4.4. Fase Konstruksi IV-34
4.4.1. Pemodelan Bisnis IV-34
4.4.2. Kebutuhan Sistem IV-36
4.5. Fase Transisi IV-38
4.5.1. Pemodelan Bisnis IV-38
4.5.2. Kebutuhan IV-38
4.5.3. Pengujian IV-39
4.6. Kesimpulan IV-52

BAB V HASIL DAN ANALISIS PENELITIAN V-1
5.1. Pendahuluan V-1
5.2. Data Uji Hasil Penelitian V-1
5.2.1. Hasil Pengujian Pengklasifikasian pada Teknik VSM *Bag-Of-Words* V-1
5.2.2. Hasil Pengujian Pengklasifikasian pada Teknik VSM *Doc2Vec* V-4
5.3. Analisis Hasil Uji Penelitian V-8
5.3.1. Analisis Hasil Pengujian Pengklasifikasian pada Teknik VSM *Bag Of Words* V-8
5.3.2. Analisis Hasil Pengujian Pengklasifikasian pada Teknik VSM *Doc2Vec* V-21
5.3.3. Analisis Hasil Pengujian Pengklasifikasian pada Metode Klasifikasi *Logistic Regression* V-31
5.3.4. Analisis Hasil Pengujian Pengklasifikasian pada Metode Klasifikasi *Support Vector Machine* V-35
5.4. Analisis Hasil Penelitian V-40
5.4.1. Analisis Penentuan Parameter pada Metode Klasifikasi *Logistic Regression* V-40
5.4.2. Analisis Perbandingan Metode Klasifikasi berdasarkan Pembentukan VSM V-41

5.5. Kesimpulan V-43

BAB VI KESIMPULAN DAN SARAN VI-1
6.1. Pendahuluan VI-1
6.2. Kesimpulan VI-1
6.3. Saran VI-3

DAFTAR PUSTAKA xxiii

DAFTAR TABEL

Halaman

Tabel II-1. Skema 10 *fold cross validation*..... II-19
Tabel III-1. Rancangan Tabel Hasil Klasifikasi..... III-5
Tabel III-2. Rancangan Tabel Hasil Analisis Klasifikasi..... III-8
Tabel III-3. Penjadwalan Penelitian dalam Bentuk *Work Breakdown Structure* (WBS)..... III-11
Tabel IV-1. Definisi Aktor *Use Case*..... IV-3
Tabel IV-2. Definisi *Use Case*..... IV-3
Tabel IV-3. Skenario *Use Case* melakukan praproses data teks..... IV-5
Tabel IV-4. Skenario *Use Case* Pembentukan VSM menggunakan *Bag-Of-Words*..... IV-7
Tabel IV-5. Skenario *Use Case* Pembentukan VSM menggunakan *Doc2Vec*..... IV-9
Tabel IV-6. Skenario *Use Case* Melakukan Klasifikasi Sentimen dengan *Logistic Regression* IV-12
Tabel IV-7. Skenario *Use Case* Melakukan Klasifikasi Sentimen dengan *Multinomial Naive Bayes* IV-14
Tabel IV-8. Skenario *Use Case* Melakukan Klasifikasi Sentimen dengan *Support Vector Machine* IV-16
Tabel IV-9. Skenario *Use Case* Mengevaluasi Hasil Pengklasifikasian Menggunakan AUROC IV-18
Tabel IV-10. Tabel Implementasi Kelas..... IV-36
Tabel IV-11. Rencana Pengujian *Use Case* melakukan Praproses Data Teks..... IV-39
Tabel IV-12. Rencana Pengujian *Use Case* melakukan Pembentukan VSM *Bag-Of-Words*..... IV-40
Tabel IV-13. Rencana Pengujian *Use Case* melakukan Pembentukan VSM *Doc2Vec*..... IV-40

