



Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk menggubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

**ANALISIS SENTIMEN MEDIA *TWITTER*
MENGUNAKAN *NAIVE BAYESIAN*
STUDI KASUS: PREDIKSI *POLLING* KANDIDAT
PRESIDEN AMERIKA SERIKAT 2016**

SKRIPSI



Diajukan Guna Memenuhi Persyaratan Memperoleh Gelar
Sarjana Komputer (S.Kom.)

Elvyna Tunggawan

12110310024

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNOLOGI INFORMASI DAN KOMUNIKASI
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG

2016

PERNYATAAN TIDAK MELAKUKAN PLAGIAT

Dengan ini, saya,

Nama : Elvyna Tunggawan

NIM : 12110310024

Program Studi : Sistem Informasi

Menyatakan bahwa skripsi ini merupakan hasil ide yang saya buat dan dikerjakan sendiri, serta bukan merupakan hasil pekerjaan atau penelitian yang dilakukan oleh orang, peneliti, organisasi, dan / atau perusahaan lain yang kemudian saya ambil atau tiru. Semua data yang saya ambil dari buku atau karya tulis orang atau lembaga lainnya seluruhnya saya cantumkan pada bagian Daftar Pustaka.

Apabila ditemukan bahwa adanya kecurangan atau kutipan yang saya lakukan di dalam skripsi ini, saya bersedia untuk dinyatakan GAGAL atau TIDAK LULUS untuk mata kuliah skripsi yang saya tempuh ini.

Tangerang, 22 Juni 2016

Elvyna Tunggawan

PERSETUJUAN LAPORAN SKRIPSI

ANALISIS SENTIMEN MEDIA *TWITTER* MENGGUNAKAN *NAIVE BAYESIAN*
STUDI KASUS: PREDIKSI *POLLING* KANDIDAT PRESIDEN AMERIKA
SERIKAT 2016

Oleh :

Nama : Elvyna Tunggawan
NIM : 12110310024
Fakultas : Teknologi Informasi dan Komunikasi
Program Studi : Sistem Informasi

Telah disetujui untuk diujikan pada Sidang Skripsi

Tangerang, 24 Mei 2016

Ketua Program Studi

Dosen Pembimbing

(Wira Mungana, S.Si., M.Sc.)

(Yustinus Eko Soelistio, S.Kom., M.M.)

PENGESAHAN LAPORAN SKRIPSI

**ANALISIS SENTIMEN MEDIA *TWITTER* MENGGUNAKAN *NAIVE BAYESIAN*
STUDI KASUS: PREDIKSI *POLLING* KANDIDAT PRESIDEN AMERIKA SERIKAT 2016**

Skripsi yang dibuat dengan judul

**”Analisis Sentimen Media *Twitter* Menggunakan *Naive Bayesian*
Studi Kasus: Prediksi *Polling* Kandidat Presiden Amerika Serikat 2016”**

oleh

Elvyna Tunggawan - 12110310024

Telah diujikan pada hari Selasa, tanggal 14 Juni 2016

Pukul 14.00 s.d. 15.30 dan dinyatakan lulus

dengan susunan penguji sebagai berikut

Pembimbing

Penguji

(Yustinus Eko Soelistio, S.Kom., M.M.)

(Ir. Raymond Sunardi Oetama, MCIS)

Ketua Sidang

(Friska Natalia Ferdinand, Ph.D.)

Disahkan oleh

Ketua Program Studi Sistem Informasi

(Wira Mungguna, S.Si., M.Sc.)

