Testimonial:

- *"Buku ini sangat tepat untuk Mahasiswa khususnya dibidang Pendidikan Komputer dan Teknik Informatika dalam mempelajari dan mengefektivitaskan konsep dasar e-learning melalui diskusi on-line. Penerapan Metode TF-IDF dan Naive Bayes menambah optimalitas pelaksanaan diskusi tersebut. Selain mahasiswa, buku ini juga dapat dijadikan pedoman bagi instruktur ataupun tenaga didik dalam memonitor dan mengevaluasi pelaksanaan diskusi on-line dalam proses pembelajaran".*

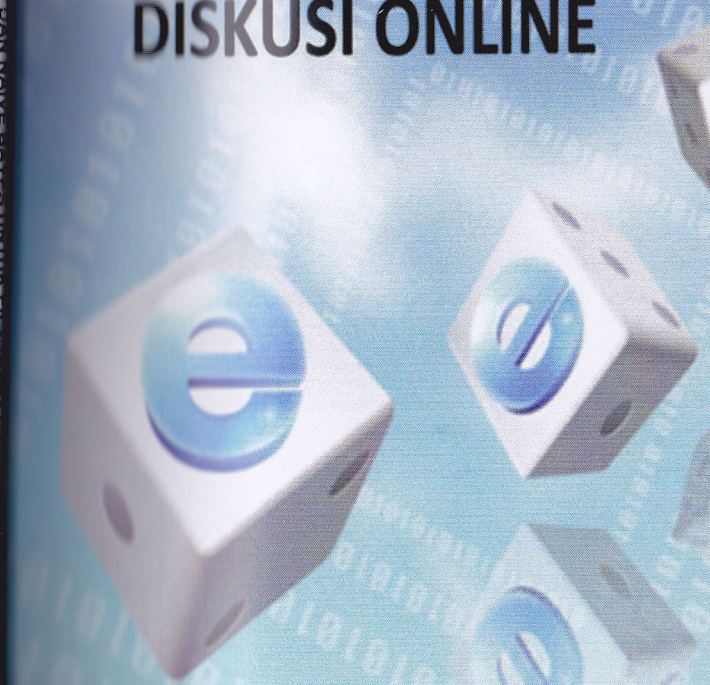
Dr. Saktioto, Mphil (Dosen FMIPA UNRI)

Penerbit: Al-Mujtahadah Press
Jl. Handayani Gg. Ros No. 52 Kel. Maharatu
Kec. Marpoyan Damai Pekanbaru-Riau
Website: al-mujtahadahpress.blogspot.com
E-mail: almujtahadahprss@gmail.com
Hp: 081365662636, 081378712088

ISBN 978-602-9392-82-1

OKFALISA

DAN NAÏVE BAYES UNTUK RING DAN MONITORING DISKUSI ONLINE



OKFALISA

DR. OKFALISA, MSC

TF-IDF DAN NAÏVE BAYES
UNTUK FILTERING DAN MONITORING DISKUSI ONLINE

TF-IDF DAN NAÏVE BAYES UNTUK FILTERING DAN MONITORING DISKUSI ONLINE

Penulis :
DR. OKFALISA, MSC

Lay Out : Eko Syahputra
Desain Cover: Okfalisa
Percetakaan : Utama Grafika

Penerbit :
Al-Mujtahadah Press
Jl. Handayani Gg. Ros No. 52 Kel. Maharatu Kec. Marpoyan Damai
Pekanbaru-Riau Email: almujtahadahprss@gmail.com
Hp. 0813 65662636 Hp. 0813 78712088

Cetakan :
Pertama, November 2016

Katalog Dalam Terbitan (KDT)

OKFALISA,
TF-IDF DAN NAÏVE BAYES UNTUK FILTERING DAN MONITORING DISKUSI ONLINE/Oleh: Okfalisa,
--Cet. Pekanbaru : Al-Mujtahadah Press, 2016
viii, 65 hlm.; ilus, 21 cm

ISBN 978-602-9392-82-1

1. **Diskusi di Internet** I. Judul

004.693

KATA PENGANTAR

Puji syukur kepada Allah SWT, penulisan buku dengan tema Content Based Spam Filtering pada Diskusi Online dengan Menggunakan Metode TF-IDF dan Naïve Bayes telah selesai dilaksanakan. Buku ini memberikan kontribusi baru terhadap dunia pendidikan dan pembelajaran berbasis ICT melalui proses penjaringan dan ekstraksi informasi dan data pada forum diskusi online. Hal ini memberikan pencerahan pada proses pembelajaran online untuk dapat dimanfaatkan menjadi lebih efektif dan efisien. Melalui aplikasi ini pihak pendidik dapat menilai kemampuan dan karakter siswa melalui tanggapan dan informasi yang diberikan sehingga proses akusisi pengetahuan dapat terbentuk dengan optimal. Secara analisis penerapan algoritma kedua metode TF-IDF dan Naïve Bayes menggambarkan effektivitas penelusuran yang baik dalam proses filtering resume konten.

Pekanbaru, 25 Agustus 2016

Penulis

DAFTAR ISI

KATA PENGANTAR iii
DAFTAR ISI iv
DAFTAR GAMBAR vi
DAFTAR TABEL viii

BAB I TEORI DASAR
1.1 Pendahuluan 1
1.1.1 Latar Belakang 1
1.1.2 Batasan Masalah 8
1.1.3 Tujuan 8
1.1.4 Signifikansi 8
1.2 Diskusi Online 10
1.3 Content Based Spam Filtering 11
1.4 Algoritma TF-IDF 12
1.5 Algoritma Naïve Bayes 12
1.6 Penelitian Terkait 13

BAB II METODOLOGI
2.1 Identifikasi Permasalahan 16
2.2 Pembangunan Instrumen 17
2.3 Studi Pendahuluan 18
2.4 Analisa dan Perancangan Model 19
2.5 Impelementasi dan Pengujian 20
2.6 Kesimpulan dan Saran 20

BAB III PEMBANGUNAN INSTRUMENT
3.1 Rancangan Instrument dan Penelusuran Proses 21
3.2 Representasi Dokumen/Pembobotan tf-df 30

3.3 Klasifikasi One-Class Naïve Bayes 31

BAB IV ANALISA DAN PERANCANGAN
4.1 Analisa Sistem 35
4.2 Perancangan Sistem 39
4.3 Context Diagram 39
4.3.1 Data Flow Diagram Level 1 40
4.3.2 Data Flow Diagram Level 2 Proses 2 Preprocessing 41
4.3.3 Data Flow Diagram Level 2 Proses 3 Klasifikasi 42
4.4 ERD (*Entity Relation Diagram*) 43
4.4.1 Struktur Database 44
4.5 Tampilan Menu Sistem 45

BAB V IMPLEMENTASI DAN PENGUJIAN
5.1 Implementasi 49
5.1.1 Tampilan menu utama system 49
5.2 Pengujian 50
5.2.1 Pengujian akurasi system 50

BAB VI KESIMPULAN DAN SARAN
6.1 Kesimpulan 59
6.2 Saran 60

DAFTAR PUSTAKA 63

DAFTAR GAMBAR

Gambar 2.1. Skema Penelitian 15
Gambar 2.2. Logika Konsep Content Based Spam Filtering 18
Gambar 3.1. Flowchart Klasifikasi Relevansi Komentar dan Topik 22
Gambar 3.2. Preprocessing 23
Gambar 3.3. Contoh Proses *Tokenization* 24
Gambar 3.4. Contoh Proses *Case Folding* 25
Gambar 3.5. Flowchart Spelling Normalization 26
Gambar 3.6. Contoh Proses *Spelling Normalization* 27
Gambar 3.7. Contoh Hasil *Filtering* 28
Gambar 3.8. Proses *Stemming* dengan Algortima Porter Stemmer 29
Gambar 3.9. Proses Klasifikasi dengan *One-Class Naive Bayes* 32
Gambar 4.1. *Context Diagram* 39
Gambar 4.2. *Data Flow Diagram I* (DFD) Level 1 40
Gambar 4.3. DFD Level 2 Proses 2 *Preprocessing* 42
Gambar 4.4. DFD Level 2 Proses 3 Klasifikasi 43
Gambar 4.5. *Entity Relation Diagram* (ERD) 44
Gambar 4.6. Tampilan Utama Sistem 48
Gambar 5.1. Tampilan Utama Sistem 50
Gambar 5.2. Topik diskusi 51
Gambar 5.3. Komentar Partisipan 51
Gambar 5.4. Ekstraksi dan Klasifikasi Website Indowebster 52
Gambar 5.5. Hasil Ekstraksi Topik Diskusi 52
Gambar 5.6. Hasil Ekstraksi Komentar 53
Gambar 5.7. Hasil Tokenisasi 53

Gambar 5.8. Hasil Stemming 54
Gambar 5.9. Hasil Indeks Topik 54
Gambar 5.10. Hasil Indeks Komentar 55
Gambar 5.11. Hasil Klasifikasi Komentar 55

DAFTAR TABEL

Tabel 3.1 Daftar Kamus Kata Tidak Baku 26
Tabel 3.1 Pembobotan tf-d 30
Tabel 3.2 Nilai Probabilitas Setiap Kata Terhadap Kelas Topik 32
Tabel 3.3 Probabilitas Relevansi Komentar Terhadap Topi 33
Tabel 4.1 Data topik_komenta 44
Tabel 4.2 Data Komentar 45
Tabel 4.3 Data Token 45
Tabel 4.4 Data kamus_kata_tidak_baku 46
Tabel 4.5 Data kamus_stem 46
Tabel 4.6 Data stem 46
Tabel 4.7 Data indeks_top 47
Tabel 4.8 Data indeks kata komentar 48
Tabel 5.1 Matrix Confusion Klasifikasi Website Indowebster 56
Tabel 5.2 Hasil Pengujian 57

BAB I
TEORI DASAR

1.1 Pendahuluan

1.1.1. Latar Belakang

Teknologi Information *Communication and Technology* (ICT) sangat erat kaitannya dengan dunia pendidikan, terutama dalam proses belajar mengajar. Pesatnya perkembangan teknologi mendorong para tenaga pendidik (akademisi) untuk lebih memahami dan memanfaatkan teknologi tersebut dalam proses belajar mengajar aktif baik didalam maupun diluar kelas. Hal ini menjadi suatu tantangan yang besar bagi sebuah institusi perguruan tinggi dalam menghadapi kompetisi global dalam memberikan pelayanan pendidikan yang tidak terbatas.Karena kita pahami bersama bahwa diperguruan tinggi inilah transformasi pengetahuan dan komunikasi dapat dibangun tanpa batasan waktu dan tempat, tidak hanya melalui pertemuan *face to face* dikelas saja, namun juga didunia maya.

Internet sebagai salah satu teknologi ICT telah banyak dimanfaatkan untuk melakukan komunikasi dikehidupan sehari-hari. Komunikasi dalam bentuk *face to face* digantikan dengan komunikasi berbasis teknologi. Bahkan banyak sekali media sosial yang dapat dimanfaatkan untuk berkomunikasi, tukar menukar pengalaman, informasi dan pengetahuan (Murray, 2008). Beberapa sosial media yang cukup digemari dalam melakukan komunikasi virtual antara lain adalah Facebook, Linked, Twitter, Youtube, MySpace, Google Plus, Classroom 2.0, Plurk, etc. Bahkan pengguna sosial media ini setiap tahunnya terus meningkat, terutama

Facebook (Toprak et al., 2009).Hal ini tentunya menjadi perhatian bagi para akademisi termasuk peneliti, guru dan siswa itu sendiri.

Apabila dikaitkan bagaimana Alquran memahami perkembangan teknologi dapat dilihat pada Surah An-Nahl (Q16:89).

وَيَوْمَ نَبْعَثُ فِي كُلِّ أُمَّةٍ شَهِيدًا عَلَيْهِم مِّنْ أَنفُسِهِمْ ۖ وَجِئْنَا بِكَ
شَهِيدًا عَلَىٰ هَٰؤُلَاءِ ۚ وَنَزَّلْنَا عَلَيْكَ الْكِتَابَ تِبْيَانًا لِّكُلِّ شَيْءٍ
وَهُدًى وَرَحْمَةً وَبُشْرَىٰ لِلْمُسْلِمِينَ ﴿٨٩﴾

(Dan ingatlah) akan hari (ketika) Kami bangkitkan pada tiap-tiap umat seorang saksi atas mereka dari mereka sendiri dan Kami datangkan kamu (Muhammad) menjadi saksi atas seluruh umat manusia. Dan Kami turunkan kepadamu Al Kitab (Al Quran) untuk menjelaskan segala sesuatu dan petunjuk serta rahmat dan kabar gembira bagi orang-orang yang berserah diri.

dan Surah Al-Isra (Q17:12).

وَجَعَلْنَا الَّيْلَ وَالنَّهَارَ آيَتَيْنِ ۖ فَمَحَوْنَا آيَةَ الَّيْلِ وَجَعَلْنَا آيَةَ النَّهَارِ
مُبْصِرَةً لِّتَبْتَغُوا فَضْلًا مِّن رَّبِّكُمْ وَلِتَعْلَمُوا عَدَدَ السِّنِينَ
وَالْحِسَابَ ۚ وَكُلَّ شَيْءٍ فَصَّلْنَاهُ تَفْصِيلًا ﴿١٢﴾

Dan Kami jadikan malam dan siang sebagai dua tanda, lalu Kami hapuskan tanda malam dan Kami jadikan tanda siang itu terang, agar kamu mencari kurnia dari Tuhanmu, dan supaya kamu mengetahui bilangan tahun-tahun dan perhitungan. Dan segala sesuatu telah Kami terangkan dengan jelas.

