INISIASI *NATURAL LANGUAGE PROCESSING (NLP)* DAN KLASIFIKASI PERTANYAAN UNTUK PEMBUATAN "STATBOT" (STATISTIKA *CHATTER BOT*)

(Studi Kasus:Informasi Seputar Program Studi Statistika Universitas Islam Indonesia)

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana
Program Studi Statistika



Disusun oleh :

Nita Tri Anggraini
15 611 061

JURUSAN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS ISLAM INDONESIA
Yogyakarta
2019

INISIASI *NATURAL LANGUAGE PROCESSING (NLP)* DAN KLASIFIKASI PERTANYAAN UNTUK PEMBUATAN "STATBOT" (STATISTIKA *CHATTER BOT*)

(Studi Kasus:Informasi Seputar Program Studi Statistika Universitas Islam Indonesia)

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana
Program Studi Statistika



Disusun oleh :

Nita Tri Anggraini
15 611 061

JURUSAN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS ISLAM INDONESIA
Yogyakarta
2019

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul	:	Inisiasi *Natural Language Processing* (NLP) dan Klasifikasi Pertanyaan untuk Pembuatan "STATBOT" (Statistika *Chatter Bot*) (Studi kasus : Informasi Seputar Program Studi Statistika Universitas Islam Indonesia)
Nama Mahasiswa	:	Nita Tri Anggraini
Nomor Mahasiswa	:	15 611 061

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta, 15 Februari 2019

Menyetujui,

Dosen Pembimbing

(Ayundyah Kesumawati, S.Si., M.Si)

HALAMAN PENGESAHAN

TUGAS AKHIR

INISIASI *NATURAL LANGUAGE PROCESSING (NLP)* DAN KLASIFIKASI PERTANYAAN UNTUK PEMBUATAN "STATBOT" (STATISTIKA *CHATTER BOT*)

(Studi Kasus : Informasi Seputar Program Studi Statistika Universitas Islam Indonesia)

Nama Mahasiswa : Nita Tri Anggraini

Nomor Mahasiswa : 15 611 061

TUGAS AKHIR INI TELAH DIUJIKAN

PADA TANGGAL, 22 Februari 2019

Nama Penguji	**Tanda Tangan**
1. Dr. Jaka Nugraha, S.Si., M.Si	.....
2. Arum Handini Primandari, S.Pd.Si., M.Sc	.....
3. Ayundyah Kesumawati, S.Si., M.Si.	.....

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

(Prof. Riyanto, S.Pd.,M.Si., Ph.D)

KATA PENGANTAR

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

Dengan menyebut nama Allah SWT yang Maha Pengasih lagi Maha Panyayang, Kami panjatkan puja dan puji syukur atas kehadirat-Nya, yang telah melimpahkan rahmat, hidayah, dan inayah-Nya kepada kami, sehingga kami dapat menyelesaikan makalah ilmiah tentang limbah dan manfaatnya untuk masyarakat.

Tugas Akhir dengan judul : "**Inisiasi *Natural Language Processing* (NLP) dan Klasifikasi untuk Pembuatan "STATBOT" (Statistika *Chatter Bot*)"** ini telah penulis susun dengan maksimal dan mendapatkan bantuan dari berbagai pihak sehingga dapat memperlancar pembuatan makalah ini. Oleh karena itu, dalam kesempatan ini penulis menyampaikan ucapan terima kasih kepada :