Tabel IV-14. Rencana Pengujian *Use Case* Melakukan Klasifikasi dengan *Logistic Regression* ……..... IV-41
Tabel IV-15. Rencana Pengujian *Use Case* Melakukan Klasifikasi dengan *Multinomial Naive Bayes* ……..... IV-42
Tabel IV-16. Rencana Pengujian *Use Case* Melakukan Klasifikasi dengan *Support Vector Machine* ……..... IV-42
Tabel IV-17. Rencana Pengujian *Use Case* Mengevaluasi Hasil Pengklasifikasian dengan AUROC ……..... IV-43
Tabel IV–18. Pengujian Fitur Praproses Data Teks……..…….... IV-44
Tabel IV–19. Pengujian Fitur Pembentukan VSM *Bag-Of-Words*…..... IV-45
Tabel IV–20. Pengujian Fitur Pembentukan VSM *Doc2Vec*……..…..... IV-46
Tabel IV–21. Pengujian Fitur Pengklasifikasian dengan Metode Klasifikasi *Logistic Regression*……..... IV-47
Tabel IV–22. Pengujian Fitur Pengklasifikasian dengan Metode Klasifikasi *Multinomial Naive Bayes* ..……..... IV-49
Tabel IV–23. Pengujian Fitur Pengklasifikasian dengan Metode Klasifikasi *Support Vector Machine*……..... IV-50
Tabel IV–24. Pengujian Fitur Evaluasi Hasil Pengklasifikasian Menggunakan AUROC…….……..... IV-51
Tabel V-1. Hasil Pengujian Parameter Jumlah Iterasi dengan *Learning Rate* 0.1…….…….……..... V-2
Tabel V-2. Hasil Pengujian Parameter Learning Rate dengan Jumlah Iterasi 20…….…….……..... V-3
Tabel V-3. Hasil Pengujian Pengklasifikasian pada Teknik VSM *Bag Of Words* …….…….……..... V-4
Tabel V-4. Hasil Pengujian Parameter Jumlah Iterasi dengan *Learning Rate* 0.1…….…….……..... V-5
Tabel V-5. Hasil Pengujian Parameter Learning Rate dengan Jumlah Iterasi 20…….…….…….…..... V-6
Tabel V-6. Hasil Pengujian Pengklasifikasian pada Teknik VSM *Doc2Vec*…….…….……..... V-7

Tabel V-7. Hasil Uji Normalitas Penentuan Jumlah Iterasi Optimum *Logistic Regression* pada Teknik VSM *Bag Of Words*…….. V-9
Tabel V-8. Hasil Uji Kruskal-Wallis H Penentuan Jumlah Iterasi Optimum *Logistic Regression* pada Teknik VSM *Bag Of Words*…….. V-10
Tabel V-9. Rata-rata Hasil Akurasi pada Jumlah Iterasi Masukan *Logistic Regression* pada Teknik VSM *Bag Of Words*…….. V-10
Tabel V-10. Hasil Uji Normalitas Penentuan *Learning Rate* Optimum *Logistic Regression* pada Teknik VSM *Bag Of Words*…….. V-11
Tabel V-11. Hasil Uji Kruskal-Wallis H Penentuan *Learning Rate* Optimum *Logistic Regression* pada Teknik VSM *Bag Of Words*…….. V-12
Tabel V-12. Rata-rata Hasil Akurasi pada Nilai *Learning Rate* Masukan *Logistic Regression* pada Teknik VSM *Bag Of Words*…….. V-13
Tabel V-13. Hasil Uji Normalitas Akurasi Pengklasifikasian pada Teknik VSM *Bag Of Words*…….. V-13
Tabel V-14. Hasil Uji Kruskal-Wallis H Akurasi Pengklasifikasian pada Teknik VSM *Bag Of Words.*…….. V-14
Tabel V-15. Hasil Uji Mann-Whitney U pada Hasil Akurasi Metode Klasifikasi *Multinomial Naive Bayes* dengan *Logistic Regression*…….. V-15
Tabel V-16. Hasil Uji Mann-Whitney U pada Hasil Akurasi Metode Klasifikasi *Multinomial Naive Bayes* dengan *Support Vector Machine*…….. V-15
Tabel V-17. Hasil Uji Mann-Whitney U pada Hasil Akurasi Metode Klasifikasi *Support Vector Machine* dengan *Logistic Regression*…….. V-16
Tabel V-18. Mean Hasil Akurasi Pengklasifikasian Teknik VSM *Bag Of Words* …….. V-16
Tabel V-19. Hasil Uji Normalitas Akurasi Pengklasifikasian pada Teknik VSM *Bag Of Words*…….. V-17
Tabel V-20. Hasil Uji Kruskal-Wallis H Waktu Pengklasifikasian pada Teknik VSM *Bag Of Words*…….. V-18
Tabel V-21. Hasil Uji Mann-Whitney U pada Waktu Pengklasifikasian Metode Klasifikasi *Multinomial Naive Bayes* dengan *Logistic Regression*..... V-19