Kedua surah diatas, An-Nahl (Q16:89) dan Al-Isra (QS17:12) menjelaskan bahwa Alquran adalah sumber dari segala pengetahuan termasuk didalamnya teknologi dan aplikasinya yang dapat meningkatkan kualitas hidup manusia. Dalam islam pendidikan haruslah membentuk keseimbangan antara karakter siswa dengan kemampuan agama yang cukup untuk menjalankan syariat-syariat islam. Tidak dapat dihindari lagi bahwa berbagai metode pembelajaran berbasis teknologi telah memberikan daya tarik yang luar biasa terhadap siswa. Berbagai pembuktian telah dilakukan bahwa teknologi dapat meningkatkan kemampuan pemahaman siswa dalam proses belajar mengajar (Azwan *et al.*,2005; Chong *et al.*, 2005; Samuel and Zaintun, 2006). Selain itu, metode ini juga dapat membantu para akademisi dalam memberikan teknik pembelajaran yang interaktif terhadap siswanya.

Salah satu bentuk teknologi pembelajaran lainnya yang berbasis ICT adalah *E-learning* (Pembelajaran Jarak Jauh). Teknik pembelajaran dengan mengirimkan materi melalui media elektronik misalkan internet, intranet ataupun extranet serta dapat juga melalui perangkat media tape audio dan video, satelit, tv interaktif, CD-ROM dan lain sebagainya (Shee dan Wang, 2008). Secara konsep, *E-learning* hanya merubah

budaya pembelajaran yaitu dari model pembelajaran yang bersifat tradisional (intruksi) kepada model paradigma yang bersifat pembelajaran berbasis teknologi (modern). Ada beberapa keutamaan dalam penggunaan *E-learning* bagi pengguna antara lain adalah dari segi efektifitas biaya yang jauh lebih murah dengan pertemuan didunia maya dibandingkan dengan harus bertemu langsung dengan siswa (jarak jauh); konten yang selalu terbaharukan dan akses yang bersifat fleksibel dapat diunggah kapan dan dimana saja (Yang, dkk, 2008). Pada *E-learning* juga terdapat konsep yang memungkinkan setiap peserta dan trainer (tenaga pendidik) dapat berinteraksi secara bersama melalui Diskusi *Online*.

Diskusi *Online* sering sekali mendapatkan sorotan terutama pada pemanfaatan arsip digital atau material yang ada, sehingga dapat dikatakan materi atau arsip digital yang ada belum dapat dimanfaatkan secara maksimal (Lui dkk, 2007). Jika ditinjau dari analisa konten pada diskusi *online*, isi dari konten komunikasi dapat digunakan untuk proses analisis yang bersifat objektif untuk dapat menyimpulkan latar belakang ataupun hasil dari komunikasi. Sehingga pentingnya dari isi komunikasi pada diskusi *online* memberikan pengaruh terhadap proses pembelajaran, karakteristik peserta dan hasil belajar. Oleh karena itu, tidak jarang sekarang ini, model diskusi *online* sering digunakan sebagai konsep untuk proses belajar mengajar, dan dapat dipergunakan sebagai bahan pertimbangan untuk melakukan proses penilaian, evaluasi dan juga bahan referensi pada bidang pendidikan dan pengajaran.

Selain manfaat yang besar dari proses pembelajaran *E-Learning*, kelemahan dan berbagai permasalahan juga bermunculan pada saat penerapan metode tersebut, terutama pada proses diskusi *online*. Hal tersebut tentunya tentunya cukup menghambat proses belajar mengajar. Bahkan waktu yang seharusnya digunakan untuk berdiskusi sering kali tidak dimanfaatkan secara maksimal yang menyebabkan analisis konten terhadap informasi yang disampaikan oleh peserta maupun tenaga didik tidak mencapai hasil yang diharapkan (Lui. dkk, 2007). "Lain yang disampaikan oleh tenaga didik, lain pula yang didiskusikan dalam forum tersebut". Bahkan, tidak sedikit materi yang dibicarakan dalam diskusi *online* tidak relevan dengan konten materi yang diajukan (*spam content*). Sehingga bukan forum diskusi yang diharapkan terbentuk, melainkan terjadinya perdebatan yang simpang siur di dalam forum tersebut. Perdebatan yang positif dan saling mendukung tentunya akan membentuk karakter siswa dalam menghasilkan suatu pengetahuan yang baru berdasarkan informasi yang ada. Sebaliknya, perdebatan kusir yang terjadi dalam diskusi *online* akan membentuk opini dan penerjemahan yang salah dan tidak tepat bagi para siswa. Sehingga bukan kebaikkan yang diperoleh, namun mudharatnya yang dirasakan sebagai akibat dari proses belajar mengajar tersebut.

Berikut petikan beberapa hadist mengenai seruan untuk menghindari debat kusir[1].

Rasulullah -*shallallohu 'alaihi wasallam*- bersabda:

[1]http://www.arrahmah.com/kajian-islam/mau-rumah-di-surga-jauhi-debat-kusir-walaupun-anda-dalam-kebenaran.html

أَنَا زَعِيمٌ بِبَيْتٍ فِى رَبَضِ الْجَنَّةِ لِمَنْ تَرَكَ الْمِرَاءَ وَإِنْ كَانَ مُحِقًّا وَبِبَيْتٍ فِى وَسَطِ الْجَنَّةِ لِمَنْ تَرَكَ الْكَذِبَ وَإِنْ كَانَ مَازِحًا وَبِبَيْتٍ فِى أَعْلَى الْجَنَّةِ لِمَنْ حَسَّنَ خُلُقَهُ

*"Aku menjamin sebuah rumah di pinggir jannah (surga) bagi siapa saja yang meninggalkan perdebatan berkepanjangan meskipun ia dalam kebenaran (al haq), juga sebuah rumah di tengah jannah bagi siapa saja yang meninggalkan berbohong walaupun ia sedang bercanda, serta sebuah rumah di puncak jannah bagi siapa saja yang berakhlak mulia."**(HR. Abu Dawud, Dinyatakan Hasan shahih oleh Syaikh Al Albani)***

Umar Bin Khattab berkata :

لا يجد عبد حقيقة الإيمان حتى يدع المراء وهو محق ويدع الكذب في المزاح وهو يرى أنه لو شاء لغلب

"Seseorang tidak akan merasakan hakikat iman sampai ia mampu meninggalkan perdebatan yang berkepanjangan meskipun ia dalam kebenaran, dan meninggalkan berbohong meskipun hanya bercanda padahal ia tahu seandainya ia mau ia pasti menang dalam percebatan itu"***(Kanzul Ummal juz 3 hal 1165)***

Imam Ishaq bin Isa berkata :

المِراء والجِدال في العلم يَذهبُ بنور العلم من قلب الرجل

"Imam Malik bin Anas mengatakan : "Debat kusir dan pertengkaran dalam masalah ilmu akan menghapuskan cahaya ilmu dari hati seseorang"

Imam Ibnu Wahab berkata : "Aku mendengar Imam Malik bin Anas mengatakan :

المراء في العلم يُقسّي القلوب ، ويورّث الضغن

"Perdebatan dalam ilmu akan mengeraskan hati dan menyebabkan kedengkian"***(Jaami' al Uluum wak Hikam 11/16)***

Hal ini jika dikaitkan dengan istilah teknologi informasi (ICT), khususnya pada buku ini debat kusir sama dengan informasi yang tidak bermanfaat *(Spam)*. Konten yang bersifat *spam* dapat dilabel berdasarkan informasi yang disampaikan oleh pemilik materi. Konten Spam ini dapat dianalisa berdasarkan relevansi dan pengelompokkan konten terhadap informasi antara tenaga didik dan siswa (Yang, dkk, 2008 dan Lui, dkk, 2007). Hal ini menjadi tanggung jawab administrator pembelajaran *e-learning* dan juga tenaga pendidik untuk memastikan informasi yang salah tidak berkembang lebih jauh. Sehingga forum diskusi *online* dapat terkontrol dengan baik, melalui fasilitas filterisasi konten.

Berdasarkan latar belakang diatas, buku dengan judul "Content Based Spam Filtering pada Diskusi Online dengan Menggunakan Metode TF-IDF dan Naïve Bayes" dirancang. Dalam buku ini Metode TF-IDF dan Naïve Bayes diaplikasikan untuk menjaring dan mengekstraksi informasi dan data antara materi pendidik

dan hasil diskusi siswa melalui proses *Content Based Spam Filtering* (Wang dkk, 2008; Almeida dan Yamakami, 2010; Lee dan Kim, 2008).

1.1.2. Batasan Masalah

Agar buku ini memiliki arah yang jelas dan tepat, maka dibutuhkan beberapa batasan yaitu:

1. Proses *Content Based Spam Filtering* dibatasi pada aplikasi *E-Learning* berbasis Web misalkan Blog, Forum dan Portal.
2. Proses *Content Based Spam Filtering* dilakukan berdasarkan resume komentar siswa terhadap keterkaitannya dengan materi yang disampaikan. Relavansi ditinjau dari segi bahasa atau kosa kata yang digunakan tidak berdasarkan makna kata.

1.1.3. Tujuan

Adapun tujuan dari buku yang dilakukan sebagai berikut:

1. Mempelajari bagaimana metode TF-IDF dan Naive Bayes pada diskusi *online* diterapkan untuk menyaring *Content Based Spam Filtering* berdasarkan komentar siswa terhadap materi yang diberikan.
2. Membangun aplikasi *Content Based Spam Filtering* yang mampu mengontrol dan memonitor proses diskusi pada diskusi *online*.

1.1.4. Signifikansi

Dilihat dari signifikansinya, buku ini dapat memberikan kontribusi secara teori, praktek maupun metodologi.

Kontribusi secara teori:

1. Buku ini mempelajari bagaimana metode TF-IDF dan Naive Bayes dapat diterapkan untuk menyaring berdasarkan Content Based Spam Filtering pada proses pembelajaran *online*.
2. Model ini merupakan pengembangan dari pendekatan buku dibidang *Information Retrieval*, Algoritma TF-IDF dan Naive Bayes serta konsep Spam Filtering.
3. Selanjutnya, Aplikasi metode TF-IDF dan Naive Bayes pada proses filtering ini akan memberikan hasil analisa baru terhadap penerapan konsep Content Based Spam Filtering pada diskusi *online*.

Kontribusi secara praktek:

1. Aplikasi ini dapat digunakan oleh pihak pendidik ataupun administrator sistem *e-learning* untuk memonitor diskusi *online* yang terjadi.
2. Aplikasi ini mampu membantu pihak pendidik untuk menilai kemampuan dan karakter siswa melalui pemberian tanggapan dan informasi yang terkait dengan materi ajar. Tentunya hal ini akan berpengaruh pada pembentukan akuisisi pengetahuan pada siswa.
3. Melalui penggunaan aplikasi ini, pihak tenaga pendidik dapat menilai dan menentukan startegi ataupun metode pembelajaran yang lebih tepat untuk siswa didiknya. Sehingga pemahaman siswa terhadap materi menjadi lebih baik.
4. Bagi pihak universitas tentunya aplikasi ini memberikan wajah baru dalam proses monitoring

dan evaluasi secara lebih inovatif dan effektif pada pembelajaran yang diberikan.

5. Bagi siswa sendiri, tentunya dengan adanya filtering informasi akan memudahkan mereka dalam mendapatkan informasi yang tepat dan relevan terkait materi pembelajaran yang diberikan.

Kontribusi secara metodologi:

1. Buku ini menggambarkan efektivitas penelusuran metode Algoritma TF-IDF dan Naive Bayes dalam proses filtering konten diskusi *online*.
2. Buku ini menunjukkan efektivitas penggunaan Spam Filtering pada diskusi *online*.

1.2. Diskusi *Online*

Diskusi *online* pada prinsipnya adalah memudahkan proses untuk membangun pengetahuan dan saling bertukar ide, pengalaman dan sudut pandang yang berbeda pada tiap-tiap peserta (Jamaluddin, Hashim, Hanafiah, Mohd Zahari, & Zulkifly, 2010). Selain itu juga, diskusi *online* dapat dijadikan sebagai wadah bagi peserta untuk berdiskusi saling mengeluarkan pendapat dan proses pembentukan pengetahuan baru. Diskusi *online* juga dapat digunakan sebagai media untuk meningkatkan performansi partisipasi peserta khususnya bagi peserta yang mengalami masalah komunikasi di ruang kelas atau *face to face class*. Oleh karena itu Diskusi *online* dapat dijadikan alternative bagi mereka yang ingin mendapatkan ataupun meningkatkan kemampuan dalam berkomunikasi dan menganalisis permasalahan.

Berikut ini beberapa manfaat dari fasilitas diskusi *online* pada proses belajar mengajar.

- Membangun suasana kelas yang memiliki keunggulan diskusi mengenai topik yang diajarkan.
- Memberikan waktu dan ruang kepada siswa untuk memperdalam tentang pengetahuan mereka sebelum ikut berpartisiapasi dalam diskusi *online*.
- Memfasilitasi proses belajar mengajar dengan memungkinkan siswa untuk meriview karya/pendapat orang lain.
- Mengembangkan pemikiran dan ketrampilan menulis
- Memungkinkan peserta untuk berpartisipasi dengan postingan informasi dan tanggapan dari setiap pertanyaan. (TeacherStream, 2012)

1.3. Content Based Spam Filtering

Istilah *spam* biasanya digunakan pada e-mail (Almeida & Yamakami, Content-Based Spam Filtering, 2010) ataupun twitter. Karena *spam* tersebut biasanya berkaitan dengan hal-hal yang tidak ada hubungannya dengan pemilik informasi. Namun pada buku ini, konsep *content based spam filtering* akan digunakan untuk memfilter isi dari konten yang disampaikan oleh pemateri ataupun pengguna (Yu & Chen, 2012) dalam suatu diskusi *online*. Ketika isi dari konten yang disampaikan oleh pengguna tidak relevan dengan materi yang disampaikan, maka akan diberikan label spam oleh pemilik materi ataupun administrator secara otomatis.

