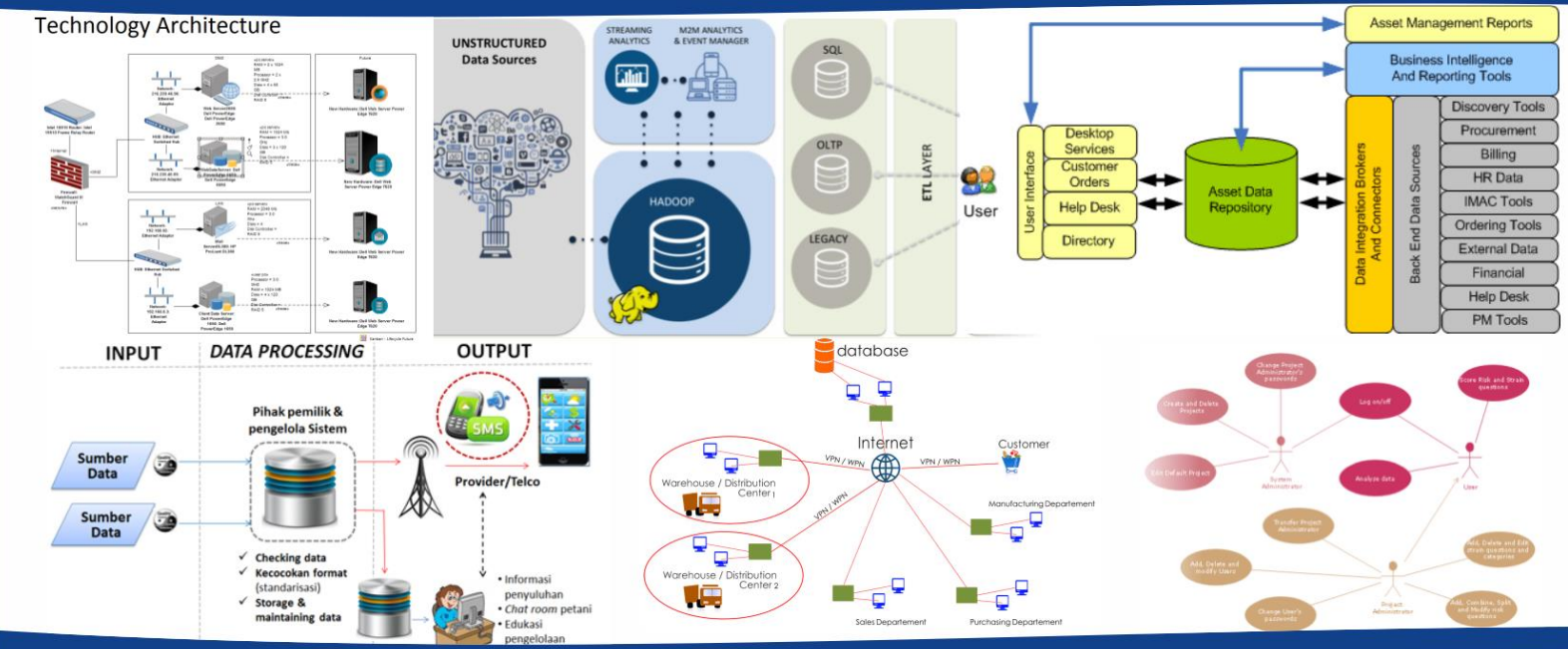
ISSN : 2614-1701 (*Cetak*)
ISSN : 2614-3739 (*Online*)

Volume 3, Nomor 2, Juni 2020

MISI

JURNAL MANAJEMEN INFORMATIKA & SISTEM INFORMASI



Diterbitkan Oleh LPPM STMIK Lombok

Jln. Basuki Rahmat No.105 Praya, Lombok Tengah - NTB
Telp dan Fax (0370) 654310 - e-journal.stmiklombok.ac.id/jsi
email. lppm@stmiklombok.ac.id

DEWAN REDAKSI

DAFTAR ISI

KLASIFIKASI KOMENTAR PUBLIK TERHADAP KEBIJAKAN PEMERINTAH PADA FACEBOOK FRONTPAGE KOMPAS MENGUNAKAN CLUSTRING K-MEANS, FURTHEST FIRST