Tabel V-22. Hasil Uji Mann-Whitney U pada Waktu Pengklasifikasian Metode Klasifikasi *Multinomial Naive Bayes* dengan *Support Vector Machine*..... V-19
Tabel V-23. Hasil Uji Mann-Whitney U pada Waktu Pengklasifikasian Metode Klasifikasi *Support Vector Machine* dengan *Logistic Regression*…..... V-20
Tabel V-24. Mean Waktu Pengklasifikasian Teknik VSM *Bag Of Words*.. V-20
Tabel V-25. Hasil Uji Normalitas Penentuan Jumlah Iterasi Optimum *Logistic Regression* pada Teknik VSM *Doc2Vec*…….. V-22
Tabel V-26. Hasil Uji Homogenitas Penentuan Jumlah Iterasi Optimum *Logistic Regression* pada Teknik VSM *Doc2Vec*…….. V-22
Tabel V-27. Hasil Uji ANOVA Jumlah Iterasi Optimum *Logistic Regression* pada Teknik VSM *Doc2Vec*…….. V-23
Tabel V-28. Rata-rata Hasil Akurasi pada Jumlah Iterasi Masukan *Logistic Regression* pada Teknik VSM *Doc2Vec*…….. V-23
Tabel V-29. Hasil Uji Normalitas Penentuan *Learning Rate* Optimum *Logistic Regression* pada Teknik VSM *Doc2Vec*…….. V-24
Tabel V-30. Hasil Uji Homogenitas Penentuan *Learning Rate* Optimum *Logistic Regression* pada Teknik VSM *Doc2Vec*…….. V-25
Tabel V-31. Hasil Uji ANOVA *Learning Rate* Optimum *Logistic Regression* pada Teknik VSM *Doc2Vec*…….. V-25
Tabel V-32. Rata-rata Hasil Akurasi pada Nilai *Learning Rate* Masukan *Logistic Regression* pada Teknik VSM *Doc2Vec*..…….. V-26
Tabel V-33. Hasil Uji Normalitas Akurasi Pengklasifikasian pada Teknik VSM *Doc2Vec*…….. V-27
Tabel V-34. Hasil Uji *Independent Sample T-Test* Akurasi Pengklasifikasian pada Teknik VSM *Doc2Vec*…….. V-28
Tabel V-35. Mean Hasil Akurasi Pengklasifikasian Teknik VSM *Doc2Vec*.V-29
Tabel V-36. Hasil Uji Normalitas Akurasi Pengklasifikasian pada Teknik VSM *Doc2Vec*…….. V-29
Tabel V-37. Hasil Uji Mann-Whitney U pada Waktu Pengklasifikasian Metode Klasifikasi *Support Vector Machine* dengan *Logistic Regression*…….. V-30
Tabel V-38. Mean Waktu Pengklasifikasian Teknik VSM *Doc2Vec*…….. V-30

