



**KLASIFIKASI TOPIK PADA WEB BERITA *ONLINE* DENGAN
PENDEKATAN TEKS *MINING*.**

SKRIPSI

M. Edo Noprian

161410264

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS ILMU KOPMUTER

UNIVERSITA BINADARMA

PALEMBANG

2020



**KLASIFIKASI TOPIK PADA WEB BERITA *ONLINE* DENGAN
PENDEKATAN TEKS *MINING*.**

M. Edo Noprian

161410264

**Skripsi ini diajukan sebagai syarat memperoleh gelar Sarjana
Komputer**

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS ILMU KOPMUTER

UNIVERSITA BINADARMA

PALEMBANG

2020

HALAMAN PENGESAHAN

**KLASIFIKASI TOPIK WEB PADA BERITA ONLINE DENGAN
PENDEKATAN TEKS MINING**

M. EDO NOPRIAN

161410264

**Telah diterima sebagai salah satu syarat untuk memperoleh gelar
Sarjana Komputer pada Program Studi Sistem Informasi**

**Palembang, 08 September 2020
Fakultas Ilmu Komputer
Universitas Bina Darma
Dekan,**

Pembimbing



Dr. Edi Surya Negara, M.Kom.



Dedy Syamsuar, S.Kom., M.I.T., Ph.D.

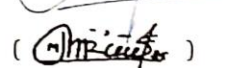
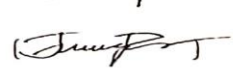
HALAMAN PERSETUJUAN

Skripsi Berjudul "Klasifikasi Topik pada Web Berita *Online* dengan Pendekatan Teks *Mining*." Oleh M. Edo Noprian (161410264) telah dipertahankan didepan komisi penguji pada Selasa tanggal 08 September 2020.

Komisi Penguji

1. Ketua : Dr. Edi Surya Negara, M.Kom.
2. Anggota : Maria Ulfa, M.Kom.
3. Anggota : Firammon Syakti, M.M., M. Kom.



()
()
()

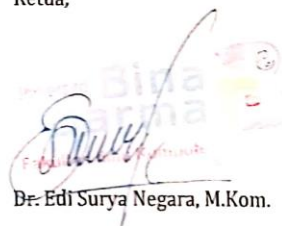
Mengetahui,

Program Studi Sistem Informasi

Fakultas Ilmu Komputer

Universitas Bina Darma

Ketua,



Dr. Edi Surya Negara, M.Kom.

SURAT PERNYATAAN

Saya yang bertanda tangan dibawah ini :

Nama : M. Edo Noprian

NIM : 161410264

Dengan ini menyatakan bahwa :

1. Karya tulis saya (Skripsi) adalah asli dan belum pernah diajukan untuk mendapatkan gelar akademik (Sarjana) di Universitas Bina Darma ataupun perguruan tinggi lainnya ;
2. Karya tulis ini murni gagasan, rumusan dan penelitian saya dengan arahan dari tim pembimbing ;
3. Di dalam karya tulis ini tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan jelas dikutip dengan mencantumkan nama pengarang dan memasukkan ke dalam daftar rujukan ;
4. Saya bersedia tugas skripsi, di cek keasliannya menggunakan plagiarism checker serta di unggah ke internet, sehingga dapat diakses secara daring ;
5. Surat pernyataan ini saya tulis dengan sungguh-sungguh dan apabila terbukti melakukan penyimpangan atau ketidakbenaran dalam pernyataan ini maka saya bersedia menerima sanksi dengan peraturan dan undang-undang yang berlaku ;

Demikian surat pernyataan ini saya buat agar dapat dipergunakan sebagaimana mestinya.

Pelambang, 08 September 2020

Yang membuat pernyataan,

A yellow postage stamp with the text "NETRAL TEMPEL" at the top, a serial number "484AAHF682641374", the value "6000", and "ENAM RIBU RUPIAH" at the bottom. A signature is written over the stamp.