I Wayan Dikse Pancane[1], I Wayan Suriana[2]

Program Studi Teknik Elektro, Fakultas Teknik dan Informatika, Universitas Pendidikan Nasional
Jl. Bedugul No. 39 Sidakarya, Denpasar, Bali 80224
diksapancane@undiknas.ac.id, wayansuriana@undiknas.ac.id

ABSTRACT

The number of Facebook users in Indonesia is very high. This condition is used as an opportunity by the online media like news portal by making a page on facebook. The content of their wall is a preview or short description about news / articles which is posted on their official website. By using k-Means and Farthest-First, the wall can be automatically grouped by similarity topics. Preprocess in this grouping or clustering is done by Porter Stemmer Stemmer and Naizef Stemmer. Amount of data to be tested in this study were 466 wall. The best clustering quality produced by the k-Means k = 2 without pre-stemming process. Accuracy achieved for the labeling of "national" and "non-national" is 92.92%. Clustering wall labeled "national" into the label of "corruption" and "non-corruption", produced 77.78% accuracy. This result was achieved by the k-means clustering with k = 2 pre-process Stemming Nazief.

Keywords: Clustering, K-Means, Farthest-First, Porter Stemmer, Nazief Stemmer

ABSTRAK

Begitu tingginya jumlah pengguna Facebook di Indonesia membuat media cetak nasional di Indonesia juga membuat Facebook *page* dengan isi *wall* adalah cuplikan dari berita yang ada di websitenya. Dengan menggunakan k-Means dan Farthest-First, wall tersebut dapat dikelompokkan secara otomatis berdasarkan kesamaan topik bahasannya. Pre-proses dari pengelompokkan ini menggunakan Porter *Stemmer* dan Naizef *Stemmer*. Dari hasil uji coba 466 data *wall* facebook, cluster terbaik didapatkan dengan k-Means k=2 tanpa pre-proses *stemming*. Akurasi yang dicapai untuk pelabelan "nasional" dan "non-nasional" adalah 92.92%. Clustering wall "nasional" terbaik juga dihasilkan menggunakan k-Means k=2 dengan menerapkan pre-proses Nazief. Akurasi yang didapatkan dengan label "korupsi" dan "non-korupsi" adalah 77.78%.

Kata kunci: Pengelompokan, K-Means, Farthest-First, Porter Stemmer, Nazief Stemmer

1. PENDAHULUAN

Disaat internet sudah merupakan kebutuhan sehari-hari di negara kita seperti saat ini maka kebutuhan informasi terkini sangat tinggi.Pusat penyedia informasi seperti media TV, Radio dan portal berita di internet menjadi point-point media yang paling dicari oleh masyarakat.Media informasi seperti media cetak harian, mingguan, bulanan apalagi tiga bulanan sudah menjadi sesuatu yang tidak relevan lagi mengingat keterkinian beritanya sangat tertinggal.

Media TV dan Radio walaupun mampu menyajikan informasi terupdate masih menjadi media informasi yang mahal.Ketersediaan perangkat untuk melakukan pemancaran siaran serta daya jangkau pemancara masih menjadi kendala. Dengan mahalnya dana yang dibutuhnkan maka stasiun TV masih berpikir ulang untuk melakukan siaran langsung.Media Internet adalah satu-satunya yang paling tepat untuk mengatasi berbagai kesulitan tersebut, Internet mampu membawa informasi langsung ke orang yang membutuhkan dengan cepat dan memungkinkan real time.Apalagi saat ini semua provider seluler telah membuka layanan internet dengan murah dan cepat.

Hal ini membuat media-media informasi cetak akhirnya perlahan mulai melakukan manuver dengan menduplikasi conten dari media cetaknya ke

dalam bentuk digital. Hampir semua koran nasional saat ini mempunyai koran versi onlinenya. Contoh Harian Kompas (kompas.com), Harian JawaPos (jawapos.co.id), Majalah Tempo (majalah.tempointeraktif.com) dan masih banyak lagi. Belum lagi majalah teknologi, kesehatan dan yang lainnya. Kebutuhan pembaca akan informasi yang terbatas pada versi cetak diakomodir pada versi onlinenya. Selain menjadi lebih update biaya yang dikeluarkan masyarakat untuk mendapatkannya pun jauh lebih murah. Hanya bermodalkan HP, laptop, komnputer atau tablet serta koneksi internet baik via modem HP, telpon rumah bahkan wifi gratis, masyarakat sudah langsung bisa mendapatkan informasi.