Tabel V-39. Hasil Uji Normalitas Perbandingan Akurasi Pengklasifikasian pada Metode Klasifikasi *Logistic Regression*…….. V-32
Tabel V-40. Hasil Uji Mann Whitney U Perbandingan Akurasi Pengklasifikasian pada Metode Klasifikasi *Logistic Regression*…….. V-32
Tabel V-41. Mean Hasil Akurasi Pengklasifikasian pada Metode Klasifikasi *Logistic Regression*…….. V-33
Tabel V-42. Hasil Uji Normalitas Perbandingan Waktu Pengklasifikasian pada Metode Klasifikasi *Logistic Regression*…….. V-34
Tabel V-43. Hasil Uji Mann-Whitney U Perbandingan Waktu Pengklasifikasian pada Metode Klasifikasi *Logistic Regression*…….. V-34
Tabel V-44. Mean Waktu Komputasi Pengklasifikasian pada Metode Klasifikasi *Logistic Regression*…….. V-35
Tabel V-45. Hasil Uji Normalitas Perbandingan Akurasi Pengklasifikasian pada Metode Klasifikasi *Support Vector Machine*…….. V-36
Tabel V-46. Hasil Uji *Independent Sample T-Test* Perbandingan Akurasi Pengklasifikasian pada Metode Klasifikasi *Support Vector Machine*…….. V-37
Tabel V-47. Mean Hasil Akurasi Pengklasifikasian pada Metode Klasifikasi *Support Vector Machine*…….. V-37
Tabel V-48. Hasil Uji Normalitas Perbandingan Waktu Pengklasifikasian pada Metode Klasifikasi *Support Vector Machine*…….. V-38
Tabel V-49. Hasil Uji *Independent Sample T-Test* Perbandingan Waktu Pengklasifikasian pada Metode Klasifikasi *Support Vector Machine*…….. V-39
Tabel V-50. Mean Waktu Pengklasifikasian pada Metode Klasifikasi *Support Vector Machine*…….. V-40

DAFTAR GAMBAR

Halaman

Gambar II-2. *Decision bound* yang membagi dua kelas..... II-2

Gambar II-1. Pembentukan *hyperplane* pada *Support Vector Machine*..... II-6

Gambar II-3. Pembentukan vektor dokumen dengan *Bag-Of-Words*..... II-11

Gambar II-4. Arsitektur *Doc2Vec*..... II-13

Gambar II-5. Matriks *one-hot-encoded* dokumen pada indeks ke-2..... II-14

Gambar II-6. Matriks *input layer* dikalikan bobot dimensi $C \; X \; N$..... II-14

Gambar II-7. Matriks *hidden layer* dikalikan bobot dimensi $N \; X \; V$..... II-14

Gambar II-8. Matriks *output layer* dikalikan *k* jumlah kata target..... II-15

Gambar II-9. Matriks dikonversi ke dalam bentuk probabilitas *softmax*..... II-15

Gambar II-10. Nilai eror yang didapat untuk memperbaiki nilai bobot..... II-15

Gambar II-11. Kurva AUROC yang tidak saling *overlap*..... II-17

Gambar II-12. Kurva AUROC yang saling *overlap*..... II-17

Gambar III-1. Tahapan Pengujian Penelitian..... III-7

Gambar III-2. Penjadwalan untuk Tahap Menentukan Ruang Lingkup dan Unit Penelitian..... III-16

Gambar III-3. Penjadwalan untuk Tahap Menentukan Dasar Teori yang Berkaitan dengan Penelitian dan Menentukan Kriteria Pengujian..... III-17

Gambar III-4. Penjadwalan untuk Tahap Menentukan Alat yang Digunakan untuk Pelaksanaan Penelitian Fase Insepsi..... III-18

Gambar III-5. Penjadwalan untuk Tahap Menentukan Alat yang Digunakan untuk Pelaksanaan Penelitian Fase Elaborasi..... III-19

Gambar III-6. Penjadwalan untuk Tahap Menentukan Alat yang Digunakan untuk Pelaksanaan Penelitian Fase Konstruksi..... III-20