1.4. Algoritma TF-IDF (*Term Frequency – Inverse Document Frequency*)

Dasar dari TF-IDF merupakan sebuah pemodelan yang bersifat teoritis, yaitu sebuah kata di dalam dokumen dapat dipisahkan dengan atau tanpa penggunaan klasifikasi (Pramono, Rohman, & Hindersah, 2013). Klasifikasi kata dari dokumen akan dievaluasi oleh TF (*Term Frequency*) dan IDF (*Inverse Document Frequency*) dengan cara mengukur tingkat kepentingan dan *similarity* kata tersebut dalam konteks dokumen. Berikut bentuk Penggunaan algoritma TF-IDF pada buku ini:

$$\text{idf} = \log \frac{N}{df} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$W_{dt} = TF_{dt} * IDF_t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Keterangan:
N : Jumlah komentar yang berisi term (t)
df : Jumlah kemunculan kata terhadap N
d : Kalimat ke-d
t : Kata (term) ke-t
TF: Frekuensi Kata
W : bobot kalimat ke-d terhadap kata (term) ke-t
IDF: *Inverse Document Frequency*

1.5. Algoritma Naïve Bayes

Algoritma *Naive Bayes* merupakan salah suatu algoritma yang sering digunakan untuk proses klasifikasi atau pengelompokan. Pengklasifikasian dalam algoritma ini menggunakan metode probabilitas dan statistik berdasarkan kesamaan ciri dari masing-masing variabel.

Klasifikasi *Naive bayes* mengasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya.
Penggunaan algoritma Naïve bayes pada buku ini adalah sebagai berikut:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(X)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$P(Y)\prod_{i=1}^{d} P(X_i|Y) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$P(X|Y) = P(X_1|Y) * P(X_2|Y) * P(X_i|Y) \quad \ldots\ldots\ldots\ldots\ldots (3)$$

Keterangan:
P(X |Y) = nilai seluruh variabel atau atribut
X = variabel atau atribut
(X_i) = Peluang X
Y = kelas kategori

1.6. Penelitian Terkait

Untuk memastikan buku ini memiliki *track record* yang jelas, berikut kami jabarkan beberapa penelitian yang terkait dengan *content based spam filtering* pada diskusi *online*.

1. Burdescu, D. M. Mihaescu, dan B. Logofatu (2008) dengan penelitiannya yang berjudul "*Employing Bayes Classifier for Improving Learner's Proficiency*". Penggunaan metode naïve bayes pada penelitian ini adalah sebagai penentu apakah sebuah chapter dapat digunakan untuk pelajar tertentu. Hasil penelitian ini adalah bagi pelajar yang mengikuti rekomendasi sistem memiliki kemampuan belajar yang lebih baik dalam waktu yang lebih singkat.

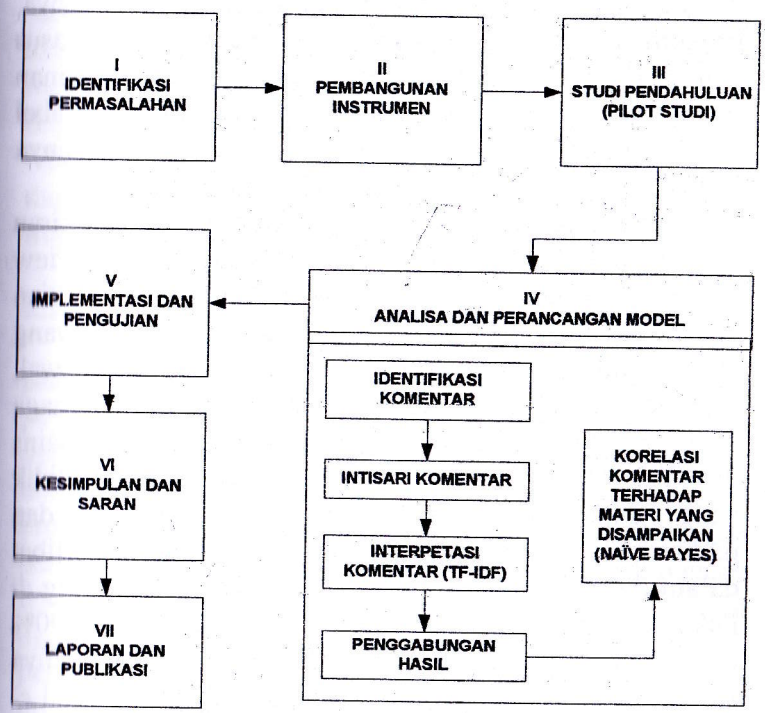
2. Phuc D dan Nguyen Thi Kim Phung (2007) dengan penelitiannya yang berjudul "*Using Naive Bayes Model and Natural Language Processing for Classifying Messages on Online Forum*". Penelitian ini mengklasifikasikan pesan yang disampaikan dalam forum diskusi *online* melalui penerapan Algoritma Bayesan dan Natural Languange Processing. Kombinasi kedua metode ini menghasilkan tingkat korelasi yang tinggi antar kata dalam setiap pesan yang disampaikan.

3. Lee Sungjick dan Kim Han-joon (2008) pada penelitiannya yang berjudul "*News Keyword Extraction for Topic Tracking*". Penelitian ini mendiskusikan penerapan metode TF-IDF dalam menentukan dan menetapkan kata kunci dari sebuah artikel berita.

4. Lui Andrew Kwok-Fai, Li Siu Cheung, dan Choy Sheung On (2007) pada penelitiannya yang berjudul "*An Evaluation of Automatic Text Categorization in Online Discussion Analysis*". Penelitian ini menggunakan konsep *text categorization* dalam mengevaluasi, memeriksa dan menganalisis kata dalam diskusi *online*.

BAB II
METODOLOGI

Dalam proses pelaksanaan penelitian, agar buku ini lebih terarah dan tepat sasaran, beberapa tahapan yang dilakukan antara lain adalah sebagai berikut (Gambar 2.1 Skema Penelitian):



Gambar 2.1: Skema Penelitian

2.1. Identifikasi Permasalahan

Tahap ini dilalui dengan melakukan serangkaian aktivitas meliputi tinjauan pustaka dan interview dan observasi lapangan. Tinjauan pustaka dilakukan dengan mempelajari beberapa konsep pendukung buku diantaranya adalah *Content Based Spam Filtering*, Metode *TF-IDF* dan *Naive Bayes* serta diskusi *online*. Berbagai isu yang berkaitan dengan konsep pendukung dipelajari baik melalui buku, jurnal ataupun publikasi ilmiah lainnya. Bagaimana penerapan konsep ini pada setiap kasus dianalisis untuk melihat prosedur dan kemungkinan penerapan konsep pada kasus yang akan diteliti. Tabel perbandingan ketiga konsep dan teori pendukung lainnya akan dibangun untuk memperkuat landasan teori.

Selanjutnya, guna memperkuat hasil pemikiran konseptual dari landasan teori yang dibangun, interview dan beberapa observasi relavan terhadap kasus penelitian dilakukan. Interview dilakukan kepada responden yang terlibat pada pembelajaran *e-learning*. Untuk tahap awal, interview dilakukan kepada beberapa dosen/tenaga pendidik dan mahasiswa dilingkungan Fakultas Sains dan Teknologi. Responden dibatasi pada tenaga didik yang dapat berfungsi sebagai adminitrator dan mahasiswa sebagai peserta didik yang pernah terlibat dalam diskusi *online*. Meskipun penerapan *e-learning* di Fakultas ini masih belum maksimal namun hampir 90% dosen dan mahasiswa pernah terlibat didalamnya walaupun dalam berbagai platform seperti web, portal *e-learning*, blog ataupun facebook.

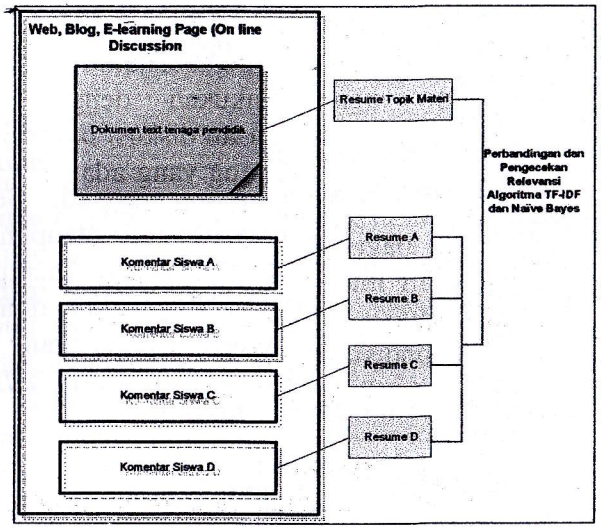
Untuk tahap ini satu (1) orang dosen dari masing-masing jurusan (5 jurusan) akan diwawancarai. Sementara jumlah mahasiswa yang diwawancarai untuk

tahap awal ini sebanyak 10 orang yang disebar proposional berdasarkan jumlah mahasiswa aktiv dari masing-masing jurusan. Adapun tujuan dari *face to face* interview (Creswell,2003) tersebut adalah untuk menggali berbagai informasi dan permasalahan yang muncul dalam pelaksanaan pembelajaran berbasis *online*. Observasi juga dilakukan dengan cara mempelajari bagaimana pelaksanaan diskusi *online* pada web, portal *e-learning* ataupun blog yang ada. Berbagai kondisi yang terlihat pada aktivitas web ataupun portal menjadi masukan bagi peneliti dalam memperkuat tinjauan pustaka dan interview yang sudah dilakukan. Analisa awal yang diperoleh digunakan untuk mengeneralisasi kondisi kegiatan diskusi *online* di lingkungan Universitas Islam Suska Riau.

2.2. Pembangunan Instrumen

Dari identifikasi masalah yang telah dirumuskan pada tahap 1, instrument yang akan digunakan dalam pembangunan dan perancangan aplikasi didefinisikan dengan lebih rinci. Berdasarkan teori, hasil interview dan observasi beberapa variabel yang mempengaruhi dalam proses *Contents Based Spam Filtering* dirumuskan dengan lebih tepat. Berbagai faktor yang menjadi puncak permasalahan dan kemungkinan solusi yang dapat digunakan diurai dengan lebih lengkap dalam pembangunan instrumen ini, seperti materi yang digunakan sebagai acuan resume ataupun proses filtering yang mungkin dilakukan maupun bentuk dan jenis karakter yang bisa digunakan sebagai standart filtering. Beberapa variabel yang digunakan pada Algoritma TF-IDF maupun Naïve Bayes sudah dapat didefinisikan

pada pembangunan instrumen ini. Logika konsep Content Based Spam Filtering dapat digambarkan pada Gambar 2.2 dibawah ini.



Gambar 2.2. Logika Konsep Content Based Spam Filtering

2.3. Studi Pendahuluan (Pilot Studi)

Pada tahap ini, sebuah studi pendahuluan berdasarkan variabel yang telah dirumuskan di instrumen akan diimplementasikan dalam bentuk prototipe dokumen *tracing* prosedur algoritma. Sebagai pilot studi tiga kasus *online discussion* yang terjadi pada web, blog ataupun *e-learning* dipelajari dan dibandingkan untuk mendapatkan hasil analisa yang lebih maksimal. Tabel perbandingan hasil berdasarkan variabel uji akan dibangun.

2.4. Analisa dan Perancangan Model

Dari hasil pilot studi yang telah dilakukan, berbagai analisa awal telah dapat dirumuskan untuk membangun dan merancang model aplikasi *Content Based Spam Filtering* berbasis Algoritma TF-IDF dan Naïve Bayes. Analisa terhadap kebutuhan sistem *(system requirement)* didefinisikan dalam bentuk dokumen kebutuhan sistem. Disini, berbagai fungsi aplikasi yang diperlukan di jelaskan legkap dengan logika proses yang mungkin akan terjadi. Pembangunan sistem dilakukan dengan menerapkan model *Propotyping* yang merupakan kombinasi dan penyempurnaan dari model *Waterfall* yang biasa digunakan. Analisis dan perancangan dilakukan dengan mengggunakan *Data Flow Diagram (DFD)*, *Entities Relationship Diagram* (ER Diagram), Database menggunakan *MySQL* dan bahasa pemograman yang digunakan adalah *PHP*. Perangkat Keras standar yang digunakan dalam pembangunan prototipe ini adalah Processor Celeron, Memory 4 GB, Harddisk berkapasitas 500 GB, Monitor dan Keyboard.

Untuk perancangan interface, tool standar yang digunakan adalah Adobe Photoshop dan Dreamweaver. Interface sistem dirancang untuk memperlihatkan secara jelas bagaimana proses tahapan yang dilakukan pada Algoritma TF-IDF dan Naïve Bayes. Pada aplikasi ini, Algoritma TF-IDF digunakan untuk menemukan resume pembicaraan berdasarkan komentar siswa terhadap uploadan materi yang disampaikan oleh tenaga pendidik. Selanjutnya, dengan menggunakan Algoritma Naive Bayes, tingkat relevansi dari hasil resume tersebut akan diukur untuk dikatagorikan dalam *Spam* atau tidak.

2.5. Implementasi dan Pengujian

Pengujian sistem dilakukan dengan 2 cara, yaitu pengujian Black Box dan User Accaptance Test (UAT). Pada pengujian Balck Box, serangkaian kondisi input dan output dalam sistem dengan berbagai persyaratan fungsional program akan diuji. Hal ini dilakukan untuk melihat apakah fungsi yang dijalankan sudah benar, adakah kesalahan interface, adakah kesalahan dalam struktur data dan akses database eksternal, adakah kesalahan kinerja sistem dan bagaimana proses inisialisasi dan kesalahan terminasi. Sementara itu, pengujian UAT dilakukan dengan melakukan survey kecil berkaiatan dengan penggunaan sistem. Software ini akan digunakan oleh administrator web, blog ataupun e-learning, tenaga pendidik dan siswa atau peserta.

2.6. Kesimpulan dan Saran

Pada tahap ini semua aktivitas yang dilakukan pada buku dirangkum untuk dibuat kesimpulan.Berbagai saran juga diajukan guna pengembangan buku ini selanjutnya.

BAB III
PEMBANGUNAN INSTRUMENT

Pada bab ini menjelaskan proses pembangunan instrument sebagai tools dalam operasional penelitian, analisa dan perancangan aplikasi.