Walaupun mampu mengatasi segala faktor keterkinian, teknologi juga datang tanpa terlepas dari ekses dan efek negatifnya. Di Internet, siapa saja dapat menjadi siapa saja dan apa saja. Jika pengguna internet tidak waspada dan berhati-hati tentunya selain mendapatkan keuntungan bisa jadi juga mendapatkan kerugian.Demikian juga dengan berjamurnya keberadaan media-media online di internet.Tidak jarang situs-situs informasi yang ada di internet menyajikan informasi yang tidak akurat.Beberapa situs pribadi seperti blog lebih sering mengungkapkan opini dibandingkan fakta.Tentunya hal ini menjadi sangat berbahaya jika pembaca akhirnya mengikuti bentuk-bentuk opini yang kemungkinan jauh melenceng dari fakta.

Keuntungan lain dari media-media informasi online ini adalah adanya interaktif antara pembaca dan media. Hal ini sebenarnya terjadi juga dengan media cetak. Surat pembaca, opini, dan tajuk adalah contoh–contoh cara menjembatani pembaca dengan media. Tetapi sifatnya sangat kaku jika dibandingkan dengan media internet.Pada media informasi internet, pembaca bisa langsung mengkomentari setiap berita yang muncul.Hampir di seluruh media informasi internet menyediakan fitur komentar yang memungkinkan pembaca memberikan komentar dengan mudah dan tanpa banyak aturan. Langsung tulis komentar menggunakan ID di internet apakah itu menggunakan akun media social seperti facebook atau twitter atau ada media yang mengharuskan pembaca untuk menjadi anggota gratis dulu untuk bisa menuiskan komentar.

Hal ini menjadi sangat menarik karena memperlihatkan bagaimana masyarakat kita selain haus akan informasi kekinian juga ternyata sangat concern masalah-masalah sosial yang timbul. Masyarakat bahkan berlomba-lomba memberika solusi yang tertepat dalam berbagai kasus.Masyarakat juga tidak ragu-ragu mengecam, menghujat atau justru mendukung terhadap sebuah kejadian. Untuk media-media internet yang berstatus koran nasional dan sudah terpercaya di masyarakat kita, hal ini sebenarnya bisa dimanfaatkan juga oleh pemerintah dalam hal mendapatkan reaksi yang cepat dan tepat dari masyarakat.

Dengan latar belakang inilah penulis tertarik untuk mengeksplorasi secara lebih detail dari semua komentar pembaca yang ada pada portal berita kompas yang berhubungan dengan pemerintahan dan situasi politik di Indonesia.Komentar akan dikoleksi menjadi sebuah data untuk dianalisa dan didapatkan cluster-cluster komentar dari masyarakat tersebutdengan harapan dapat membantu pemerintah didalam mengetahui secara cepat tingkat keberpihakan masyarakat terhadap program-program pemerintah atau kasus-kasus tertentu yang sedang terjadi.

2. Tinjauan Pustaka

2.1 Text Mining

Text mining dapat diartikan sebagai penemuan informasi yang baru dan tidak diketahui sebelumnya oleh komputer, dengan secara otomatis mengekstrak informasi dari sumbersumber yang berbeda. Kunci dari proses ini adalah menggabungkan informasi yang berhasil diekstraksi dari berbagai sumber (Hearst,2003). Sedangkan menurut (Milkha Harlian) text mining memiliki definisi menambang data yang berupa teks dimana sumber data biasanya didapatkan daridokumen, dan tujuannya adalah mencari kata kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen.Dengan *text mining* tugastugas yang berhubungan dengan penganalisaan teks dengan jumlah yang besar, penemuan pola serta penggalian informasi yang mungkin berguna dari suatu teks dapat dilakukan.