Gambar III-7. Penjadwalan untuk Tahap Menentukan Alat yang Digunakan untuk Pelaksanaan Penelitian Fase Transisi..... III-20

Gambar III-8. Penjadwalan untuk Tahap Melakukan Pengujian Penelitian, Analisis Hasil Pengujian Penelitian dan Membuat Kesimpulan..... III-21

Gambar IV-1. Diagram *Use Case*…….. IV-2

Gambar IV-2. *Sequence diagram* melakukan praproses dokumen..... IV-20

Gambar IV-3. *Sequence diagram* melakukan pembentukan VSM dengan teknik *Bag Of Words*…….. IV-21

Gambar IV-4. *Sequence diagram* melakukan pembentukan VSM dengan teknik *Doc2Vec*…….. IV-22

Gambar IV-5. *Sequence diagram* melakukan klasifikasi dengan *logistic regression* ……..…….. IV-23

Gambar IV-6. *Sequence diagram* melakukan klasifikasi dengan *multinomial naive bayes*……...……...…….. IV-24

Gambar IV-7. *Sequence diagram* melakukan klasifikasi dengan *support vector machine*……...……...…….. IV-25

Gambar IV-8. *Sequence diagram* melakukan evaluasi hasil pengklasifikasian dengan AUROC……...……...……...……... IV-25

Gambar IV-9. Rancangan Antar Muka Perangkat Lunak…...……..…… IV-26

Gambar IV-10. Rancangan Muat Dokumen……...…...……..…… IV-26

Gambar IV-11. Rancangan Fitur Pembentukan VSM …….……..…..... IV-27

Gambar IV-12. Rancangan Fitur Pengklasifikasian …….……..……..... IV-27

Gambar IV-13. Kotak/*field* yang menampilkan informasi dari proses tahapan-tahapan yang dilakukan ……..…….……..……..... IV-28

Gambar IV-14. Rancangan Fitur Evaluasi Hasil Pengklasifikasian…..... IV-29

Gambar IV-15. *Activity diagram* melakukan praproses dokumen …..... IV-30

Gambar IV-16. *Activity diagram* melakukan pembentukan VSM *Bag Of Words* …..... IV-30

Gambar IV-17. *Activity diagram* melakukan pembentukan VSM *Doc2Vec* …..... IV-31

Gambar IV-18. *Activity diagram* melakukan klasifikasi dengan *multinomial naive bayes*……...……...…….. IV-31

Gambar IV-19. *Activity diagram* melakukan klasifikasi dengan *logistic regression* ……..…….. IV-32

Gambar IV-20. *Activity diagram* melakukan klasifikasi dengan *support vector machine*……...……...…….. IV-33

Gambar IV-21. *Activity diagram* melakukan evaluasi hasil pengklasifikasian dengan AUROC……...……...……...……... IV-33

Gambar IV-22. Diagram Kelas……...…...……..…… IV-35

DAFTAR LAMPIRAN

Halaman

Lampiran 1. Kode Program……..... xxv

BAB I

PENDAHULUAN

1.1. Pendahuluan

Pada bab ini membahas latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian serta batasan masalah. Bab ini akan memberikan penjelasan umum mengenai keseluruhan penelitian.

Pendahuluan dimulai dengan penjelasan mengenai tantangan dan tujuan proses menemukan pengetahuan baru pada *text data mining*. Serta penelitian yang berkaitan dengan perbandingan metode-metode klasifikasi untuk analisis sentimen berdasarkan teknik pembentukan *vector space model* yang menjadi latar belakang dari penelitian ini.