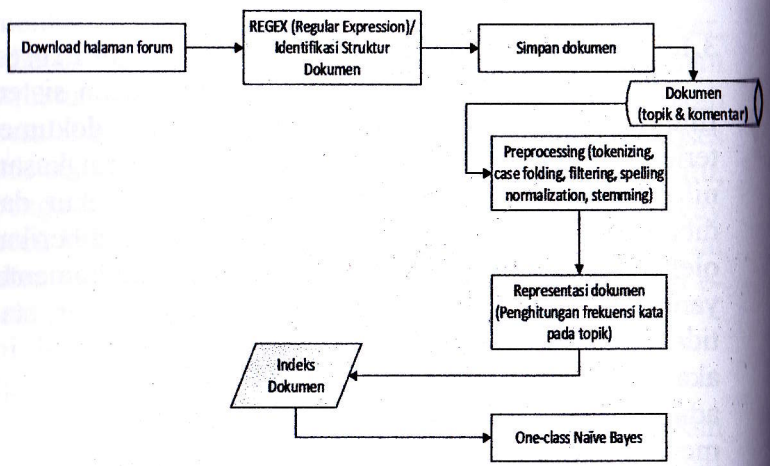
3.1. Rancangan Instrument dan Penelusuran Proses

Sistem yang akan dibangun ini merupakan sistem filtering dokumen melalui peringkasan materi dokumen terlebih dahulu. Setelah melewati proses peringkasan, nilai bobot relevansi materi (konten) akan diukur dan dibandingkan dengan setiap komentar yang diberikan oleh peserta, sehingga dapat ditentukan apakah komentar yang diberikan oleh peserta diskusi online relevan atau tidak terhadap topik (konten) yang diposting. Hal ini akan dijadikan sebagai bahan pertimbangan bagi administrator sistem ataupun pemateri untuk mengkatagorikan komentar tersebut kedalam spam atau bukan *spam* (*ham*).

Pengukuran *spam* dilakukan berdasarkan bobot terendah dari nilai keseluruhan perbandingan postingan dengan komentar. Selanjutnya administrator yang akan memutuskan apakah komentar tersebut termasuk *spam* atau bukan *spam* (*ham*). Ketika komentar tersebut dideteksi sebagai spam oleh administrator, maka komentar akan dihapus dan sistem akan memberikan pesan kepada peserta (si pemberi komentar) sebagai peringatan ataupun *warning*.

Tahapan diawali dengan phase *text preprocessing*, pembobotan *TF-IDF*, pembobotan *relevance query*, pembobotan *similarity* kalimat dan

pengelompokkan dengan menggunakan Algoritma Naïve Bayes. Dokumen yang diposting oleh administrator dalam diskusi online akan melewati beberapa tahapan text processing, meliputi pemecahan kalimat, *case folding, filtering, tokenizing* kata, *stemming* seperti pada Gambar 3.1.



Gambar 3.1. Flowchart Klasifikasi Relevansi Komentar dan Topik

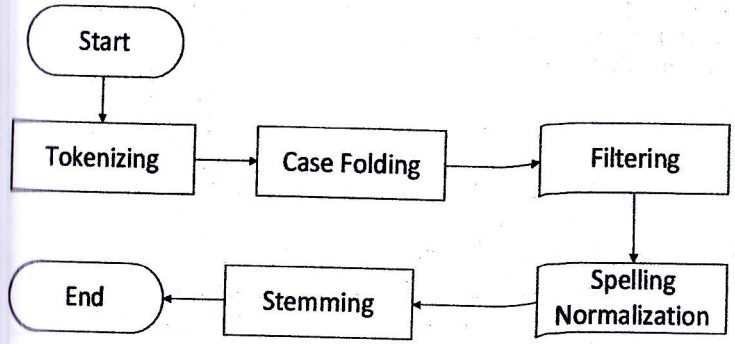
a. Download Halaman Forum dan Proses Identifikasi Struktur Dokumen (Parsing).

Proses download halaman forum akan menghasilkan file dengan format HTML. Selanjutnya file tersebut akan dilakukan proses *parsing*, yaitu identifikasi struktur dokumen html yang terdiri dari tag dan elemen tag untuk proses ekstraksi dokumen topik dan komentar yang terdapat dalam tag dan elemen dari dokumen html tersebut. Setelah proses ekstraksi, maka

dokumen akan disimpan dalam database untuk selanjutnya digunakan untuk proses *preprocessing*.

b. Preprocessing

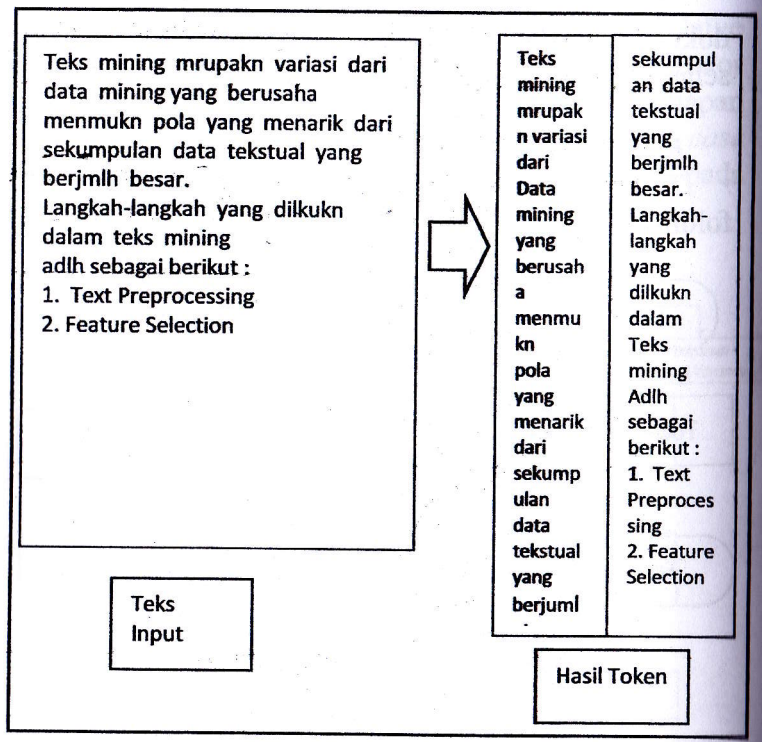
Tahapan preprocessing terdiri dari tokenizing, case folding, filtering, spelling normalization dan stemming.



Gambar 3.1. Preprocessing

c. Tokenizing

Dokumen hasil ekstraksi akan dipecah menjadi kata atau token. *Tokenization* dilakukan untuk mendapatkan token atau potongan kata yang akan menjadi entitas yang memiliki nilai dalam penyusunan matriks dokumen pada proses selanjutnya. Gambar 3.3 adalah ilustrasi dari pemecahan dokumen menjadi kata:



Gambar 3.3. Contoh Proses *Tokenization*

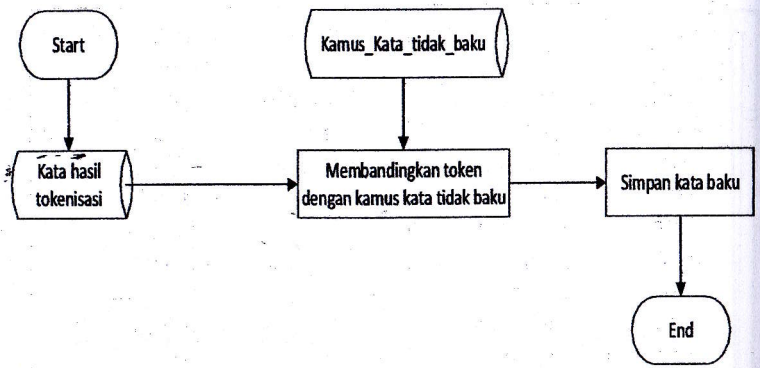
d. Case Folding

Pada tahapan *case folding* setiap huruf di dalam daftar token akan diseragamkan baik dengan mengubah menjadi huruf kecil semua atau huruf besar semua.

Gambar 3.4. Contoh Proses *Case Folding*

e. Spelling Normalization

Merupakan perbaikan dan subtitusi kata-kata yang salah eja ataupun disingkat dengan bentuk tertentu. Subtitusi kata dilakukan untuk menghindari jumlah perhitungan dimensi kata yang melebar. Perhitungan dimensi kata akan melebar jika kata yang salah eja atau disingkat tidak diubah karena kata tersebut sebenarnya memiliki kontribusi dalam merepresentasikan dokumen tetapi akan dianggap sebagai entitas yang berbeda proses penyusunan matriks.

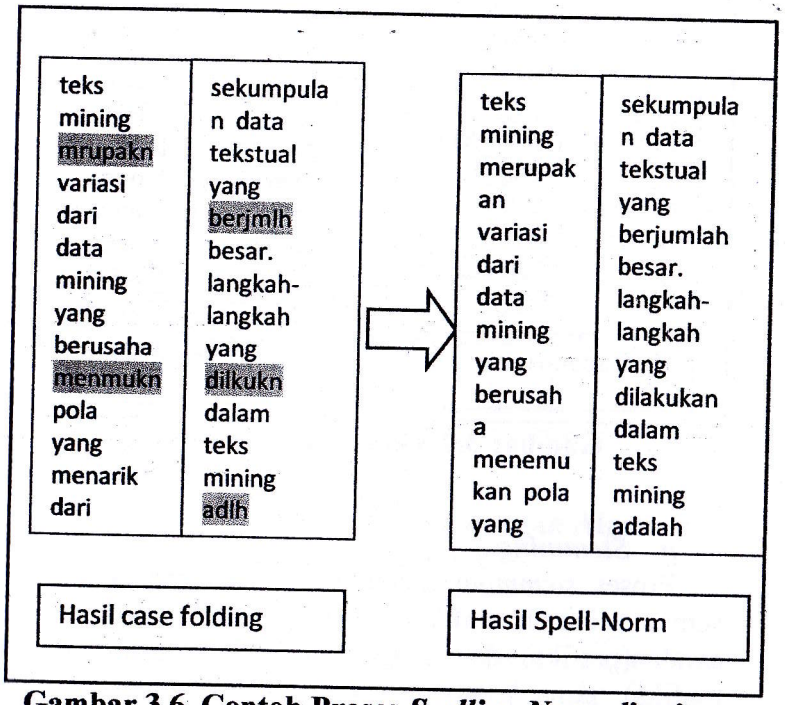


Gambar 3.5. Flowchart Spelling Normalization

Tabel 3.1. Daftar Kamus Kata Tidak Baku

Ad~ada	Blum~belum	Menjdi~menjadi
Adany~adanya	Blz~balas	Mnjd~menjadi
Adek~adik	Bnar~benar	Mnjdi~menjadi
Adlh~adalah	Bndg~bandung	Mnjdikan~menjadikan
aer~air	Bner~benar	Mnrima~menerima
aj~saja	Bnget~sangat	Mnrm~menerima
Aja~saja	Bngkus~bungkus	Mnrma~menerima
Ajah~saja	Bngung~bingung	Mnrt~menurut
Ak~aku	Bngun~bangun	Mnta~minta
Aq~aku	Bngunan~bangunan	Mnunjukan~menjunjukkan
Akn~akan	Bnjr~banjir	Mnum~minum
Alesan~alasan	Prhtungan~perhitungan	Mo~mau
Alamt~alamat	Barang2~barang-barang	Moga~semoga

Kndala~kendala		Mtor~motor
Sy~saya		Mnerapkn=menerapkan



Gambar 3.6. Contoh Proses *Spelling Normalization*

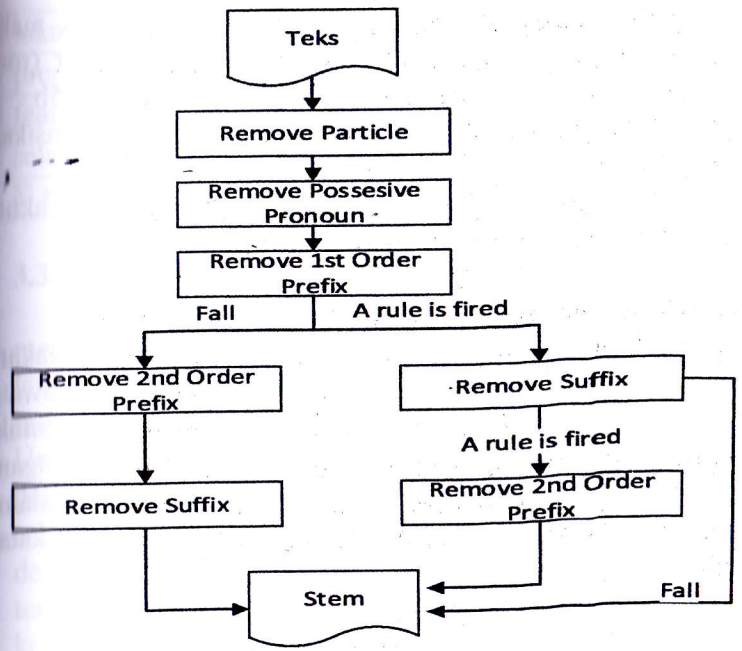
f. Filtering

Pada tahap *filtering* , kata, url, *hashtag(#)* maupun tanda baca tertentu yang tidak berkontribusi secara signifikan dalam merepresentasikan dokumen akan disaring menggunakan algoritma *stop-word*dan *stop-character*.

Gambar 3.7. Contoh Hasil *Filtering*

g. Stemming

Proses *stemming* dilakukan dengan menghilangkan semua imbuhan(afiks) baik yang terdiri dari awalan(prefiks) dan sisipan (infiks) maupun akhiran (sufiks) dan kombinasi dari awalan dan akhiran (konfiks). *Stemming* digunakan untuk mengganti bentuk dari suatu kata menjadi kata dasar sesuai dengan struktur morfologi bahasa Indonesia yang baik dan benar. Pengambilan akar kata dilakukan untuk mengurangi dimensi matriks yang dihasilkan oleh kemunculan entitas berbeda yang sebenarnya mempunyai akar kata yang sama.