KATA PENGANTAR

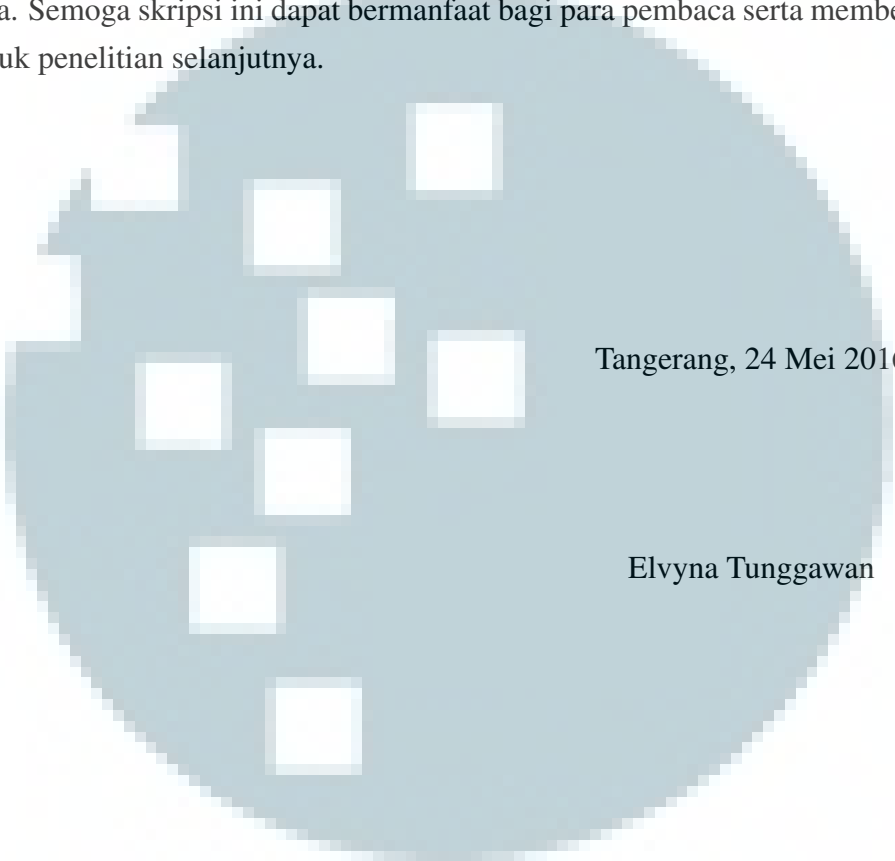
Puji syukur penulis ucapkan kepada Tuhan Yang Maha Esa karena berkat kuasa dan penyertaan-Nya penulis dapat memulai dan menyelesaikan skripsi dengan judul "Analisis Sentimen Media *Twitter* Menggunakan *Naive Bayesian* (Studi Kasus: Prediksi *Polling* Kandidat Presiden Amerika Serikat 2016)" tepat dengan batas waktu yang telah ditentukan. Skripsi ini dibuat sebagai salah satu syarat kelulusan penulis dalam Program Strata 1, Program Studi Sistem Informasi, Fakultas Teknologi Informasi dan Komunikasi, Universitas Multimedia Nusantara.

Penulis ingin menyampaikan terima kasih kepada pihak-pihak yang turut membantu penulis dalam pelaksanaan skripsi ini, yaitu

1. Keluarga penulis, khususnya Alm. Mama, atas didikan, dukungan dan motivasi untuk terus berkarya,
2. Bapak Yustinus Eko Soelistio selaku pembimbing penulisan skripsi, atas bimbingan dan masukan yang diberikan selama penelitian berlangsung,
3. Alm. Bapak Sigit Surendra, atas masukan dalam menentukan topik penelitian,
4. Bapak Wira Mungguna dan Bapak Johan Setiawan selaku Ketua dan Sekretaris Program Studi Sistem Informasi,
5. Novita Belinda Wunarso dan Gina Akmalia selaku rekan seperjuangan penulis selama studi,
6. Roderick Markus Irawan, Alexander Samuel, Daniel Gunawan, Kenny Supangat, Yuniar Berlian Ananda Panjaitan, Milka Veronika, Melisa Mulyasari, dan Johanna Tjokrosasmito atas semangat dan dukungan agar penulis segera menyelesaikan penelitian,
7. Prasasya Dira Pinasthika, Kelvin Chandra, Eric Bagus, Galih Prakoso dan teman-teman narasumber lainnya atas bantuan dalam menyelesaikan penelitian ini, dan
8. teman-teman mahasiswa Sistem Informasi, khususnya angkatan 2012, atas

pengalaman perkuliahan yang tidak terlupakan.

Penulis juga mengucapkan terima kasih kepada pihak-pihak lain yang tidak dapat penulis sebutkan satu per satu, yang turut membantu penulis dengan berbagai cara. Semoga skripsi ini dapat bermanfaat bagi para pembaca serta memberikan ide untuk penelitian selanjutnya.



Tangerang, 24 Mei 2016

Elvyna Tunggawan

UMN

ABSTRAK

Nama : Elvyna Tunggowan

NIM : 12110310024

Amerika Serikat merupakan salah satu negara yang paling berpengaruh di dunia. Kebijakan yang diambil ditentukan oleh pemerintah dan bergantung pada sudut pandang presidennya. Oleh karena itu, penting untuk mengetahui siapa yang berpeluang menjadi Presiden Amerika Serikat berikutnya. Penelitian sebelumnya telah melakukan prediksi hasil Pemilihan Presiden Amerika Serikat 2008 dan 2012 berdasarkan data *Twitter* menggunakan metode *preprocessing* data yang relatif kompleks. Pada penelitian ini, penulis memanfaatkan data *Twitter* yang melalui tahap *preprocessing* lebih sederhana untuk memprediksi hasil *polling* kandidat Presiden Amerika Serikat 2016.