1.2. Latar Belakang Masalah

Luasnya penggunaan internet mengakibatkan orang-orang telah mampu memberikan berbagai informasi kepada publik. Informasi ini sering menyertakan pendapat atau sentimen terhadap beberapa layanan atau produk. Dengan menganalisis pendapat individu yang terkait dengan suatu layanan atau produk, maka akan dapat didapatkan apakah mereka memiliki sifat positif terhadap layanan atau produk tersebut. Tugas utama dalam analisis sentimen ialah melakukan klasifikasi terhadap sentimen. Dalam melakukan klasifikasi sentimen tersebut dapat menggunakan metode-metode klasifikasi yang terdapat pada algoritma pembelajaran mesin. Metode-metode klasifikasi ini memiliki pendekatan

pengklasifikasian yang berbeda-beda. Diantaranya terdapat metode klasifikasi dengan pendekatan pada probabilitas kemunculan suatu atribut di dalam suatu kelas. Metode klasifikasi ini ialah *Multinomial Naive Bayes*. Kemudian metode klasifikasi dengan pendekatan pada jarak suatu atribut sehingga menghasilkan jarak pembatas yang memisahkan antar kelas. Metode klasifikasi ini antara lain *Support Vector Machine* dan *k-Nearest Neighbor.* Selanjutnya metode klasifikasi dengan pendekatan *binary decision* pada suatu atribut. Metode klasifikasi ini ialah *Logistic Regression.*

Algoritma pembelajaran mesin dalam melakukan pengklasifikasian membutuhkan masukan dengan representasi vektor atau masukan yang berbentuk numerik. Oleh karena itu, masukan yang berupa teks tersebut perlu ditransformasikan ke dalam bentuk vektor atau numerik. Teknik mentransformasi masukan teks ke dalam bentuk vektor ini disebut juga *vector space model* (VSM). Teknik ini merupakan salah satu bagian dari tahapan pengolahan data awal dan merupakan teknik mendasar dalam proses *text mining.* Teknik *vector space model* yang paling umum digunakan ialah representasi *Bag-Of-Words* (BOW). Representasi BOW sangat sederhana dan efisien dalam melakukan klasifikasi teks, namun representasi ini memperlakukan kata-kata sebagai unit yang independen (Rudkowsky, et al. 2018). Representasi BOW bekerja pada tingkatan kata, sehingga representasi ini akan mengabaikan urutan pada kata-kata dalam suatu teks. Dalam tingkatan dokumen, terdapat representasi *Doc2Vec* yang dapat memperoleh representasi numerik yang terdistribusi dari suatu frase, paragraf maupun dokumen.

Penelitian dilakukan oleh Barhoumi, et al. (2017) dengan menggunakan teknik VSM pada metode klasifikasi yang berbeda. Penelitian tersebut menerapkan representasi BOW dengan metode klasifikasi *multinomial Naive Bayes* dan representasi *Doc2Vec* dengan metode klasifikasi *Logistic Regression* pada analisis sentimen bahasa Arab. Hasil akurasi yang didapatkan representasi BOW dengan metode klasifikasi *multinomial Naive Bayes* menghasilkan akurasi sebesar 45%. Hasil akurasi yang didapatkan representasi *Doc2Vec* dengan metode klasifikasi *Logistic Regression* menghasilkan nilai akurasi sebesar 32,38%. Hasil akurasi yang didapatkan dari penelitian tersebut tidak dapat dijadikan perbandingan secara langsung dikarenakan teknik VSM diterapkan pada metode klasifikasi yang berbeda. Hasil yang didapatkan juga sesuai dengan kesulitan bahasa Arab dibandingkan dengan bahasa Inggris dikarenakan bahasa Arab kaya akan morfologi.

Selanjutnya penelitian yang dilakukan oleh Widyaningtyas, Adiwijaya dan Faraby (2018) dalam mengklasifikasi sentimen pada *review* film berbahasa Inggris menerapkan representasi *Doc2Vec* dengan metode *Support Vector Machine.* Data yang digunakan merupakan data yang telah diklasifikasi menjadi dua kelas (positif dan negatif) yang dikumpulkan dari *IMDb* (*Internet Movie Database).* Dari penelitian yang dilakukan memiliki nilai *F1-Measure* sebesar 54,1872%. Pada penelitian tersebut, komposisi dan jumlah dimensi yang ada pada data latih berpengaruh pada akurasi yang dihasilkan. Representasi *Doc2Vec* dapat menghasilkan representasi numerik dari tingkat dokumen. Representasi ini akan

bekerja dengan sangat baik dengan dokumen atau korpus yang besar (Lau dan Baldwin 2016) dan (Bilgin dan Senturk 2017).