M. Edo Noprian

161410264

ABSTRAK

Pesatnya perkembangan dalam dunia informasi digital menyebabkan peningkatan volume informasi yang berbentuk teks seperti dokumen berita. Di sosial media dokumen berita sangat banyak di unggah dalam rentan waktu yang begitu cepat dan salah satunya adalah *Twitter*. *Twitter* adalah layanan sosial media yang sudah sangat banyak melayani pengguna sehingga menjadikannya sebagai salah satu sosial media yang memiliki data yang sangat besar. Dari data yang sangat besar tersebut dapat dimanfaatkan sebagai sumber dokumen berita untuk web berita *online*. Akan tetapi dengan banyaknya topik yang di ekstraksi pada data *Twitter* membuat data yang masuk memiliki beragam topik yang menyebabkan kesulitan dalam mengidentifikasi topik dari kumpulan data yang diambil dan akan membutuhkan waktu yang banyak jika harus dilakukan secara manual oleh manusia. Sedangkan, data tersebut berpotensi dibutuhkan untuk memberikan informasi secepat mungkin.

Dengan dilakukannya penelitian ini, bertujuan untuk mengklasifikasikan topik pada data yang diambil dari *Twitter* secara otomatis sehingga dapat membuat klasifikasi pada dokumen berita yang diambil, dapat lebih efektif dan efisien dan tidak membutuhkan waktu sebanyak dilakukan manual oleh manusia. Penelitian dilakukan dengan metode *Latent Dirichlet Allocation (LDA)*. Dokumen berita yang akan di klasifikasi adalah dokumen berita bahasa Indonesia dan akan di klasifikasikan kedalam topik yang akan ditentukan. Eksperimen pemodelan topik dengan metode LDA menyimpulkan bahwa jumlah topik yang dibentuk dari 9094 data tweet adalah 10 topik. Hasil eksperimen ini *Latent Dirichlet Allocation* dapat digunakan untuk *text mining*, tetapi data tweet yang dihasilkan belum layak untuk dijadikan informasi dalam pengambilan topik berita. Karena hasil dari metode ini masih membutuhkan peninjauan kembali oleh manusia untuk mengetahui fakta dari *topic* yang dihasilkan.

Kata Kunci : Klasifikasi Dokumen Berita, *Latent Dirichlet Allocation (LDA)*

ABSTRACT

The rapid development in the world of digital information has led to an increase in the volume of information in the form of *text* such as news documents. On social media, news documents are very much uploaded in a very vulnerable time and one of them is Twitter. Twitter is a social media service that has served very many users, making it one of the social media that has huge data. From this very large data, it can be used as a *source* of news documents for *online* news websites. However, with the many topics extracted from the Twitter data, the incoming data has a variety of topics which causes difficulty in identifying topics from the data collection taken and will require a lot of time if it has to be done manually by humans. Meanwhile, the data is potentially needed to provide information as quickly as possible.

By doing this research, it aims to classify topics on data taken from Twitter automatically so that it can classify the news documents that are taken, can be more effective and efficient and do not require as much time as done manually by humans. The research was conducted using the Latent Dirichlet Allocation (LDA) method. News documents to be classified are Indonesian news documents and will be classified into topics to be determined. The topic modeling experiment with the LDA method concluded that the number of topics that were formed from 9094 tweet data was 10 topics. The results of this experiment, Latent Dirichlet Allocation, can be used for text mining, but the resulting tweet data is not suitable for information in taking news topics. Because the results of this method still require human review to find out the facts of the resulting topic.

Keywords: News Document Classification, *Latent Dirichlet Allocation (LDA)*

KATA PENGANTAR



Puji syukur kehadiran Tuhan Yang Maha Esa karena berkat rahmat dan karunia-nya skripsi yang berjudul “**Klasifikasi Topik Pada Web Berita Online Dengan Pendekatan Teks Mining**” dapat diselesaikan dengan baik untuk memenuhi salah satu syarat mendapatkan gelar Sarjana Komputer di Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Bina Darma.

Dalam penulisan skripsi ini, tentunya masih jauh dari sempurna. Hal ini dikarenakan keterbatasannya pengetahuan yang dimiliki. Oleh karena itu dalam rangka melengkapi kesempurnaan dari penulisan skripsi ini diharapkan adanya saran dan kritik yang diberikan bersifat membangun.