1. Bapak Prof. Riyanto, S.Pd.,M.Si.,Ph.D. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia.
2. Bapak Dr. Edy Widodo, S.Si.,M.Si. selaku Ketua Jurusan Statistika beserta jajarannya.
3. Ibu Ayundyah Kesumawati, S.Si., M.Si, selaku dosen Pembimbing yang sudah membimbing dan memberikan dukungan serta motivasi kepada saya dari awal sampai akhir ini.
4. Seluruh staff pengajar Program Studi Statistika Universitas Islam Indonesia yang telah memberikan bekal ilmu kepada saya.
5. Orang tua tercinta beserta keluarga besar saya atas kasih sayang, dukungan dan doanya sehingga dapat menyelesaikan Tugas Akhir ini dengan baik.
6. Sahabat-sahabat tercinta saya (Listari, Fitri Khairani, Dita Apriani, Halimah Setio Rini, dan Vika Azkiya Dihni) yang selalu memberikan energi positif sehingga penulis termotivasi untuk selalu mengerjakan Tugas Akhir ini
7. Sahabat-sahabat di grup Abang Oh Abang (Listari, Fitri Khairani, Dita Apriani, Halimah Setio Rini, Vika Azkiya Dihni, Tihat Jaya Kusuma, Rohmat Apriyanto, Taufan Anindita Pradana, M. Dicky Hidayat, dan Muhammad Rafii) yang selalu memberikan dukungan, motivasi-motivasi

dan semangat dikala lelah dan malas dalam pengerjaan Tugas Akhir ini menghampiri.

8. Sahabat se-hobby saya, yang selalu mengerti di kala saya membutuhkan *refreshing* dan membuat *mood* saya kembali baik agar mempermudah penyelesian Tugas Akhir ini
9. Teman-teman bimbingan TA (Listari, Maymunah, Nida Nurhafidzah, Amalia Dwi N, Hadyati Utami, Meila Ika P, Haniatul Maghfuroh, Resa Mahardhika) yang sudah sama-sama berjuang, saling mengingatkan dan memberi motivasi serta dorongan untuk menyelasaikan Tugas Akhir ini
10. Teman-teman yang pernah menjadi rekan kepanitiaan ataupun organisasi saya, terimakasih terlah memberi pelajaran dan pengalaman yang sangat berharga bagi saya.
11. Teman-teman Statistika UII Angkatan 2015 (AFFINISTIC) yang bersama-sama menjadi pejuang gelar S.Stat dan Toga UII, yang selalu memberi informasi yang sangat bermanfaat, terimakasih semangat dan bantuannya
12. Serta semua pihak lainya yang tidak bisa dituliskan penulis satu per satu yang telah membantu selama pembuatan Tugas Akhir ini.

Penulis menyadari bahwa tugas akhir ini masih jauh dari kata sempurna dan masih banyak kekurangan. Hal tersebut dikarenakan keterbatasan ilmu dan pengetahuan yang dimiliki penulis semata. Oleh karena itu penulis mengharapkan kritik dan saran yang membangun dari pembaca untuk menyempurnakan penulisan tugas akhir ini. Semoga Tugas Akhir ini dapat memberikan manfaat bagi penulis khususnya dan umumnya bagi semua pihak yang membutuhkan. Akhir kata, semoga Allah SWT senantiasa melimpahkan rahmat serta hidayah-Nya kepada kita semua, Amin amin ya robbal ‘alamiin.

Yogyakarta, 22 Januari 2019

(Nita Tri Anggraini)

DAFTAR ISI

HALAMAN PERSETUJUAN PEMBIMBING ii
HALAMAN PENGESAHAN iii
KATA PENGANTAR iv
DAFTAR ISI vi
DAFTAR TABEL viii
DAFTAR GAMBAR ix
DAFTAR LAMPIRAN xi
HALAMAN PERNYATAAN xii
INTISARI xiii
ABSTRACT xiv
BAB I PENDAHULUAN 1
1.2 Latar Belakang 1
1.2 Rumusan Masalah 5
1.3 Batasan Masalah 5
1.4 Tujuan Penelitian 6
1.5 Manfaat Penelitian 6
BAB II TINJAUAN PUSTAKA 7
2.1 Penelitian Sebelumnya 7
BAB III LANDASAN TEORI 14
3.1 Konsep Sistem Informasi 14
3.2 Pengertian Informasi Kemahasiswaan 14
3.3 Pengertian Akademik dan Non-Akademik 14
3.4 Analisis Deskriptif 15
3.5 *Artificial Intellegent* 16
3.5.1 *Machine learning* 18
3.5.2 Pengertian *Natural Language Processing* (NLP) 20
3.6 *Chatter Bot (Chatbot)* 23
3.6.1. *Tkinter* 24
3.7 NLTK (*Natural Language Toolkit*) 26
3.8 *Text Mining* 27
3.8.1 *Text Preprocessing* 27