Sebagai bentuk aplikasi dari *text mining*, sistem klasifikasi berita menggunakan berita sebagai sumber informasi dan informasi klasifikasi sebagaiinformasi yang akan diekstrak dari sumber informasi. Informasi klasifikasi dapatberbentuk angka angka probabilitas, set aturan atau bentuk lainnya.

Walaupun inti dari suatu sistem klasifikasi adalah tahap penemuan pola (*pattern discovery*) namun secara lengkap proses *text mining* dibagi menjadi 3 tahap utama, yaitu proses awal terhadap teks (*text preprocessing*), transformasi teks ke dalam

bentuk antara (*text transformation/feature generation*), dan penemuan pola (*pattern discovery*). (Even dan Zohar, 2002). Masukan awal dari proses ini adalah suatu data teks dan menghasilkan keluaran berupa pola sebagai hasil interpretasi.

2.2 *Text Preprocessing*

Tahapan awal dari *text mining* adalah *text preprocessing* yang bertujuan untuk mempersiapkan teks menjadi data yang akan mengalami pengolahan pada tahapan berikutnya. Beberapa contoh tindakan yang dapat dilakukan pada tahap ini, mulai dari tindakan yang bersifat kompleks seperti *partofspeech* (pos) *tagging*, *parse tree*, hingga tindakan yang bersifat sederhana seperti proses parsing sederhana terhadap teks, yaitu memecah suatu kalimat menjadi sekumpulan kata. Selain itu pada tahapan ini biasanya juga dilakukan *casefolding*, yaitu pengubahan karakter huruf menjadi huruf kecil. Proses *partofspeech*melakukan parsing terhadap seluruh kalimat dalamteks kemudian memberikan peran kepada setiap kata, misalnya : petani (subyek)pergi (predikat) ke (kata hub) sawah (keterangan). Hasil dari *partofspeechtagging* dapat digunakan untuk *parse tree*, di mana masing masingkalimat berdiri sebagai sebuah pohon mandiri.Untuk proses parsing sederhana tidak dibangun *parse tree* seperti carasebelumnya. Pada proses parsing sederhana sistem akan memecah teks menjadisekumpulan katakata,yang kemudian akan dibawa sebagai input untuk tahapberikutnya pada proses *text mining*.

2.3 *Text Transformation (feature generation)*

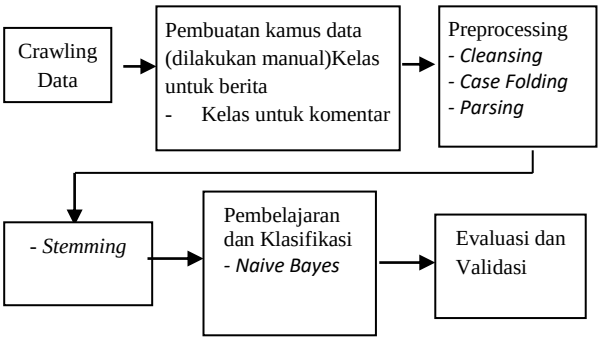
Pada tahap ini hasil yang diperoleh dari tahap *text preprocessing* akanmelalui proses tranformasi. Adapun proses transformasi ini dilakukan denganmengurangi jumlah katakatayang ada dengan penghilangan *stopword* dan jugadengan mengubah katakatake dalam bentuk dasarnya (*stemming*).*Stopword* adalah katakatayang bukan merupakan ciri (kata unik) darisuatu dokumen seperti kata sambung, kata kepunyaan. Memperhitungkan*stopword* pada transformasi teks akan membuat keseluruhan sistem *text mining*bergantung kepada faktor bahasa. Hal ini menjadi kelemahan dari prosespenghilangan *stopword*. Namun proses penghilangan *stopword* tetap digunakankarena proses ini akan sangat mengurangi beban kerja system. Denganmenghilangkan *stopword* dari suatu teks maka sistem hanya akanmemperhitungkan katakatayang dianggap penting.