Pada kesempatan yang baik ini, tak lupa penulis menghaturkan terima kasih kepada semua pihak yang telah memberikan bimbingan, pengarahan, nasehat dan pemikiran dalam penulisan skripsi penelitian ini, terutama kepada :

1. Dr. Sunda Ariana, M.Pd., M.M. selaku Rektor Universitas Bina Darma Palembang.
2. Dedy Syamsuar, Ph.D. selaku Dekan Fakultas Ilmu Komputer.
3. Dr. Edi Surya Negara, M.Kom. selaku Ketua Program Studi Sistem Informasi, sekaligus pembimbing dalam menyelesaikan skripsi ini.
4. Kepada Ibu Maria Ulfa, M.Kom dan Firamon Syakti, M.M., M. Kom. sebagai penguji.
5. Kepada seluruh dosen dan mahasiswa Universitas Bina Darma yang telah membantu atas terlaksananya skripsi tersebut.
6. Kepada orang tua yang selalu memberikan semangat dan do’a sehingga dapat menyelesaikan skripsi ini.
7. Kepada kedua adik-adikku “**Syahrul**” dan “**Wulan**” yang selalu memberikan dukungan.

8. Kepada teman satu bimbingan yaitu Fajar, Tiara, Aldo, Fitra, Romadon, Siska, Yaya.
9. Kepada **Fajar** yang tida henti memberikan semangat.
10. Kepada Rekan-rekan kepengurusan HIMSIF Universitas BinaDarma.

Palembang, Agustus 2020

Penulis

DAFTAR ISI

HALAMAN PENGESAHAN.....	i
HALAMAN PERSETUJUAN.....	ii
SURAT PERNYATAAN	iii
ABSTRAK.....	iv
ABSTRACT	v
KATA PENGANTAR	vi
DAFTAR ISI.....	viii
DATAR GAMBAR	xii
DAFTAR <i>SOURCE CODE</i>	xiv
BAB I	1
PENDAHULUAN.....	1
1.1. Latar Belakang.....	1
1.2. Identifikasi Masalah	4
1.2.1. Web berita yang terkendala dengan banyaknya waktu digunakan untuk klasifikasi topik berita.....	4
1.2.2. Belum adanya klasifikasi topik otomatis pada web berita <i>online</i>	4
1.3. Rumusan Masalah	4
1.3.1. Bagaimana mengembangkan Aplikasi pengumpulan pada data <i>Twitter</i>	4
1.3.2. Bagaimana melakukan klasifikasi topik berita pada web berita <i>online</i> menggunakan metode <i>Latent Dirichlet Allocation (LDA)</i>	4
1.4. Batasan Masalah	4
1.5. Manfaat Penelitian	5

1.6. Tujuan Penelitian	5
1.6.1. Untuk mengembangkan aplikasi pengumpulan dokumen berita data Twitter untuk berita pada web berita <i>online</i>	5
1.6.2. Mengklasifikasi topik pada dokumen berita untuk web berita <i>online</i> menggunakan metode <i>Latent Dirichlet Allocation (LDA)</i>	5
BAB II	6
Tinjauan Pustaka.....	6
2.1. Landasan Teori	6
2.1.1. Media Sosial.....	6
2.1.2. Twitter.....	9
2.1.3. Web Berita	10
2.1.4. <i>Text preprocessing</i>	11
2.1.5. <i>Stemming</i>	11
2.1.6. <i>Stopword</i>	12
2.1.7. Bahasa Pemrograman Python.....	12
2.1.8. <i>Text mining</i>	12
2.1.9. <i>Topic Modelling</i>	16
2.1.10. <i>Latent Dirichlet Allocation</i>	16
2.1.11. <i>Wordcloud</i>	19
2.2. Penelitian Sebelumnya	19
BAB III	22
METODOLOGI PENELITIAN.....	22
3.1. Metode Penelitian.....	22
3.1.1. Waktu dan Tempat	22
3.1.2. Alat dan Bahan	22
3.1.3. Tahapan Penelitian.....	23