3.8.2 Pengertian *Term Frequency – Invers Document Frequency* (TF-IDF) 29
3.8.3 Simulasi Pembobotan TF-IDF.....30
3.8.4 *Topic Modelling*.....33
3.8.5 *Latent Dirichlet Allocation* (LDA).....34
3.8.6 *Topic Cohorence*.....36
3.9 Data *Training* dan Data *Testing* 38
3.10 Klasifikasi.....38
3.10.1 *Support Vector Machine* (SVM).....38
3.10.3 *Kernel Trick* dan *Non Linear SVM* 41
3.10.3 *Confusion Matrix* 44
BAB IV METODOLOGI PENELITIAN 46
4.1 Populasi dan Sampel 46
4.1.1 Teknik Pengambilan Sampel.....47
4.2 Sumber dan Jenis Data 48
4.3 Variabel Penelitian 48
4.4 Metode Pengumpulan Data 49
4.5 Alat dan Cara Organisir Data 50
4.6 Diagram Alir.....51
BAB V HASIL DAN PEMBAHASAN 53
5.1 Analisis Deskriptif.....54
5.2 *Preprocessing Data* 59
5.3 Pembobotan *Term Frequency-Invers Document Frequency* (TF-IDF) 6**Error! Bookmark not defined.**
5.4 *Topic Modelling* 64
5.4.1 Hasil *Topic Modelling* 64
5.5 Klasifikasi *Support Vector Machine* (SVM) 72
5.6 Rancang Bangun *Chatbot* "Statbot".....79
5.6.1 Analisa Sistem *Chatbot* 80
5.6.2 Pengujian *Chatbot* "STATBOT".....82
BAB VI KESIMPULAN DAN SARAN.....87
6.1 Kesimpulan.....87
6.2 Saran.....88
DAFTAR PUSTAKA.....89
LAMPIRAN.....93

DAFTAR TABEL

Tabel 2.1 Tabel Penelitian Terdahulu 10
Tabel 3.1 Distribusi Frekuensi Relatif..... 15
Tabel 3.2 Tabel Kontingensi 16
Tabel 3.3 *Widget Toolkit Ttk* 25
Tabel 3.4 Hasil Perhitungan Pembobotan Masing-masing Dokumen 32
Tabel 3.5 Tabel *Confussion Matrix* 44
Tabel 4.1 Tabel Perhitungan Sampel Masing-masing Kelas 47
Tabel 4.2 Tabel Variabel Penelitian 48
Tabel 5.1 Data Awal Penelitian..... 59
Tabel 5.2 Hasil *Lower Casing* 59
Tabel 5.3 Hasil *Remove Punctuation* 60
Tabel 5.4 Hasil *Stopwords Removal*..... 60
Tabel 5.5 Hasil *Tokenizing* 61
Tabel 5.6 Perhitungan TF Manual..... 62
Tabel 5.7 Perhitungan DF Manual 63
Tabel 5.8 Perhitungan IDF Manual..... 63
Tabel 5.9 Perhitungan TF-IDF Manual..... 63
Tabel 5.10 Nilai *Cohorence Value* 65
Tabel 5.11 Pembagian Data *Training* dan Data *Testing*..... 72
Tabel 5.12 *Confussion Matrix* Data Testing *Kernel Linear* 73
Tabel 5.13 *Confussion Matrix* Data Testing *Kernel RBF* 74
Tabel 5.14 *Confussion Matrix* Data Testing *Kernel Polynomial* 75
Tabel 5.15 *Confussion Matrix* Data Testing *Kernel Sigmoid* 76
Tabel 5.16 Hasil Nilai Akurasi *Kernel RBF, Polynomial, Liner* dan *Sigmoid*..... 77
Tabel 5.17 Hasil Prediksi 10 Data Baru 77
Tabel 5.18 Pengetahuan *Chatbot*..... 80
Tabel 5.19 Tabel Pengujian Sistem *Chatbot* 82