Stemming adalah contoh tindakan lain yang dapat dilakukan pada tahaptransformasi teks. *Stemming* adalah proses untuk mereduksi kata ke bentukdasarnya. Kata yang memiliki bentuk dasar samawalaupun imbuhannya berbeda seharusnya memiliki kedekatan arti. Disampingitu juga, proses stemming akan sangat mengurangi jumlah dan beban database.Jika setiap kata disimpan tanpa melalui proses stemming, maka satu macam katadasar saja akan disimpan dengan berbagai macam bentuk yang berbeda sesuaidengan imbuhan yang mungkin melekatinya. Hal ini sangat berbeda jika kitamenerapkan proses *stemming* pada tahap ini, satu kata dasar hanya akan disimpansekali walaupun mungkin kata dasar tersebut pada sumber data sudah berubah daribentuk aslinya dan mendapatkan berbagai macam imbuhan.

Proses *stemming* dan penghilangan *stopword* dapat digunakan secaramandiri atau tergabung, dimana dilakukan proses penghilangan *stopword* terlebihdahulu yang diikuti dengan proses *stemming*. Hal ini dilakukan untuk menemukanpola dari teks dalam berita tersebut.Karena pada Tugas Akhir ini menggunakan teks berita dalam bahasaIndonesia sebagai objek penelitian maka akan dibahas proses stemming padateks/kata berbahasa Indonesia pada bahasan selanjutnya.

3. Metodologi Penelitian

3.1 Diagram alir Penelitian



Gambar 1. Blok diagram aliran kerja penelitian

3.2 Pemodelan Sistem

1. Model Probabilistic Naive Bayes

Model probabilitas untuk classifier adalah model kondisional Lihat hasil pada table 1 dan 2

$$p(C|F_1,..Fn) \quad (2.1)$$

terhadapvariabelkelasdependenCdengansejumlahkecilhasilataukelas, tergantung pada beberapa

variabel fitur F1 sampai Fn. Masalahnya adalah bahwa jika jumlah fitur n besar atau bila fitur bisa mengambil sejumlah besar nilai, maka membuat sebuah model pada tabel probabilitas adalah tidak mungkin. Oleh karena itu kita mereformulasi model untuk membuatnya lebih fleksibel.

Menggunakan teorema Bayes , kita menulis

$$p(C|F_1,..F_n)=\frac{p(C)p(F_1,..F_n|C)}{p(F_1,..F_n)} \quad (2.2)$$

Dalam bahasa Inggris persamaan di atas dapat ditulis sebagai

$$\text{posterior}=\frac{\text{prior x likehood}}{\text{evidence}} \quad (2.3)$$

Dalamprakteknyakitahanyatertarikpadapembilangdaripersamaantersebut, karenapenyebuttidaktergantungpadaCdannilai-nilaifiturFidiberikan,sehinggapenyebutsecaraefektifkonstan.Pembilanginisetaradengan model probabilitasgabungan$p(C|F1,..Fn)$ yangdapatditulisulangsebagaiberikut, menggunakan penggunaan berulang dari definisi probabilitas bersyarat:

$$\begin{aligned} &p(C|F_1,..F_n)\\ &=p(C)p(F_1,..F_n|C)\\ &=p(C)p(F_1|C)p(F_2,..F_n|C,F_1)\\ &=p(C)p(F_1|C)p(F_2|C,F_1)p(F_3,..F_n|C,F_1,F_2)\\ &=p(C)p(F_1|C)p(F_2|C,F_1)p(F_3|C,F_1,F_2)p(F_4,..F_n|C,F_1,F_2,F_3)\\ &=p(C)p(F_1|C)p(F_2|C,F_1)p(F_3|C,F_1,F_2)..p(F_n|C,F_1,F_2,F_3,..F_{n-1}) \end{aligned} \quad (2.4)$$