Gambar 3.8. Proses *Stemming* dengan Algortima Porter Stemmer

1. Menghapus partikel seperti: -kah, -lah, -tah
2. Mengapus kata ganti (Possesive Pronoun), seperti –ku, -mu, -nya
3. Mengapus awalan pertama. Jika tidak ditemukan, maka lanjut ke langkah 4a, dan jika ada maka lanjut ke langkah 4b.
4. a. Menghapus awalan kedua, dan dilanjutkan pada langkah ke 5a

b.Menghapus akhiran, jika tidak ditemukan maka kata tersebut diasumsikan sebagai kata dasar (root word). Jika ditemukan maka lanjut ke langkah 5b.

5. a. Menghapus akhiran dan kata akhir diasumsikan sebagai kata dasar (root word).

b. Menghapus awalan kedua dan kata akhir diasumsikan sebagai kata dasar (root word).

3.2. Representasi Dokumen / Pembobotan tf-df

Representasi dokumen/ pembobotan digunakan untuk mentransformasi data teks yang telah melewati preprocessing menjadi data numerik ke dalam bentuk matriks. Proses transformasi data dilakukan dengan menghitung frekuensi kemuculan kata dalam dokumen(tf) dan frekuensi dokumen yang memuat kata (df).

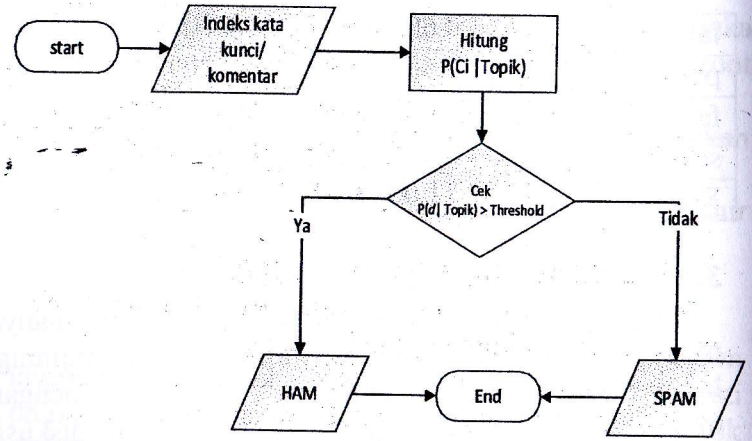
Tabel 3.2. Pembobotan tf-df

term	tf	w	idf	df
teks	4	0,49975495	0,124938737	3
mining	3	0,90308999	0,301029996	2
variasi	1	0,60205999	0,602059991	1
data	2	1,20411998	0,602059991	1
usaha	1	0,60205999	0,602059991	1
temu	1	0,60205999	0,602059991	1
pola	1	0,60205999	0,602059991	1
tarik	1	0,60205999	0,602059991	1
kumpul	1	0,60205999	0,602059991	1
jumlah	1	0,60205999	0,602059991	1
besar	1	0,60205999	0,602059991	1
langkah	2	1,20411998	0,602059991	1
laku	1	0,60205999	0,602059991	1
preprocessing	1	0,60205999	0,602059991	1
feature	1	0,60205999	0,602059991	1
selection	1	0,60205999	0,602059991	1
$\sum$ tf	23	11,0358048		

3.3. Klasifikasi One-Class Naive Bayes

One-Class Naive Bayes pada penelitan ini hanya menggunakan dokumen topik sebagai data training, langkah awal dalam proses klasifikasi dengan menghitung P(C_i|Topik) untuk setiap topik diskusi. Untuk komentar yang tidak terkait dengan topik, diasumsikan memiliki probabilitas 1/m. Jika diberikan sebuah data test D, dapat dibandingkan P(C_i|Topik) dengan P(C_i|~Topik), semakin besar rasio P(C_i|Topik) terhadap P(C_i|~Topik), menujukkan bahwa Komentar ke-i relevan terhadap topik.

Menerapkan *One-Class Naive Bayes* pada dataset yang spesifik juga sangat sederhana. Ketika setiap dokumen test (komentar) terdapat 100 data, maka probabilitas komentar yang tidak relevan terhadap topik adalah sama untuk setiap data test. jadi tidak diperlukan lagi penghitungan untuk probabilitas komentar yang tidak relevan. Penghitungan komentar yang relevan terhadap topik dapat digunakan sebuah threshold untuk memutuskan apakah komentar ini relevan atau tidak. Secara umum proses *One-Class Naive Bayes* dijelaskan pada gambar 3.9.



Gambar 3.9. Proses Klasifikasi dengan *One-Class Naive Bayes*

a. Menghitung Probabilitas Kata Terhadap Topik

$P(C_i|\text{Topik})$ adalah probabilitas kata C_i dalam kelas topik dalam data pelatihan, maka dapat dihitung $P(d|\text{Topik})$ pada dokumen test *d*.

Tabel 3.3. Nilai Probabilitas Setiap Kata Terhadap Kelas Topik

term	tf	w	idf	df	$P(C_i\|Topik)$
teks	4	0,49975495	0,124938737	3	0,128205
mining	3	0,90308999	0,301029996	2	0,102564
variasi	1	0,60205999	0,602059991	1	0,051282
data	2	1,20411998	0,602059991	1	0,076923
usaha	1	0,60205999	0,602059991	1	0,051282
temu	1	0,60205999	0,602059991	1	0,051282
pola	1	0,60205999	0,602059991	1	0,051282
tarik	1	0,60205999	0,602059991	1	0,051282
kumpul	1	0,60205999	0,602059991	1	0,051282

jumlah	1	0,60205999	0,602059991	1	0,051282
besar	1	0,60205999	0,602059991	1	0,051282
langkah	2	1,20411998	0,602059991	1	0,076923
laku	1	0,60205999	0,602059991	1	0,051282
preprocessing	1	0,60205999	0,602059991	1	0,051282
feature	1	0,60205999	0,602059991	1	0,051282
selection	1	0,60205999	0,602059991	1	0,051282

b. Menghitung Probabilitas Relevansi Komentar Terhadap Topik

$P(C_i|\text{Topik})$ adalah probabilitas kata C_i dalam kelas topik dalam data pelatihan, dalam hal ini data test adalah komentar partisipan yang terlibat pada diskusi.$P(C_i|\text{Topik})$ digunakan untuk menghitung $P(d|\text{Topik})$ pada dokumen test *d*atau dokumen komentar. Untuk komentar yang tidak terkait dengan topik, diasumsikan memiliki probabilitas 1/m, m adalah jumlah total frekuensi kata pada kelas topik.

Tabel 3.4. Probabilitas Relevansi Komentar Terhadap Topik

	Term	Topik(mn)
	hubung	0,025641
	teks	0,128205
$P(C_i\|Topik)$	mining	0,010519
	data	0,076923
	Π	2,66E-06
	Σ	0,241289

Untuk menghitung relevansi komentar terhadap topik dengan menggunakan klasifikasi *One-Class Naive Bayes,* komentar akan dikatakan relevan jika probabilitas komentar terhadap topik melebihi nilai *threshold* (λ) yang telah ditentukan. Misalnya jika nilai *threshold* yang diambil adalah 30% dari total nilai probabilitas topik, maka komentar diatas dikatakan tidak relevan, karena kurang dari nilai *threshold.*

BAB IV
ANALISA DAN PERANCANGAN

Pada Bab ini akan dijelaskan analisa dan perancangan yang telah dilakukan dalam penelitian dibuku ini.

4.1. Analisa Sistem

Analisa sistem awal diperoleh berdasarkan hasil pilot studi melalui wawancara terhadap beberapa orang dosen sebagai tenaga pengajar berbasis on-line dan beberapa orang mahasiswa sebagai peserta diskusi online dilingkugan Fakultas Sains dan Teknologi. Pilot studi dilakukan dengan menggunakan tiga bentuk media diskusi online yaitu Web, Blog dan E-Learning. Dengan menggunakan teknik Summarization hasil wawancara disimpulkan.

Dari hasil analisis diperoleh beberapa variable yang perlu menjadi pertimbangan dalam melakukan diskusi online, yaitu:

a. Konflik dalam Diskusi

Konflik dalam diskusi online sering sekali terjadi dalam berbagai bentuk apakah dengan mengeluarkan kata-kata yang tidak sopan ataupun melalui konflik ide, pemahaman ataupun saran yang diberikan. Hal ini dipicu juga oleh suatu kondisi dimana peserta diskusi online bersifat *"invisible"* tidak tampak wujud fisiknya, mimik wajahnya ataupun intonasi suara pada saat melakukan diskusi. Hal ini ternyata cukup mempengaruhi suasana diskusi online yang dilakukan. Guna mengatasi hal tersebut, pemahaman akan pentingnya sebuah diskusi yang sehat, pengontrolan emosional, pendekatan yang dapat digunakan dalam menyampaikan ide di forum

online perlu di tekankan oleh administrator ataupun tenaga pengajar sebelum membuka forum diskusi ataupun selama proses diskusi berlangsung. Artinya kehadiran dan peranan Administrator ataupun tenaga pengajar adalah sangat penting guna memastikan diskusi tersebut berjalan dengan baik dan positive. Apabila ada hal yang perlu didiskusikan secara lebih mendalam peserta bisa menghubungi peserta lainnya secara personal.

b. Munculnya Penyerangan dan Pembulian Personal

Seringkali kasus pembulian terhadap peserta diskusi online terjadi. Hal tersebut bisa saja disebabkan karena perbedaan pemahaman, kepercayaan, ataupun masalah pribadi lainnya. Dimana peserta tersebut akan diserang oleh sekelompok peserta lainnya dalam bentuk komentar-komentar yang tidak sehat yang bisa menjatuhkan mental dari peserta yang dibuli. Tentunya diskusi yang seperti ini sudah mengarah kepada diskusi yang negative dan bisa menimbilkan trauma hingga menyebabkan peserta keluar dari forum diskusi tersebut. Guna mengatasi hal ini, peranan Administrator ataupun Tenaga Pengajar sangat penting untuk tidak membiarkan proses pembulian tersebut terus berlangsung baik melalui teguran personal hingga kepada hukuman. Aturan main selama diskusi online perlu dibuat dan dipatuhi oleh semua peserta diskusi.

c. Keengganan Peserta untuk Terlibat Aktiv dalam Diskusi

Hal ini timbul pada beberapa peserta yang merasa malu untuk terlibat dalam diskusi online, adanya perasaan

minder untuk bergabung dalam diskusi terbuka ataupun merasa kemampuannya kurang saat melihat posting peserta lainnya yang bagus sehingga mengakibatkan terbatasnya peserta yang terlibat aktive dan kecendrungan peserta yang itu-itu saja. Guna mengatasi hal tersebut Administrator dan Tenaga Pengajar mampu memotivasi peserta baik melalui pemberian *reward* ataupun pujian penyemangat atas setiap posting yang dilakukannya. Sementara bagi peserta yang masih jarang terlibat dalam diskusi, perlu dilakukan pendekatan baik melalui email ataupun langsung guna menunjukkan perhatian sebagai wujud pernyataan pentingnya diskusi online ini.

d. Adanya *plagarizm* Posting

Seringkali dalam diskusi online terjadinya plagarizm postingan, yang diambil dari postingan orang lain baik dari dalam diskusi itu sendiri maupun dari blog ataupun web site yang lain. Hal ini tentunya akan memberikan dampak yang tidak baik, baik terhadap peerta itu sendiri maupun peserta yang lain. Disini, sebagai Administrator ataupun tenaga pengajar perlu untuk melakukan teguran dan penekanan bahwa *cheating* tetap bukanlah hal yang baik terutama apabila hanya untuk mengharap nilai yang tinggi. Hal ini bisa diatasi dengan cara menghubungi peserta tersebut secara personal untuk dinasehati betapa buruknya pengaruh plagarizm.

e. Keluar dari Topik Diskusi

Meskipun dari tahap awal topik diskusi sudah dijelaskan, namun seringkali diskusi keluar dari pembahasan. Tentunya hal ini memerlukan proses pengontrolan yang

baik terhadap setiap posting yang dilontarkan oleh peserta. Hal ini dapat dilakukan oleh Tenaga Pengajar dengan memberikan komentar atau postingan yang merangsang ataupun memberikan informasi yang lebih mendalam dari setiap postingan yang sudah diajukan oleh peserta. Sehingga proses pembelajaran menjadi lebih hidup dan luas.

f. Postingan Peserta yang Sangat Minim

Hal ini sering terjadi dalam diskusi online, dimana peserta hanya berkomentar "Ya, Saya Setuju" atau "Mantap" yang tidak memberikan pengaruh yang besar terhadap isi diskusi dan pembahasan. Hal ini juga akan mempengaruhi peserta lainnya untuk tidak terlibat aktif karena kurang merasakan manfaat dari forum diskusi tersebut. Guna mengatasi hal tersebut, tenaga pengajar dapat memberikan contoh bagaimana memberikan komentar yang berbobot dari sekedar kalimat singkat yang kurang bermakna.

Dari ke-6 variabel diatas, menjadi pijakan dasar dalam merancang dan membangun sebuah diskusi online yang baik dan sehat. Sehingga apa yang menjadi tujuan dan harapan dalam forum tersebut dapat tercapai. Kondisi pilot studi ini yang tampak dilingkungan Fakultas Sains dan Teknologi yang dapat mengeneralisasikan lingkungan Universitas Islam Suska Riau. Khusus pada buku ini, filtering dilakukan pada diskusi yang tidak sehat dikarenakan ketidaksinkronan dengan materi yang dijabarkan dalam postingan diskusi baik dalam bentuk postingan kalimat yang tidak sopan, konflik ide yang sudah keluar dari area pembahasan dan pembulian yang

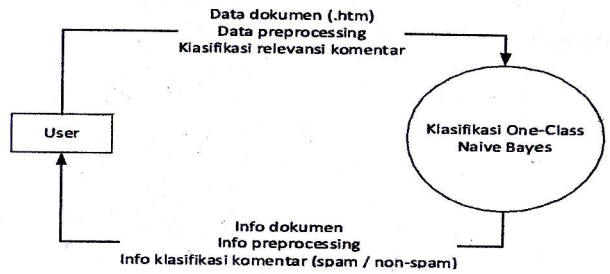
keluar dari konteks diskusi. Namun komentar yang minim dan pengecekan plagarizm tidak dibahas dalam bentuk rancangan sistem pada buku ini.