DAFTAR GAMBAR

Gambar 1.1 Presentase Mahasiswa yang Menyetujui Dibuatnya *Chatbot* Statistika UII 4
Gambar 3.1 Proses Motor Inferensi 17
Gambar 3.2 Perbedaan AI dan ML 20
Gambar 3.3 Representasi Model LDA 21
Gambar 3.4 Hubungan Antara AI, ML dan NLP 23
Gambar 3.5 Ilustrasi *Hyperplane* 34
Gambar 3.6 *Hyperplane* 39
Gambar 3.7 *Margin Hyperplane* 40
Gambar 3.8 Fungsi Φ memetakan data ke ruang vektor yang berdimensi lebih tinggi 41
Gambar 3.9 Elemen Utama Sistem Pemahaman *Natural Language* 42
Gambar 4.1 Diagram Alir Penelitian 48
Gambar 4.2 Diagram Alir Penelitian 49
Gambar 4.3 Diagram Alir Penelitian 49
Gambar 5.1 Grafik Jumlah Mahasiswa yang Menilai Kemudahan Memperoleh Informasi Jurusan 54
Gambar 5.2 Grafik Cara Mahasiswa Memperoleh Informasi Jurusan 55
Gambar 5.3 Grafik Jumlah Mahasiswa yang Menilai Kelengkapan Informasi di *Website* 56
Gambar 5.4 Grafik Jumlah Mahasiswa yang Mengetahui *Chatbot* dan yang Tidak Mengetahui *Chatbot* 57
Gambar 5.5 Grafik Informasi yang Ingin Didapatkan Dari *Chatbot* Statistika UII 58
Gambar 5.6 Proses *Tokenizing* 61
Gambar 5.7 Grafik *Cohorence Value* 61
Gambar 5.8 Visualisasi *Topic Modeling* dengan *PyLDAvis* 65
Gambar 5.9 Visualisasi Topik 1 66
Gambar 5.10 Visualisasi Topik 9 68
Gambar 5.11 Visualisasi Topik 4 69
Gambar 5.12 Visualisasi Topik 7 70

Gambar 5.13 Hasil *Lower Casing, Remove Punctuation* dan *Stopwoard Remover* 71
Gambar 5.14 *Confussion Matrix* Data *Testing Kernel Linear* 73
Gambar 5.15 *Confussion Matrix* Data *Testing Kernel RBF* 74
Gambar 5.16 *Confussion Matrix* Data *Testing Kernel Polynomial* 75
Gambar 5.17 *Confussion Matrix* Data *Testing Kernel Sigmoid* 76
Gambar 5.18 Matriks Klasifikasi Dengan *Kernel Linear* 78
Gambar 5.19 *Wordcloud* 79
Gambar 5.20 Percakapan dengan *Chatbot* "STATBOT" 86

DAFTAR LAMPIRAN

LAMPIRAN 1. Data Penelitian Lengkap.....93
LAMPIRAN 2. Hasil *Term Frequency – Invers Document Frequency*.....95
LAMPIRAN 3. *Syntax Preprocessing* 97
LAMPIRAN 4. *Syntax Topic Modelling* 99
LAMPIRAN 5. *Syntax Topic Modelling* 101
LAMPIRAN 6. *Syntax* dan *output* Klasifikasi SVM.....106
LAMPIRAN 7. *Syntax* Pembuatan *Chatbot* 116
LAMPIRAN 8. Contoh Percakapan dengan *Chatbot*.....117

HALAMAN PERNYATAAN

Dengan ini saya sebagai penulis menyatakan bahwa dalam Tugas Akhir ini adalah hasil karya penulis sendiri, dan tidak terdapat karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan disuatu Perguruan Tinggi dan sepanjang pengetahuan penulis juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan diterbitkan dalam daftar pustaka.