Sekarangasumsikemandirianbersyaratyang "naif"memegangperanan. MenganggapbahwasetiapfiturFiadalahsecarakondisi independenterhadap setiap fitur lainnya Fj untuk j ≠ i. Ini berarti bahwa

$$p(F_i|C,F_j)=p(F_i|C) \quad (2.5)$$

untuk i ≠ j , sehingga joint modeldapat dinyatakan sebagai

$$p(C|F_1,..,Fn)=p(C)p(F_1|C)p(F_2|C)p(F_3|C)..$$

$$=p(C)\prod_{i=1}^{n}p(F_i|C) \quad (2.6)$$

Iniberartibahwadibawahasumsiindependendiatas,distribusibersyaratdari variabel kelas C dapat dinyatakan seperti ini :

$$=p(C|F_1,..,Fn)=\frac{1}{z}\,p(C)\prod_{i=1}^{n}p(F_i|C) \quad (2.7)$$

dimanaZ(bukti)adalahfaktorskalatergantunghanyapada $F_1,\ldots,F_n$,yaitu, sebuah konstanta jika nilai dari variabel fitur diketahui.

Modeldari bentuk ini jauh lebih mudah dikelola, karena mereka memecah menjadi class priorp(C) dan distribusi probabilitas independen $p(F_i|C)$. Jika ada k kelas dan jika model untuk masing-masingp$(F_i|C)$= c dapat dinyatakan dalam bentukparameter,makamodelnaifBayesyangsesuaimemiliki(k-1)+nrk parameter. Dalam prakteknya, sering k = 2 (klasifikasi biner) dan r = 1 (variabel Bernoulli sebagai fitur) yang umum, sehingga jumlah parameter model Bayes naif adalah2n+1,dimananadalahjumlahfiturbineryangdigunakanuntuk klasifikasi dan prediksi.

2. Estimasi Parameter

Semua model parameter (yaitu, prior kelas dan distribusi probabilitas fitur) dapatdidekatidenganfrekuensirelatifdarihimpunanpelatihan.Inimerupakan perkiraankemungkinanmaksimumdariprobabilitas. Sebuah *priorclass*dapat dihitung dengan asumsi kelas *equiprobable* (yaitu, prior = 1 / (jumlah kelas)), atau denganmenghitungperkiraanprobabilitaskelasdarihimpunanpelatihan(yaitu, (prioruntuk kelas tertentu) = (jumlah sampel di kelas) / (jumlah sampel)). Untuk memperkirakanparameteruntukdistribusifiturini,seseorangharus mengasumsikandistribusiataumenghasilkanmodelnon-Fmatrikuntukfitur-fitur darisetpelatihan.Jikaseseorangberhadapan

$$\mu=\frac{1}{N}\sum_{i=1}^{N}xi \quad (2.8)$$

dengandatakontinu,asumsi khas adalah distribusi Gaussian, dengan parameter model dari mean danvarians.

Mean, μ, dihitung dengan dimanaNadalahjumlahsampeldanxiadalahnilaidarisuatucontohyang diberikan.

$$\text{Varian dihitung dengan } \sigma^2=\frac{1}{N}\sum_{i=1}^{N}(xi-\mu)^2 \quad (2.9)$$

Jika sebuah kelas tertentu dan nilai fitur tidak pernah terjadi bersama-sama dalamhimpunanpelatihanmakaestimasiprobabilitas berbasisfrekuensiakan menjadi nol. Hal ini bermasalah karena akan menghapus seluruh informasi dalam probabilitas lain ketika mereka dikalikan. Oleh karena itu sering diinginkan untuk memasukkankoreksisampelkecildalamsemuaperkiraanprobabilitasbahwa tidak ada probabilitas untuk menjadi persis nol.

3.3 Penyusunan Algoritma

3.3.1. *Crawling Data*

Data komentar facebook kompas ditarik memanfaatkan GraphAPI dari facebook.Berikut adalah hasil XMLnya.

```
ID: 398904863461485_5074958
From ID: 100000202236281
From Name: Mutia Nugraha
Message: makan pakis enak ke timbang pks hehhehehehee
Timestamp: 2012-04-07T12:50:49+0000

ID: 398904863461485_5074959
From ID: 100001648513699
From Name: Buyut Jaga Makam
Message: 2014 golput lebih baik.
Timestamp: 2012-04-07T12:50:51+0000

ID: 398904863461485_5074960
From ID: 100001594936036
From Name: Mhail Suarez
Message: pak tifatul,,,knp mesti takut,,,kalian punya power koq,,,,
Timestamp: 2012-04-07T12:50:52+0000

ID: 398904863461485_5074961
From ID: 100002163565224
From Name: Een Egj
Message: gtu dong.
Timestamp: 2012-04-07T12:50:56+0000

ID: 398904863461485_5074963
From ID: 100002244511563
From Name: Harri Mimi
Message: hahahaha
Timestamp: 2012-04-07T12:51:02+0000
```