4.2. Perancangan Sistem

Rancangan sistem dibangun berdasarkan informasi yang diperoleh dari analisis awal hingga proses aplikasi instrument terhadap proses filtering dengan konsep TF-IDF dan Naïve Bayes. Tahapan perancangan terbagi dalam beberapa bagian, diantaranya adalah perancangan Diagram konteks (*Context Diagram*), perancangan DFD (*Data Flow Diagram*), perancangan ERD (*Entity Relation Diagram*), Perancangan menu utama sistem.

4.3. Context Diagram

Context Diagram digunakan untuk menggambarkan proses kerja sistem secara umum atau gambaran operasional sistem secara garis besar.



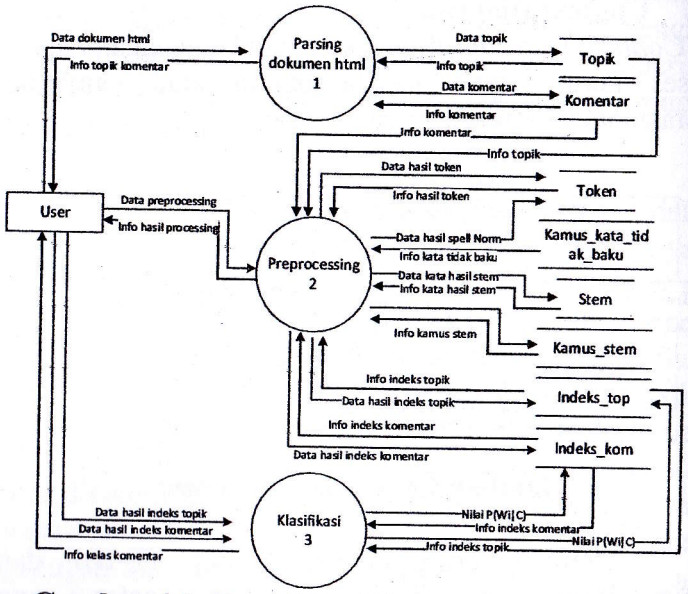
Gambar 4.1. *Context Diagram*

Sistem Filtering dengan konsep TF-IDF dan Naïve Bayes memiliki satu orang actor yang

dikatagorina kedalam user. User menginputkan data dokumen dalam bentuk .htm ke dalam sistem. Dokumen dapat berupa diskusi online yang dilakukan di blog, website ataupun e-learning untuk diproses dan diklasifikasikan berdasarkan relevansinya terhadap topik pembelajaran yang disampaikan. Sistem akan melakukan serangkaian proses dan outputnya berupa informasi klasifikasi komentar berdasarkan relevansinya terhadap topik. Sistem akan menentukan apakah komentar ini berada dalam katagori spam atau non spam.

4.3.1. Data Flow Diagram Level 1

Gambar DFD level 1 proses klasifikasi relevansi komentar partisipan terhadap topik.

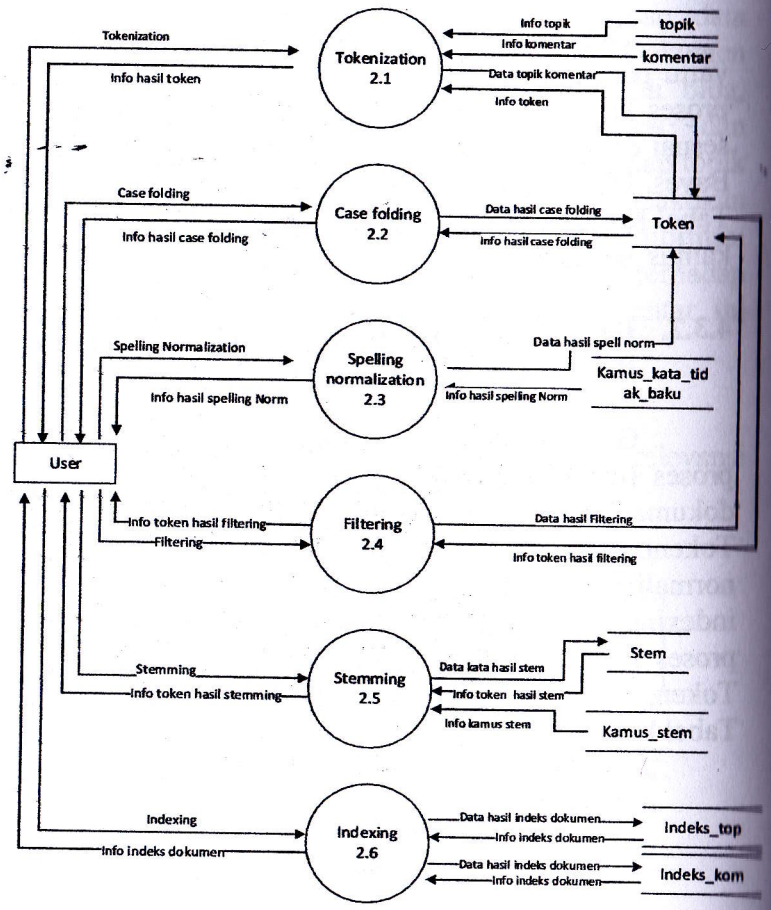


Gambar 4.2. *Data Flow Diagram I* (DFD) Level 1

DFD Level I terdiri dari beberapa proses utama yaitu parsing dokumen html (1); preprocessing (2) dan proses klasifikasi (3). Masing-masing proses dijabarkan sesuai dengan alur penerapan proses TF IDF dan Naïve Bayes. Database yang terlibat disini adalah Tabel Topik, Tabel Komentar, Token, Kamus_kata_tidak_baku, stem, kamus_stem, indeks_top dan indeks_kom.

4.3.2. Data Flow Diagram Level 2 Proses 2 Preprocessing

Gambar DFD level 2 proses 2 merupakan aliran proses secara rinci dari proses *preprocessing* terhadap dokumen sebelum proses klasifikasi terdiri dari proses Tokenization (2.1), Case folding (2.2), spelling normalization (2.3), filtering (2.4), stemming (2.5) dan indexing (2.6). Beberapa database yang terlibat pada proses ini adalah Tabel Topik, Tabel komentar, Tabel Token, Tabel Kamus_data_tidak_baku, Tabel stem, Tabel kamus_stem, Tabel indeks_top dan Indeks_kom.

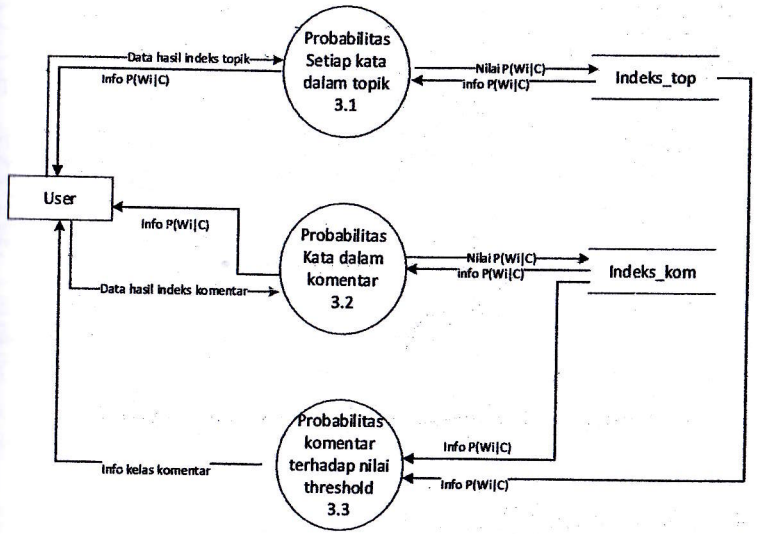


Gambar 4.3. DFD Level 2 Proses 2 *Preprocessing*

4.3.3. Data Flow Diagram Level 2 Proses 3 Klasifikasi

Gambar DFD level 2 proses 3 merupakan rancangan proses secara rinci dari klasifikasi relevansi komentar partisipan terhadap topik. Terdiri dari proses

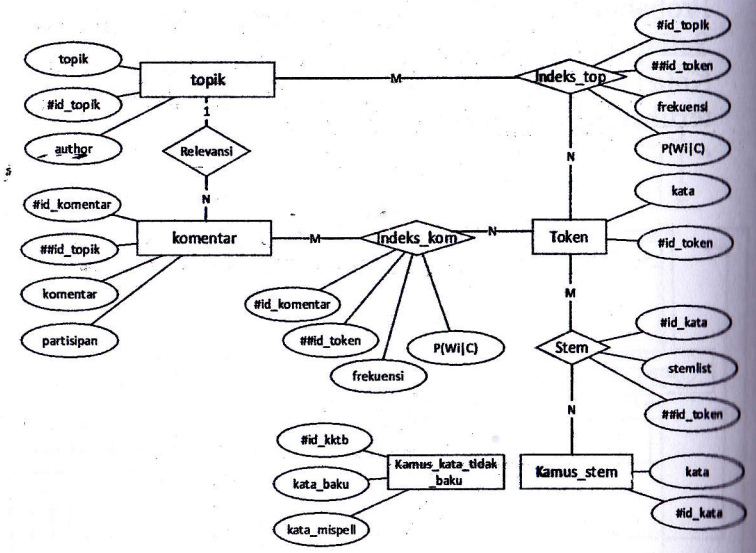
Probabilitas Setiap Kata dalam Topik (3.1), Probabilitas Setiap Kata dalam Komentar (3.2), Probabilitas Setiap Kata terhadap nilai threshold (3.3). Database yang terlibat meliputi indeks_top, indeks_kom.



Gambar 4.4. DFD Level 2 Proses 3 Klasifikasi

4.4. ERD (*Entity Relation Diagram*)

Gambar 4.5 merupakan diagaram relasai antar entitas dalam sistem, yang menggambarkan keterhubungan antar tabel pada database untuk memenuhi kebutuhan proses sistem.



Gambar 4.5 *Entity Relation Diagram* (ERD)

4.4.1. Struktur *Database*

a. Topik

Nama tabel : topik
Deskripsi isi : memuat isi topik diskusi.
Primary key : id_topik

Tabel 4.1. Data topik_komentar

NO	Nama Field	*Type & Length*	Deskripsi	Null	Default
1	Id_topik	*Varchar(30)*	Kode topik	Not	PK
2	author	*varchar(15)*	Pembuat topik	Not	
3	topik	*Text*	Isi topik	Not	

b. Komentar

Nama tabel : komentar
Deskripsi isi : memuat data komentar partisipan, nama partisipan berdasarkan topik bahasan.
Primary key :id_komentar

Tabel 4.2 Data Komentar

NO	Nama Field	*Type & Length*	Deskripsi	Null	Default
1	Id_topik	*Varchar(15)*	Kode topik	Not	FK
2	Id_komentar	*Varchar(15)*	Kode komentar	Not	PK
3	komentar	*Text*	Isi komentar	Not	
4	partisipan	*Varchar(15)*	Peserta diskusi	Not	

c. Token

Nama tabel : token
Deskripsi isi dan komentar : memuat token/ kata dari topik
Primary key : id_token

Tabel 4.3 Data Token

NO	Nama Field	*Type & Length*	Deskripsi	Null	Default
2	Id_token	*varchar(15)*	Kode token	Not	PK
3	kata	*Varchar(30)*	Kata/ token	Not	

d. Kamus Kata Tidak Baku

Nama tabel : Kamus_kata_tidak_baku
Deskripsi isi : Daftar kata salah eja dan kata bakunya.
Primary key : id_kktb

Tabel 4.4 Data kamus_kata_tidak_baku

NO	Nama Field	*Type & Length*	Deskripsi	Null	Default
1	Id_kktb	*Varchar(15)*	Kode kata tidak baku	Not	PK
2	Kata_mispell	*varchar(15)*	Kata salah eja/ tidak baku	Not	
3	Kata baku	*Varchar(30)*	Kata baku	Not	

e. Kamus Stem

Nama tabel : kamus_stem
Deskripsi isi : berisi kata dasar, kata dengan partikel tertentu sebagai pelengkap proses *stemming.*
Primary key : id_kata

Tabel 4.5 Data kamus_stem

NO	Nama Field	*Type & Length*	Deskripsi	Null	Default
1	Id_kata	*Varchar(15)*	Kode kata dasar	Not	PK
2	Kata	*varchar(30)*	Kata dasar dan kata berpartikel/ berimbuhan	Not	

Tabel 4.6 Data stem

NO	Nama Field	*Type & Length*	Deskripsi	Null	Default
1	Id_kata	*Varchar(15)*	Kode kata dasar	Not	PK
2	Id_token	*Varchar(15)*	Kode token		
3	stemlist	*varchar(30)*	Kata dasar dan kata berpartikel/ berimbuhan	Not	

f. Indeks topik

Nama tabel : indeks_top
Deskripsi isi : memuat indeks kata dari topik, frekuensi kemunculan kata, probabilitas kata terhadap topik.
Primary key : id_topik

Tabel 4.7 Data indeks_top

NO	Nama Field	Type & Length	Deskripsi	Null	Default
1	Id_topik	*Varchar(15)*	Kode topik	Not	PK
2	Id_token	*varchar(15)*	Kode token	Not	FK
3	frekuensi	*Int(11)*	Frekuensi kemunculan kata/ token	Not	
4	probability	*Float*	Probabilitas kemunculan token dalam topik	Not	

g. Indeks Komentar

Nama tabel : indeks_kom
Deskripsi isi : memuat indeks kata dari komentar, frekuensi kata tiap-tiap komentar, dan probabilitas kata komentar terhadap topik.
Primary key : id_komentar