Berdasarkan uraian tersebut, penelitian ini akan menguji perbandingan metode-metode klasifikasi untuk analisis sentimen berdasarkan teknik pembentukan *vector space model* agar didapatkan hasil pengklasifikasian yang efektif untuk proses analisis sentimen.

1.3. Rumusan Masalah

Rumusan masalah pada penelitian ini adalah apa perbedaan kinerja dari metode-metode klasifikasi untuk analisis sentimen berdasarkan teknik pembentukan *vector space model*? Untuk menjawab rumusan masalah tersebut, diuraikan beberapa pertanyaan penelitian sebagai berikut:

1. Apa perbedaan cara kerja pembentukan *vector space model* oleh *Bag-Of-Words* dan *Doc2Vec*?
2. Bagaimana peranan teknik *vector space model* terhadap kinerja beberapa teknik pengklasifikasian untuk analisis sentimen?

1.4. Tujuan Penelitian

Tujuan dilakukannya penelitian adalah sebagai berikut:

1. Mengetahui mekanisme pembentukan *vector space model* oleh *Bag-Of-Words* dan *Doc2Vec*
2. Dapat memilih teknik *vector space model* yang sesuai dengan teknik pengklasifikasian yang digunakan

1.5. Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini adalah sebagai berikut :

1. Dapat melakukan pengklasifikasian yang lebih efektif diantara pasangan teknik *vector space model* dan teknik pengklasifikasian yang digunakan
2. Dapat melakukan pengklasifikasian analisis sentimen yang lebih akurat

1.6. Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut :

1. Data yang digunakan adalah data ulasan film dalam bahasa Inggris, maka kaidah pada praproses diatur dalam tata bahasa Inggris. Data terdiri dari dua kelas, yaitu kelas positif dan negatif.
2. Metode-metode klasifikasi sentimen yang digunakan yaitu berdasarkan jenis-jenis pendekatannya. Pendekatan dengan *binary decision* menggunakan *Logistic Regression,* pendekatan dengan pemisah antar kelas menggunakan *Support Vector Machine* dan pendekatan dengan probabilitas kemunculan menggunakan *Multinomial Naive Bayes.*
3. Teknik *vector space model* yang digunakan adalah *Bag-Of-Words* dan *Doc2Vec.*
4. Ukuran kinerja hanya ditentukan oleh hasil akurasi dengan menggunakan teknik *Area Under the Receiver Operating Characteristics* (AUROC) dan waktu komputasi dalam melakukan pengklasifikasian.

1.7. Sistematika Penulisan

Penyusunan skripsi ini disusun dengan sistematika penulisan sebagai berikut :

BAB I. PENDAHULUAN

Pada bab ini diuraikan mengenai latar belakang, perumusan masalah, tujuan dan manfaat penelitian, batasan masalah atau ruang lingkup serta sistematika penulisan.

BAB II. KAJIAN LITERATUR

Pada bab ini membahas dasar-dasar teori yang digunakan dalam penelitian, seperti teknik pengklasifikasian untuk analisis sentimen, teknik *vector space model* serta beberapa kajian literatur mengenai penelitian lain yang relevan pada penelitian ini.

BAB III. METODOLOGI PENELITIAN

Pada bab ini membahas mengenai tahapan yang akan dilaksanakan pada penelitian ini. Masing-masing rencana tahapan penelitian dideskripsikan dengan rinci mengacu pada suatu kerangka kerja. Di akhir bab, berisi perancangan manajemen proyek pelaksanaan penelitian.