Gambar 3.2Contoh hasil retrieve XML komentar facebook

Data hasil retrieve ini dimasukkan ke database MySQL sebagai data training.

autoid	iduser	name	comment	date
440	100001744211876	Fatimah Wulandari	Hmmm..PKS bgai telur diujung tanduk..ywdh met mnun...	2012-04-0
441	100000362892573	Aziz Adzy	Koreksilah diri sblum mengkoreksi org laen. . .	2012-04-0
442	100001095828969	H Roesdi Effendi Nst	me juah2 ya	2012-04-0
443	100003077274325	Mohd Ajhun Piliang	ssuka hti loe lah..	2012-04-0
444	100000130773054	Dian Utami	mohon dukungannya teman2 dalam pemilihan ICON FK U...	2012-04-0
445	100002614022440	Ahmad Fahri Ryanfajriansyah	waduh gawat bngt... negara makin kacaw ajha...?	2012-04-0
446	100001284420843	Willy Kusnadi	Mo di ujung tanduk,mo di ujung dunia terserah. yg	2012-04-0
447	100003177745104	Muhdesrianto Kh	Mau bela koalisi? Boleh Mau suka PKS? Silahken M...	2012-04-0
448	1850458182	Hary Agustia	setuju bgt	2012-04-0
449	1671095682	Mas Adi Winarno	pada kenyataannya pemerintah tdk jadi menaikan buk...	2012-04-0
450	100002908309940	Arip Gunawan	Berdayung sampan ke pulau feri,menatap bintang sam...	2012-04-0
451	100000111582674	Opa Tjoek Aryana	Hwaratah.....kono.....!	2012-04-0
452	100002913184367	Ardan Dhana	PKS??	2012-04-0
453	100000062859377	Widarto Plat H	Ini yang namanya jual suara rakyat untuk cari jaba...	2012-04-0
454	100001436973462	Afrizal Kasim Caniago	>>Salah tuh Sembiring, Reshuffle itu perintah Sby...	2012-04-0

Gambar 3.3Contoh komentar sudah dimasukkan ke MySQL

3.3.2. Pembuatan Kamus Kelas

Tahapan yang dilakukan dari pembuatan kamus kelas ini adalah mengambil data hasil crawling dan melakukan klasifikasi secara manual dan disimpan menjadi sebuah kamus kelas.Ada 2 (dua) kamus kelas yaitu untuk kelas kategori berita (kebijakan) dan kelas kategori pendapat (pro, kontra).

3.3.3. Pre-Processing

Tahapan yang dilakukan dari dokumen *pre-processing* adalah sebagai berikut:

1. *Cleansing*, yaitu proses membersihkan dokumen dari kata yang tidak diperlukan untuk mengurangi noise. Kata yang dihilangkan adalah karakter HTML dan simbol.
2. *Case folding*, yaitu penyeragaman bentuk huruf serta penghapusan angka dan tanda baca. Dalam hal ini yang digunakan hanya huruf latin antara a sampai dengan z.
3. *Parsing*, yaitu proses memecah dokumen menjadi sebuah kata. Hal ini sesuai dengan fitur digunakan yaitu *unigram*.

SOURCE

Ini yang kalian bilang Demokrasi ? YANG BENER DEMOCRAZY !!!

Cleaning and Folding

ini yang kalian bilang demokrasi yang bener democrazy

3.3.4. Pemilihan dan Ekstraksi Fitur

Berikut adalah proses pemilihan dan ekstraksi fitur yang akan digunakan sebagai dasar proses klasifikasi.