Prediksi dilakukan dengan menggunakan model *Naive Bayesian* dan dilatih dengan 33.708 *tweet* yang sudah melalui tahap *preprocessing* sederhana serta diberi *label* secara manual. Metode *preprocessing* dilakukan sesederhana mungkin agar tidak mengubah makna *tweet*. Model berhasil mencapai akurasi sebesar 95,8% dan memprediksi Bernie Sanders serta Ted Cruz sebagai kandidat Partai Demokrat dan Republik yang unggul. Namun, model hanya memiliki akurasi 26,7% ketika dibandingkan dengan *RealClearPolitics.com*.

Kata kunci: analisis sentimen, *text mining*, *Twitter*, *Naive Bayes*

ABSTRACT

Name : Elvyna Tunggowan

Student Number : 12110310024

United States is one of the most influential countries in the world. U.S. policies are set by the government and are dependent on the President's point of view. Hence, it is important to know which candidate is most likely to be the next U.S. President. Previous researches have predicted the outcomes of U.S. Presidential Election in 2008 and 2012 based on *Twitter*. Most of the researches used complex data preprocessing methods to filter the tweets. Therefore, this research is focused on predicting 2016 U.S. Presidential Election poll results for all candidates based on *Twitter* data using relatively simpler data preprocessing method.

We build a Naive Bayesian model for each candidate using 33,708 manually labeled tweets. The tweets have passed a simple preprocessing stage, which does not alter their meanings. The model achieves 95,8% accuracy and predicts Bernie Sanders and Ted Cruz as the nominees of Democratic and Republican Party respectively. However, the prediction only achieves 26.7% accuracy when it is compared to *RealClearPolitics.com*.

Keywords: sentiment analysis, text mining, *Naive Bayes*, *Twitter*

DAFTAR ISI

PERNYATAAN TIDAK MELAKUKAN PLAGIAT	ii
PERSETUJUAN LAPORAN SKRIPSI	iii
PENGESAHAN LAPORAN SKRIPSI	iv
KATA PENGANTAR	v
ABSTRAK	vii
ABSTRACT	viii
DAFTAR ISI	ix
DAFTAR GAMBAR	xiii
DAFTAR TABEL	xv
1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	4
1.3 Batasan Masalah	4
1.4 Tujuan dan Manfaat Penelitian	5
1.4.1 Tujuan Penelitian	5
1.4.2 Manfaat Penelitian	5
1.5 Rencana Kegiatan	5
1.6 Sistematika Penulisan	6
1.7 Sumber Dana	7

2	LANDASAN TEORI	8
2.1	<i>Text Mining</i>	8
2.2	Analisis Sentimen	9
2.3	<i>Naive Bayes Classifier</i>	11
2.4	<i>Cross Validation</i>	13
2.5	<i>Confusion Matrix</i>	14
2.6	<i>F₁-score</i>	16
2.7	<i>Matthews Correlation Coefficient</i>	16
2.8	<i>Majority Rule</i>	18
2.9	<i>Tweepy</i>	18
2.10	<i>Scikit-Learn</i>	19
2.11	<i>Natural Language Toolkit (NLTK)</i>	19
2.12	Penelitian Terdahulu	20
2.12.1	<i>From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series</i>	20
2.12.2	<i>Predicting US Primary Elections with Twitter</i>	23
2.12.3	<i>The Predictive Power of Social Media: On the Predictability of U.S. Presidential Elections using Twitter</i>	26
2.12.4	<i>A System for Real-Time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle</i>	29
2.12.5	<i>Analyzing Twitter Sentiment of the 2016 Presidential Candidates</i>	32
3	METODOLOGI PENELITIAN	36
3.1	Gambaran Umum Objek Penelitian	36
3.2	Metode Penelitian	37
3.2.1	Pengumpulan Data	38

3.2.2	<i>Data Preprocessing</i>	40
3.2.3	<i>Data Labelling</i>	42
3.2.4	Pelatihan Model	44
3.2.5	Pengujian Akurasi Model	45
3.2.6	Pengujian Akurasi Prediksi	45
4	ANALISIS DAN PEMBAHASAN	47
4.1	Pengumpulan Data	47
4.2	<i>Data Preprocessing</i>	49
4.3	<i>Data Labelling</i>	52
4.4	Pelatihan Model	55
4.5	Pengujian Akurasi Model	55
4.6	Pengujian Akurasi Prediksi	57
4.6.1	Hasil Prediksi Model	57
4.6.2	Perbandingan Prediksi Model dengan Hasil <i>Polling</i>	59
4.7	Diskusi	62
4.7.1	Jumlah <i>Data Training</i>	63
4.7.2	Metode <i>Preprocessing</i>	66
4.7.3	Hasil <i>Labelling</i>	68
4.7.4	Hubungan Jumlah <i>Tweet</i> Positif dengan Hasil <i>Polling</i>	71
4.7.5	Hubungan Popularitas Kandidat dengan Hasil <i>Polling</i>	73
4.7.6	<i>Twitter</i>	75
5	SIMPULAN DAN SARAN	77
5.1	Simpulan	77
5.2	Saran	78