BAB IV. PENGEMBANGAN PERANGKAT LUNAK

Pada bab ini membahas mengenai analisis dan perancangan perangkat lunak yang akan digunakan sebagai alat penelitian. Dimulai dari pengumpulan dan analisa kebutuhan, rancangan dan konstruksi perangkat lunak serta pengujian untuk memastikan semua kebutuhan

pengembangan perangkat lunak sesuai dengan dengan kebutuhan. Penyusunan pada bab ini memiliki kerangka penulisan dengan fase-fase dan elemen-elemen pengembangan perangkat lunak bersifat berorientasi objek.

BAB V. HASIL DAN ANALISA PENELITIAN

Pada bab ini diuraikan hasil pengujian berdasarkan langkah-langkah yang telah direncanakan. Tabel hasil pengujian serta analisisnya disajikan sebagai basis dari kesimpulan yang akan diambil dalam penelitian ini.

BAB VI. KESIMPULAN DAN SARAN

Pada bab ini berisi kesimpulan dari semua uraian-uraian pada bab-bab sebelumnya dan juga saran-saran yang diharapkan berguna dalam melakukan pengklasifikasian dengan metode-metode klasifikasi untuk analisis sentimen berdasarkan teknik pembentukan *vector space model*.

1.8. Kesimpulan

Dari bab ini telah dibahas latar belakang masalah penelitian ini dalam melakukan pengklasifikasian sentimen dengan metode-metode klasifikasi berdasarkan teknik pembentukan *vector space model*. Karena itu, penelitian ini akan melakukan perbandingan metode-metode klasifikasi untuk analisis sentimen berdasarkan teknik pembentukan *vector space model* agar didapatkan hasil pengklasifikasian yang efektif untuk proses analisis sentimen.

Barhoumi, Amira, Yannick Esteve, Chafik Aloulou, and Lamia Hadrich Belguith. 2017. "Document embeddings for Arabic Sentiment Analysis."

Bilgin, Metin, and Izzet Fatih Senturk. 2017. "Sentiment Analysis on Twitter data with Semi-Supervised Doc2Vec."

Goldberg, Yoav. 2017. *Neural Network Methods in Natural Language Processing* . Morgan & Claypool Publishers.

Hsu, Chih-wei, Chih-chung Chang, and Chih-Jen Lin. 2016. "A Practical Guide to Support Vector Classification."

Joachims, Thorsten. 1999. "Text Categorization with Support Vector Machine: Learning with Many Relevant Features."

Lau, Jey Han, and Timothy Baldwin. 2016. "An Empirical Evaluation of Doc2Vec with Practical Insight into Document Embedding Generation ."

Le, Quoc, and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents."

Maas, Andrew L, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. "Learning Word Vectors for Sentiment Analysis." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142-150. Portland, Oregon: Association for Computational Linguistics. http://www.aclweb.org/anthology/P11-1015.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jefrey Dean. 2013. "Distributed Representations of Words and Phrases and their Compositionality."

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space."

Rudkowsky, Elena, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair. 2018. "More than Bag of Words: Sentiment Analysis with Word Embeddings."

Saraswati. 2011. "Text Mining dengan Metode Naive Bayes Classifier dan Support Vector Machine untuk Sentiment Analysis."

Sherin, Shamseera, and Sreekanth. 2016. "Word Vectors in Sentiment Analysis." *International Journal of Current Trends in Engineering & Research.*

Song, J, B Lee, K T Kim, and S Kim. 2017. "A Novel Classification approach based on Naive Bayes for Twitter Sentiment Analysis." *KSII Transaction on Internet and Information Systems.*

Widyaningtyas, Winda Christina, Adiwijaya, and Said Al Faraby. 2018. "Klasifikasi Sentiment Analysis pada Review Film Berbahasa Inggris dengan Menggunakan Metode Doc2Vec dan Support Vector Machine."

Zirpe, Sayali, and Bela Joglekar. 2017. "Polarity Shift Detection Approaches in Sentiment Analysis: A survey." *International Conference on Inventive Systems and Control.*