Yogyakarta, 22 Januari 2019



(Nita Tri Anggraini)

INISIASI *NATURAL LANGUAGE PROCESSING (NLP)* DAN KLASIFIKASI PERTANYAAN UNTUK PEMBUATAN "STATBOT" (STATISTIKA *CHATTER BOT*)

(Studi Kaus : Informasi Seputar Program Studi Statistika Universitas Islam Indonesia)

Nita Tri Anggraini

Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Islam Indonesia

INTISARI

Perkembangan teknologi saat ini mempermudah kegiatan berkomunikasi. Dimana komunikasi tidak hanya dilakukan secara tatap muka, tetapi bisa dilakukan dengan menggunakan perantara internet. Masyarakat saat ini bahkan cenderung bertanya sesuatu hal yang tidak dimengerti kepada internet. Teknologi yang keberadaannya dapat menjadikan komunikasi lebih interaktif layaknya berinteraksi dengan manusia yaitu teknologi Chatter Bot atau Chatbot. Seperti di Program studi Statistika, hal yang sering terjadi adalah ketika masa key-in, kerja praktek, tutup teori, ujian komprehensif, dan sebagainya. Banyak mahasiswa yang kurang mendapatkan informasi tersebut, sehingga menyebabkan beberapa mahasiswa harus berkali-kali menanyakan hal-hal demikian kepada pihak jurusan. Hal ini yang menjadikan penulis ingin membuat chatbot untuk jurusan Statistika UII, dimana chatbot nantinya akan membantu atau mempermudah mahasiswa Statistika UII khususnya untuk memperoleh informasi yang diinginkan dan membantu agar pihak jurusan Statistika UII tidak berulang-ulang menyampaikan informasi terkait jurusan. Teknologi chatbot akan lebih mudah dipahami jika menggunakan bahasa yang digunakan untuk komunikasi sehari-hari oleh manusia, untuk itu diperlukan Natural-language Processing (NLP). Karena data penelitian ini berbentuk teks, maka akan dilakukan pembobotan kata dengan metode Term Frequncy- Invers Document Frequency (TF-IDF) dan dilakukan Topic Modelling dengan Latent Dirichlet Allocation (LDA) untuk membagi kata-kata dalam pertanyaan ke dalam topik dan melihat keterkaitan antar kata tersebut. Kemudian pertanyaan yang ada diklasifikasikan dengan Support Vector Machine (SVM). Hasil klasifikasi yang dilakukan diperoleh hasil akurasi 0.888%. kemudian kemampuan chatbot dalam menjawab pertanyaan yang di ajukan adalah sebesar 70%.

Kata Kunci* : *Chatbot, NLP, Topic Modelling, LDA, SVM

INITIATION NATURAL LANGUAGE PROCESSING (NLP) AND CLASSIFICATION OF QUESTIONS FOR MAKING "STATBOT" (STATISTICS CHATTER BOT)

(*Study Case: Information of Departement Statistics Islamic University of Indonesia*)

Nita Tri Anggraini

Department Statistics, Faculty of Mathematics and Natural Science

Islamic University of Indonesia

ABSTRACT

Current technological developments. Where communication cannot be done officially, but it can be done using the internet. People today are wondering something that cannot be discussed on the internet. Interesting technology can make communication more interactive like interaction with humans, namely technology Chatter Bot or Chatbot. As in the Statistics Study Program, what often happens is key compilation, internship program, closing theory, examinations completed, and so on. Many students did not get this information, so that some students had to ask such things repeatedly to the department. Statistics UII, where chatbots will help or assist UII Statistics students specifically to get the information needed and help so that the UII Department of Statistics cannot be re-accessed information related to majors. Chatbot technology will be easier to use if you use a language that is used for everyday communication by humans, therefore Natural Language processing (NLP) is needed. Because this research data forms text, word weighting will be carried out using the Term Freq-Invers Document Frequency (TF-IDF) method and a Topic Modeling with Latent Dirichlet Allocation (LDA) is done to connect the words in the question into the question and see the interrelationship between the word. Then the question answers with Support Vector Machine (SVM). The results of the classification carried out obtained an evaluation of 0.888%. then the chatbot's ability to answer the questions raised is 70%.

Keywords : Chatbot, NLP, Topic Modelling, LDA, SVM