DAFTAR PUSTAKA



DAFTAR GAMBAR

2.1	<i>Confusion Matrix</i> pada <i>Classifier</i> Biner (Han & Kamber, 2006)	15
2.2	Arsitektur Sistem (Wang, Can, Kazemzadeh, Bar, & Narayanan, 2012)	30
2.3	Tampilan <i>Dashboard</i> (Wang et al., 2012)	31
2.4	Daftar <i>Emoji</i> pada Masing-masing Kelas (Chin, Zappone, & Zhao, 2015)	33
2.5	Contoh Pemetaan Sentimen (Chin et al., 2015)	35
3.1	Alur Penelitian	38
3.2	Struktur JSON <i>Tweet</i>	39
3.3	Tahap <i>Preprocessing</i>	41
3.4	Alur Proses <i>Labelling</i>	43
3.5	Menentukan Objek <i>Tweet</i>	43
3.6	Tampilan <i>Tweet</i> Setelah Diberi <i>Label</i>	44
3.7	Alur Penentuan Simpulan Sentimen	44
4.1	Distribusi Jumlah <i>Tweet</i> ($\mu = 37.126,4; \sigma = 27.823,82$)	47
4.2	Atribut JSON <i>Tweet</i>	48
4.3	Perbandingan Distribusi Jumlah <i>Tweet</i>	51
4.4	Persentase Selisih <i>Tweet</i> Sebelum dan Sesudah <i>Preprocessing</i> ($\mu = 40,87\%; \sigma = 4,98\%$)	51
4.5	Pengelompokan <i>Tweet</i>	52
4.6	Simpulan Sentimen <i>Tweet</i>	52
4.7	Sebaran Sentimen Hasil <i>Labelling</i>	53
4.8	Total Sentimen per Kandidat	54
4.9	Hasil <i>Polling</i> Partai Demokrat pada 9 Februari 2016	60
4.10	Hasil <i>Polling</i> Partai Republik pada 9 Februari 2016	60

4.11	Perkembangan Akurasi Model Berdasarkan Jumlah Data <i>Training</i>	64
4.12	Tingkat Ketidaksesuaian Jumlah <i>Tweet</i> Positif dengan Hasil <i>Polling</i> per Periode ($\mu = 0,972$)	71
4.13	Sebaran Tingkat Ketidaksesuaian Jumlah <i>Tweet</i> Positif dengan Hasil <i>Polling</i>	72
4.14	Sebaran Tingkat Ketepatan Jumlah <i>Tweet</i> Positif dengan Hasil <i>Polling</i> ($\mu = 39\%$)	72
4.15	Tingkat Ketidaksesuaian Popularitas dengan Hasil <i>Polling</i> per Periode ($\mu = 0,617$)	73
4.16	Sebaran Tingkat Ketidaksesuaian Popularitas dengan Hasil <i>Polling</i>	74
4.17	Sebaran Tingkat Ketepatan Popularitas dengan Hasil <i>Polling</i> ($\mu = 46,7\%$)	74



DAFTAR TABEL

1.1	Rencana Penelitian	5
2.1	Modul NLTK	20
3.1	Syarat Penentuan <i>Label Tweet</i>	42
4.1	Contoh <i>Tweet</i> Sebelum dan Setelah <i>Preprocessing</i>	50
4.2	Pembagian <i>Tweet</i> untuk <i>Labelling</i>	52
4.3	Jumlah Data <i>Training Model</i>	55
4.4	Hasil Pengujian Model	56
4.5	Prediksi pada Kandidat Partai Demokrat	58
4.6	Prediksi pada Kandidat Partai Republik	58
4.7	Perbandingan Urutan Prediksi dan <i>Polling</i> pada Kandidat Partai Demokrat	59
4.8	Perbandingan Urutan Prediksi dan <i>Polling</i> pada Kandidat Partai Republik	61
4.9	Jumlah <i>Tweet</i> per Kandidat Setiap <i>n Tweet</i> Pertama	64
4.10	Hasil Prediksi <i>Polling</i> 19 Januari dan 9 Februari 2016	65
4.11	Perbandingan Hasil <i>Preprocessing</i>	67
4.12	Perbandingan <i>Label</i> pada <i>Tweet</i> dengan Isi yang Sama	69
4.13	Sampel <i>Tweet</i> "Tidak Jelas" yang Disingkirkan	70
4.14	Hasil Analisis Diskusi	75