3.1.4.	Mempersiapkan Data.....	24
3.1.5.	Topik modeling dengan <i>Latent Dirichlet Allocation</i>	25
3.1.6.	Pra-pemrosesan <i>Corpus</i>	25
3.1.7.	Memodelkan topik.....	26
3.1.8.	<i>Latent Dirichlet Allocation</i>	27
BAB IV.....		28
HASIL DAN PEMBAHASAN.....		28
4.1. Implementasi.....		28
4.1.1.	Implementasi <i>Crawling</i>	28
4.1.2.	Implementasi pra-pemrosesan <i>corpus</i>	36
4.1.3.	Implementasi <i>Lowercase</i>	38
4.1.4.	<i>Tokenization</i>	39
4.1.5.	<i>Stopwords</i>	42
4.1.6.	<i>Stemming</i>	43
4.1.7.	Hapus duplikat dokumen.....	48
4.1.8.	Memodelkan <i>topic</i>	49
4.1.9.	Implementasi <i>Latent Dirichlet Allocation</i>	50
4.1.10.	Visualisasi	83
4.2. Pembahasan Uji Coba.....		89
4.3. Pembahasan Hasil.....		90
BAB V.....		91
KESIMPULAN DAN SARAN.....		91
5.1. Kesimpulan.....		91
5.2. Saran		91
Daftar Pustaka.....		92

LAMPIRAN	96
1. <i>Crawlingdata.py</i>	96
2. <i>Fungsiprosesing.py</i>	97
3. <i>Lowercase.py</i>	99
4. <i>Tokenization.py</i>	99
5. <i>Stopword.py</i>	100
6. <i>Stemming.py</i>	101
7. <i>Hapus duplikat.py</i>	104
8. <i>Membentuk model.py</i>	104
9. <i>JalankanLDA.py</i>	104
10. <i>Visualwordcloud.py</i>	105

DATAR GAMBAR

Gambar 1 <i>Latent Dirichlet Allocation (LDA)</i> menurut Blei	16
Gambar 2 Visualisasi Topic modelling dengan Metode <i>LDA</i> [Lau et al., 2012].	18
Gambar 3 Tahapan penelitian	24
Gambar 4 Tahap Topik <i>modeling dengan Latent Dirichlet Allocation (LDA)</i>	25
Gambar 5 Sub-aktivitas dari tahap pra-pemrosesan corpus (I MADE KUSNANTA BRAMANTYA PUTRA, 2017)	26
Gambar 6 Tampilan <i>output</i> dari <i>source code crawling</i> data	31
Gambar 7 bagian kecil hasil <i>crawling</i> dalam bentuk <i>json</i>	33
Gambar 8 proses rubah <i>json</i> ke <i>cmd</i>	34
Gambar 9 Hasil merubah <i>file json</i> ke <i>csv</i>	35
Gambar 10 Hasil evaluasi.....	36
Gambar 11 Sebelum Lowercase	39
Gambar 12 Sesudah Lower.....	39
Gambar 13 Hasil <i>Tokenization</i>	42
Gambar 14 Sesudah <i>Stopwords</i>	43
Gambar 15 Hasil dari <i>stemming</i>	47
Gambar 16 hasil dari <i>source code</i> seleksi data.....	48
Gambar 17 Hasil setelah menghapus duplikat data	49
Gambar 18 hasil dari pemanggilan topik.....	52
Gambar 19 Topik 1	84
Gambar 20 Topik 2	85
Gambar 21 Topik 3	85
Gambar 22 Topik 4	86
Gambar 23 Topik 5	86

Gambar 24 Topik 6	87
Gambar 25 Topik 7	87
Gambar 26 Topik 8	88
Gambar 27 Topik 9	88
Gambar 28 Topik 10	89
Gambar 29 hasil <i>Latent Dirichlet Allocaion</i>	90

DAFTAR SOURCE CODE

<i>Source code 1</i> Pengaturan hak akses twitter API.....	29
<i>Source code 2 Source code Crwling</i>	31
<i>Source code 3 source code</i> membuat fungsi pra-prmrosesan corpus.....	37
<i>Source code 4</i> Perintah membentuk <i>text</i> Lowercase	38
<i>Source code 5 source code Tokenization</i>	40
<i>Source code 6 Stopwords</i>	41
<i>Source code 7 stemming</i> menggunakan sastrawi	43
<i>Source code 8</i> Cleaning ulang.....	45
<i>Source code 9</i> menghilangkan data lain.....	47
<i>Source code 10</i> menghapus duplikat data.....	47
<i>Source code 11</i> perintah memodelkan topik	48
<i>Source code 12</i> Menentukan jumlah topik.....	49
<i>Source code 13 source code</i> Pembentukan topik.....	50
<i>Source code 14</i> pemanggilan topik	51
<i>Source code 15</i> Membuat fungsi wordcoud	52
<i>Source code 16</i> Manampilkan hasil dalam bentuk <i>wordcloud</i>	52