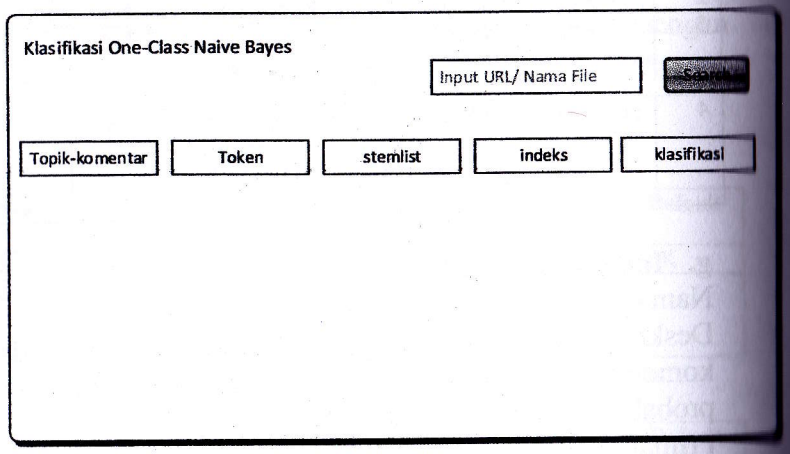
Tabel 4.8 Data indeks kata komentar

NO	Nama Field	Type & Length	Deskripsi	Null	Default
1	Id_komentar	*Varchar(15)*	Kode topik	Not	PK
2	Id_token	*varchar(15)*	Kode token	Not	FK
3	frekuensi	*Int(11)*	Frekuensi kemunculan kata/ token	Not	
4	probability	*Float*	Probabilitas kemunculan	Not	

			token dalam topik		

4.5. Tampilan Menu Sistem

Gambar 4.6 adalah rancangan tampilan menu utama sistem, yang digunakan sebagai acuan implementasi pada sistem yang akan dibangun. Sistem memiliki beberapa fungsi diantaranya adalah menu Topik komentar, menu token, menu stemlist, menu indeks, dan menu klasifikasi.



Gambar 4.6 Tampilan Utama Sistem

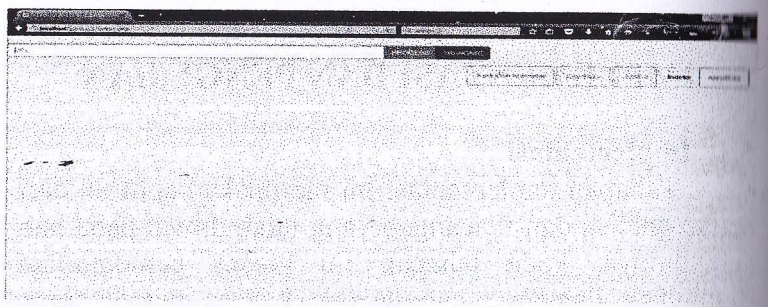
BAB V
IMPLEMENTASI DAN PENGUJIAN

5.1. Implementasi

Tahapan implementasi ini merupakan aplikasi dari setiap analisa dan rancangan yang telah dibuat pada bab sebelumnya. Pada tahapan ini proses pemograman digunakan dengan menggunakan bahasa pemograman PHP dan database mySQL. Beberapa tampilan menu sistem dapat dijabarkan dibawah ini.

5.1.1. Tampilan menu utama sistem

Seperti dijelaskan pada bab sebelumnya sistem ini memiliki beberapa menu fungsi, yaitu menu topic dan komentar yang berfungsi untuk menampilkan topic dan komenta-komentar yang akan diklasifikasikan; menu tokenisasi adalah menu yang berfungsi untuk menampilkan proses tokenisasi yang dilakukan terhadap sampel; menu stemlist berfungsi untuk menampilkan proses stemlist pada sistem; menu indeks berfungsi untuk menampilkan hasil proses indeksing yang telah dilakukan oleh sistem; dan terakhir adalah menu klasifikasi sebagai menu yang menampilkan hasil proses klasifikasi sistem yang menunjukkan tingkat relevansi komentar terhadap topik yang ada serta klasifikasi spam dan non-spam.



Gambar 5.1 Tampilan Utama Sistem

5.2. Pengujian

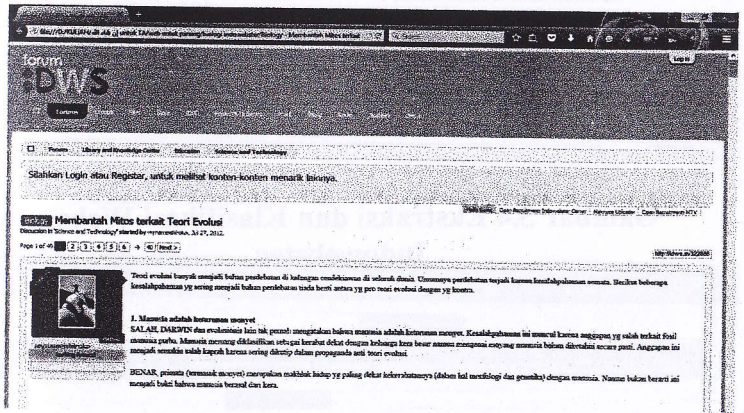
Tujuan dari pengujian adalah mencari kesalahan atau *error* dari sistem yang telah dibangun serta mengukur akurasi dari sistem dalam mengklasifikasi relevansi komentar terhadap topik pada beberapa website, blog ataupun e-learning yang diuji coba.

5.2.1. Pengujian akurasi sistem

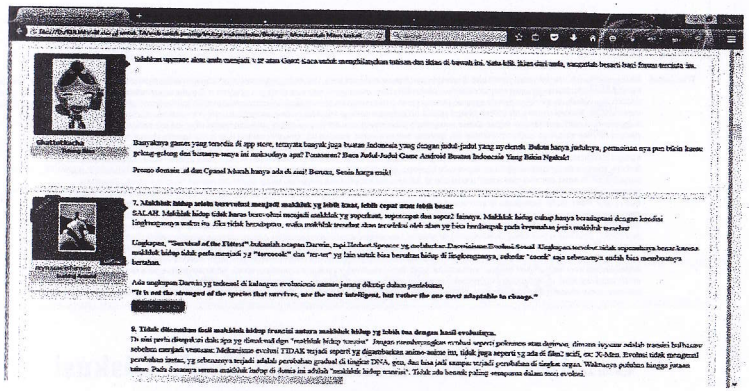
Pengujian akurasi sistem adalah pengukuran kemampuan sistem dalam mengklasifikasi komentar berdasarkan relevansi komentar terhadap topik yang dibahas pada website/blog/e-learning/forum yang diekstrak. Nilai akurasi diperoleh dari perbandingan antara jumlah dokumen yang relevan berdasarkan klasifikasi secara manual dan jumlah dokumen yang berhasil diklasifikasi oleh sistem. Sebagai sampel, pengujian buku ini dilakukan pada dua *website* forum diskusi *online* yaitu Forum Indowebster dan website Rumaysho.com.

1. Pengujian pada Website Indowebster

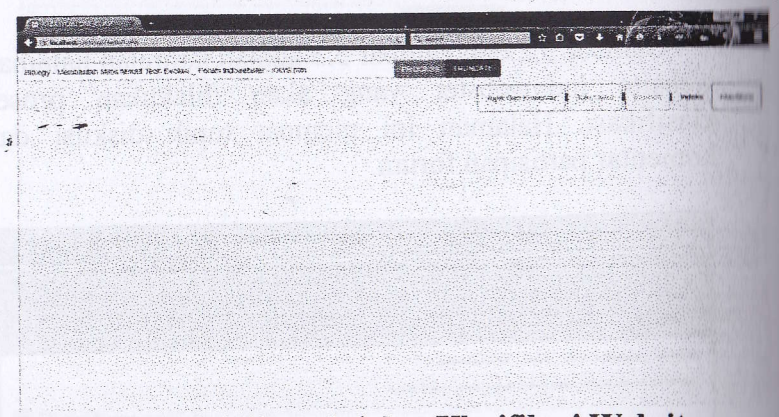
Gambar 5.2 dan gambar 5.3 adalah tampilan halaman website indowebster yang akan dilakukan proses klasifikasi. Halaman ini memuat topik diskusi dan komentar partisipan forum.



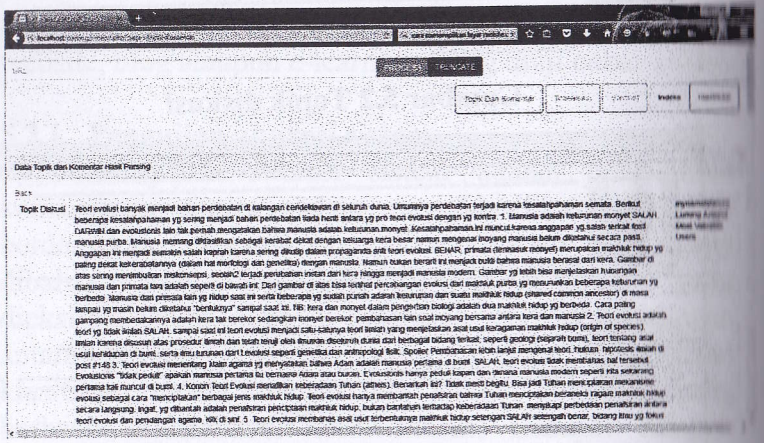
Gambar 5.2 Topik diskusi



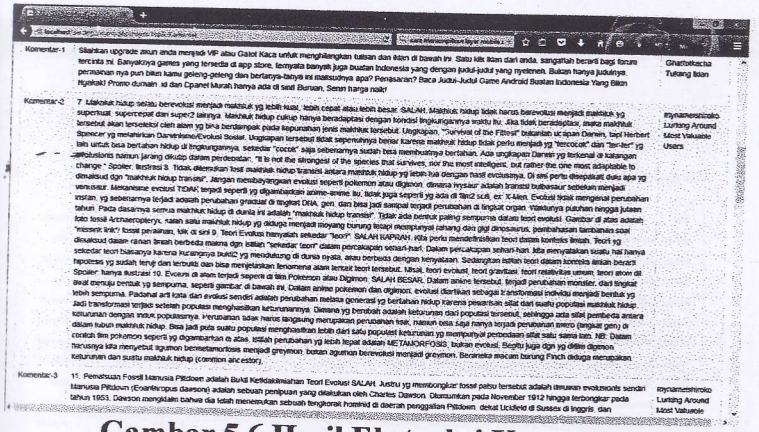
Gambar 5.3 Komentar Partisipan



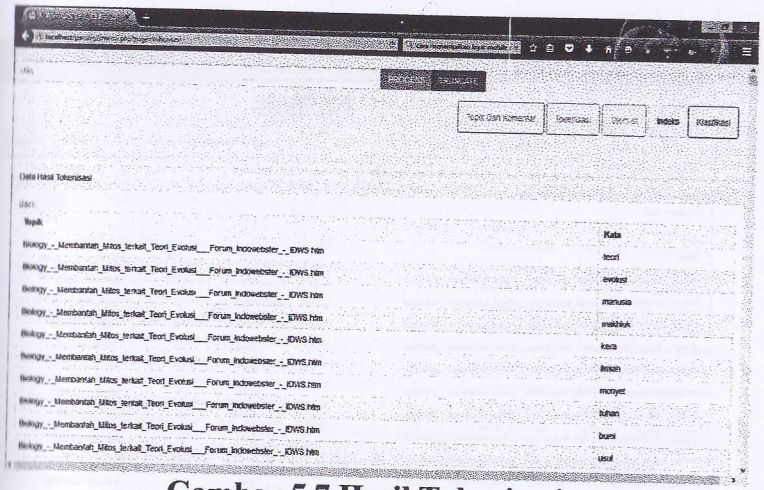
Gambar 5.4 Ekstraksi dan Klasifikasi Website Indowebster



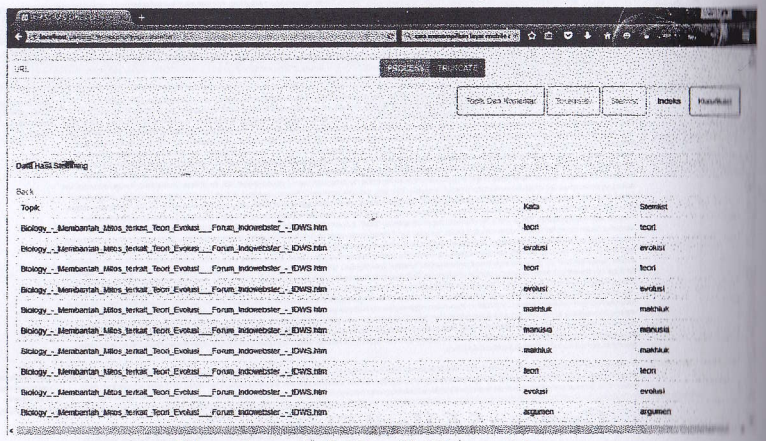
Gambar 5.5 Hasil Ekstraksi Topik Diskusi



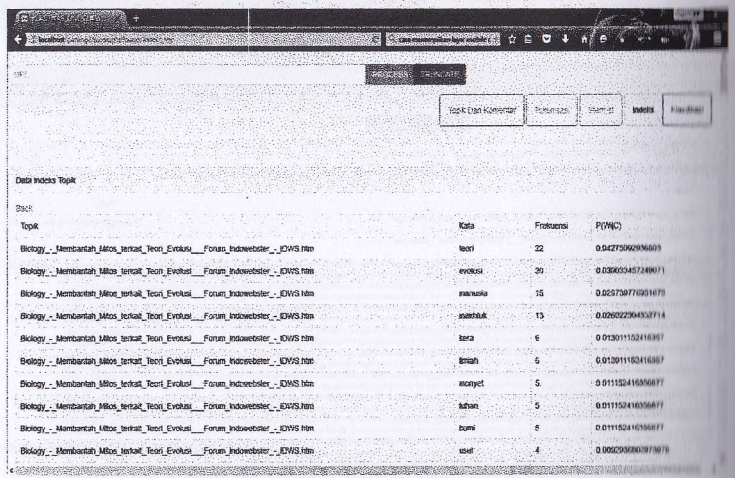
Gambar 5.6 Hasil Ekstraksi Komentar



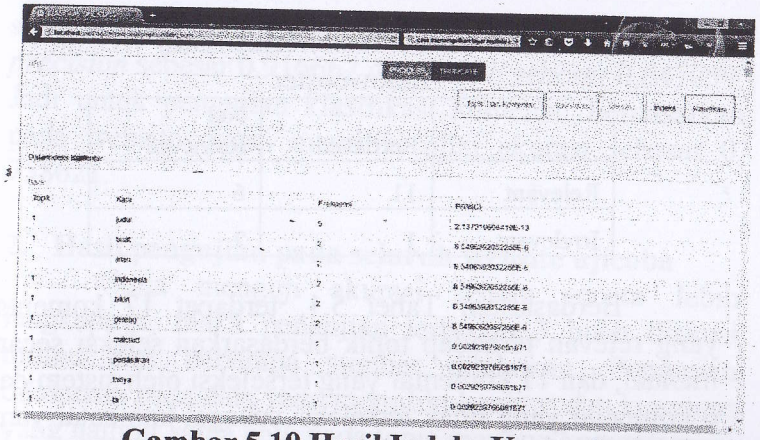
Gambar 5.7 Hasil Tokenisasi



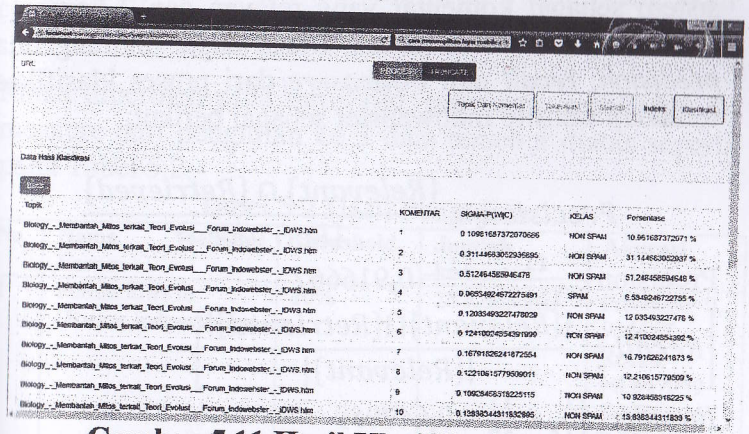
Gambar 5.8 Hasil Stemming



Gambar 5.9 Hasil Indeks Topik



Gambar 5.10 Hasil Indeks Komentar



Gambar 5.11 Hasil Klasifikasi Komentar

Dari hasil klasifikasi komentar diatas diperoleh hasil komentar yang relevan terhadap topik dan tidak relevan seperti tertera pada Tabel 5.1.

Tabel 5.1 Matrix Confusion Klasifikasi Website Indowebster

	Retrieved	Not Retrieved
Relevant	11	6
Irrelevant	1	2

Berdasarkan Tabel 5.1, terdapat 17 komentar yang relevan terhadap topik berdasarkan seleksi secara manual, dan 11 komentar yang terseleksi oleh sistem dan 6 komentar yang tidak terseleksi sebagai komentar yang relevan. Dari 3 komentar yang tidak relevan berdasarkan seleksi manual, terdapat 1 komentar yang terseleksi oleh sistem sebagai komentar yang relevan dan 2 komentar tidak terseleksi sama sekali. Maka dapat dihitung nilai *precision*, *recall* dan F-*Measure* dari proses klasifikasi terhadap website Indowebster sebagai berikut:

$$Precision = \frac{\{Relevant\} \cap \{Retrieved\}}{\{Retrieved\}}$$

$$\text{Precision} = \frac{11}{11+1} = \frac{11}{12} = 0{,}916667$$

$$\left[Recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}\right]$$

$$Recall = \frac{11}{11+6} = \frac{11}{17} = 0{,}647059$$

$$\text{F-Measure} = \frac{2\,xRxP}{R+P}$$

$$F\text{-}Measure = \frac{2\,x0{,}647059\,x0{,}916667}{0{,}647059 + 0{,}916667} = 0{,}758621$$

Hasil klasifikasi diatas diperoleh dengan menetapkan nilai *threshold* 10% dari rata-rata probabilitas topic yang menunjukkan hasil yang

signifikan. Sementara itu, nilai yang diperoleh pada *precision* sebesar 0,916667 dan *Recall* sebesar 0,647059. Jadi untuk pengujian klasifikasi komentar berdasarkan pada website diatas memiliki nilai akurasi sebesar 0,758621.

2. Hasil pengujian pada seluruh website ujicoba

Guna menguji akurasi sistem secara lebih lengkapnya, maka pengujian terhadap 3 buah diskusi online dari beberapa sumber dilakukan. Pengujian dilakukan pada data yang memiliki kualitas dan kuantitas yang hampir sama baik pada topic maupun komentarnya, misalnya dari segi ukuran atau banyaknya komentar dan panjang topic atau jumlah kata pada 57opic yang akan diuji.

Maka diperoleh nilai threshold, precision, recall dan F-measure dari pengujian ke 3 data tersebut diatas seperti pada Tabel 5.2.

Tabel 5.2 Hasil Pengujian

No	Nama Website/ Forum	Threshold	Jumlah komentar	Akurasi
1	Forum.idws.id (Forum Indowebster)	10%	20 Komentar	Precision=0,916667 Recall =0,647059 F-Measure= 0,758621
2	http://teknojurnal.com/definisi-internet-of-things	10%	51 Komentar	Precision=0,84375 Recall =0,65853659 F-Measure=0,73972603

3	https://www.bersosial.com/threads/motivasi-menulis-buku.19737/	10%	18 Komentar	Precision=1 Recall =0,727273 F-Measure=0,842105
			Total	Precision=0,920139 Recall =0,677623 F-Measure=0,780151

Dari hasil analisis ini diperoleh nilai rata-rata *precision* dari ketiga data yaitu 92% , nilai *recall* 67% dan *F-Measure* 78 %. Hal ini menunjukkan bahwa penerapan konsep algoritma TF-IDF dan Naïve bayes pada kegita web site diatas memberikan hasil yang cukup baik dengan nilai variable lebih besar dari 50%. Keakuratan sistem dalam mengklasifikasi relevansi antara topic dan komentar memberikan hasil yang baik.

BAB VI
KESIMPULAN DAN SARAN

6.1. Kesimpulan

Penyaringan komentar yang tidak relevan terhadap topik dalam suatu diskusi merupakan fitur penting yang harus tersedia untuk berbagai forum diskusi *online* baik dalam bentuk blog, website maupun e-learning. Sehingga diskusi yang dilakukan dapat memberikan manfaat yang sesuai dengan harapan baik bagi peserta maupun bagi administrator atau tenaga pengajar. Hal ini dikarenakan komentar yang tidak memiliki kontribusi terhadap topik diskusi akan menjadi penghalang tersampainya informasi atau tujuan yang efektif dari suatu diskusi. Selain itu juga akan mempengaruhi nuansa diskusi online yang aktif dan efektif. Berdasarkan hasl pengujian yang dilakukan terhadap sistem dapat diambil kesimpulan bahwa metode TF-IDF dan Naïve Bayes sangat effektif digunakan dalam memfilter dan mengklasifikasikan relevansi komentar tehadap topic pada diskusi online dengan nilai rata-rata precision, recall dan F-measure lebih besar dari standart yang diharapkan (lebih besar dari 50%). Hasil dari klasifikasi sangat ditentukan pada proses yang terjadi mulai dari tahapan *preporcessing* teks pada awal proses, terlebih lagi pada proses stemming dan normalisasi kata tidak baku. Hal ini dikarenakan data yang diklasifikasi adalah data teks yang tidak memiliki standar baku penulisan sehingga memicu munculnya banyak sekali kesalahan ejaan ataupun ejaan-ejaan yang tidak sesuai dengan Ejaan Yang Disempurnakan seperti gue, gw, elo, pg, dsb. Hal ini tentunya akan mempengaruhi kemampuan

algoritma dalam mengklasifikasi data. Dinamika kamus data baku tentunya juga dipengaruhi oleh permasalahan tersebut. Dikarenakan

Selain itu, penentuan nilai *threshold* juga akan mempengaruhi proses klasifikasi, semakin kecil nilai *threshold* akan semakin rendah kemampuan algoritma mengenali komentar yang relevan dan begitu pula sebaliknya. Dengan nilai threshold yang cukup 10% pada pengujian menunjukkan hasil yang cukup bagus sebagai pembuktian keberhasilan penerapan algoritma TF-IDF dan Naïve Bayes pada diskusi online.

6.2. Saran

Untuk pengembangan buku berkaitan dengan penerapan metode TF-IDF dan Naïve Bayes selanjutnya beberapa saran kedepan antara lain adalah:

1. Perlu dilakukan penelitian yang lebih lanjut berkaitan dengan algoritma stemming yang memiliki akurasi yang lebih baik. Sehingga pemotongan kata imbuhan yang digunakan menjadi selalu tepat. Kelebihan stemming Porter stemmer yang digunakan pada buku ini adalah waktu proses yang lebih cepat namun masih sering terdapat kesalahan pada pemotongan kata.
2. Perlunya sistem yang mampu menampung dan mendeteksi kata-kata yang tidak baku di kamus data secara otomatis. Sehingga apabila sistem menemukan kata baru sistem tetap masih bisa terdeteksi.
3. Perlu adanya nilai threshold yang bersifat dinamis, sehingga tidak ada batasan nilai threshold tertentu untuk topik dan komentar yang akan dianalisa. Hal ini disebabkan semakin tinggi nilai threshold maka akan semakin selektif proses filtering yang dilakukan.
4. Perlunya sebuah sistem yang mampu melakukan proses filtering tidak hanya bersasarkan frekuensi kata atau kalimat yang ditemukan namun juga makna dari kata atau kalimat yang digunakan.

DAFTAR PUSTAKA

Almeida, T., & Yamakami, A. (2010). Content-Based Spam Filtering. *Neural Networks (IJCNN), The 2010 International Joint Conference*, (pp. 1-7).

Azwan Ahmad, Abdul Ghani Abdullah, Mohammad Zohir Ahmad and Abd. Rahman Abd Aziz, 2005. Kesan efikasi kendiri guru sejarah terhadap amalan pengajaran berbantukan teknologi maklumat dan komunikasi (ICT). Jurnal Penyelidikan Pendidikan, 7: 14-27.

Burdescu, D., Mihaescu, M., & Logofatu, B. (2008). Employing Bayes Classifier for Improving Learner's Proficiency. *Internet and Web Applications and Services, 2008. ICIW '08. Third International Conference*, (pp. 38 - 42).

Chong, Chee Keong, Sharaf Horani and Jacob Daniel, 2005. A study on the use of ICT in mathematics teaching. Malaysian Online Journal of Instructional Technology, 2(3): 43-51.

Creswell, J.W. (2003). *Research Design: Qualitative, Quantitative and Mixed Methods Approach* (2nd *ed.)*. London: Sage.

Jamaluddin, M., Hashim, R., Hanafiah, M., Mohd Zahari, M., & Zulkifly, M. (2010). Performance Evaluation of Online Discussion. *Science and Social Research (CSSR), 2010 International Conference*, (pp. 1065-1068).

Lee, S., & Kim, H.-j. (2008). News Keyword Extraction for Topic Tracking. *Fourth International Conference on Networked Computing and*

Advanced Information Management, (pp. 554-559).

Lui, A. K.-F., Li, S. C., & Choy, S. O. (2007). An Evaluation of Automatic Text Categorization in Online Discussion Analysis. *Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007)*, (pp. 205 - 209).

Murray, C . (2008). Schools and social networking: Fear or education? *Synergy Perspectives: Local, 6*(1), 8-12.

Phuc, D., & Phung, N. T. (2007). Using Naive Bayes Model and Natural Language Processing for Classifying Messages on Online Forum. *Research, Innovation and Vision for the Future, 2007 IEEE International Conference*, (pp. 247 - 252).

Pramono, L., Rohman, A., & Hindersah, D. (2013). Modified Weighting Method in TF-IDF Algorithm for Extracting User Topic Based on Email and Social Media in Integrated Digital Assistant. *Rural Information & Communication Technology and Electric-Vehicle Technology (rICT & ICeV-T), 2013 Joint International Conference*, (pp. 1-6).

Samuel, R.I. and Zaintun, A.B., 2006. Do teachers have adequate ICT resources in promoting English language teaching and learning?. Journal of ICT, 5: 29-44.

Shee, D. Y., & Wang, Y.-S. (2008, April). Multi-criteria evaluation of the web-based e-learning system: A methodology based on learner satisfaction and its

applications. *Computers & Education, 50*(3), 894–905.

TeacherStream. (2012). *Mastering Online Discussion Board Facilitation.* TeacherStream, LLC.

Toprak, A. (2009). *Toplumsal paylasim agi Facebook.* Istanbul: Kalkedon.

Wang, Y., Wu, Z., & Wu, R. (2008). Spam Filtering System Based on Rough Set and Bayesian Classifier. *Granular Computing, 2008. GrC 2008. IEEE International Conference*, (pp. 624 - 627).

Yang, Q., Zheng, S., Huang, J., & Li, J. (2008). A Design to Promote Group Learning in e-learning by Naive Bayesian . *2008 International Symposium on Computational Intelligence and Design*, (pp. 379-382).

Yu, Y., & Chen, Y. (2012). A Novel Content Based and Social Network Aided Online Spam Short Message Filter. *Intelligent Control and Automation (WCICA), 2012 10th World Congress*, (pp. 444 - 449).