1. *Stemming*, bertujuan mengurangi variasi kata yang memiliki kata dasar sama. Proses`*stemming* dilakukan dengan menggunakan bantuan KBBI.

```
Penghapusan partikel = menggelapkan
Penghapusan milik = menggelapkan
Penghapusan prefiks1= gelapkan
Penghapusan prefiks2= gelapkan
Penghapusan sufiks= gelap
```

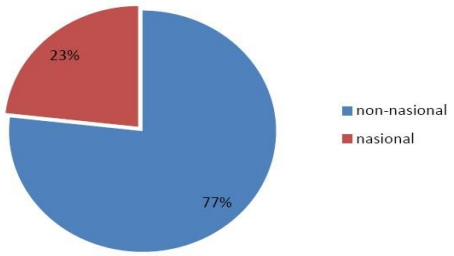
Gambar 3.5 Contoh hasil *stemming*

3.4 DATA

Data yang digunakan pada penelitian ini adalah data wall Facebook dari Facebook *page* KompasCom yang diposting selama bulan Januari sampai dengan Juli 2012. Pelabelan manual untuk melakukan interpretasi hasil *cluster*, memanfaatkan caption yang sudah ada pada setiap wall KompasCom. Jumlah data yang diolah sebesar 466 wall. Penelitian ini menggunakan 3 jenis pre-proses yaitu *wall* tanpa stem, *wall* dengan melalui Porter *stemmer* dan *wall* dengan melalui Nazief *stemmer*.Stopwords yang digunakan adalah stopwords dari penelitian Tala 2003 yaitu sebanyak 758 kata. Total wall yang diproses adalah 466 wall dengan jumlah attribut/fitur 3.665.

4. HASIL DAN PEMBAHASAN

Pengelompokan yang dilakukan oleh KompasCom pada setiap wall adalah untuk label "nasional" sebesar 108 wall dan untuk label "non-nasional" adalah 358 wall.



Gambar 2 Grafik Prosentase Jumlah *Wall* berdasarkan 2 label manual

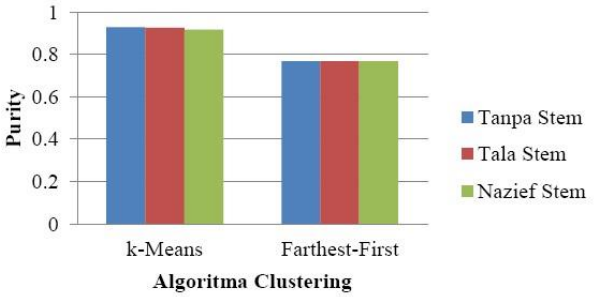
Tabel 1 Purity k-Means k=2

SEED	TANPA STEM	TALA STEM	NAZIEF STEM
10	0,7682	0,7682	0,8906
20	0,8863	0,8864	0,8864
30	0,9292	0,9270	0,9163
40	0,7682	0,7682	0,7682
50	0,7682	0,7682	0,7682
60	0,7682	0,7682	0,7682
70	0,7682	0,7682	0,7682
80	0,7682	0,7682	0,7682
90	0,7682	0,7682	0,7682
100	0,7682	0,7682	0,7682

Tabel 2 Purity Farthest-first k=2

SEED	TANPA STEM	TALA STEM	NAZIEF STEM
10	0,7682	0,7682	0,7682
20	0,7682	0,7682	0,7682
30	0,7682	0,7682	0,7682
40	0,7682	0,7682	0,7682
50	0,7682	0,7682	0,7682
60	0,7682	0,7682	0,7682
70	0,7682	0,7682	0,7682
80	0,7682	0,7682	0,7682
90	0,7682	0,7682	0,7682
100	0,7682	0,7682	0,7682

Percobaan dilakukan dengan menggunakan seed yang berbeda-beda untuk mendapatkan inisialisasi centroid yang paling baik. Setelah dilakukan percobaan menggunakan k-Means dengan membandingkan pre-proses tanpa stem, dengan Tala Stem dan Nazief Stem didapatkan seperti pada Tabel 1. Purity tertinggi dihasilkan oleh percobaan tanpa pre-proses stem dengan nilai 0.9292. Tabel 2 menunjukkan bahwa hasil purity yang dihasilkan Farthest- First adalah 0.7682. Nilai ini lebih kecil dari yang dihasilkan oleh k-Means. Perbandingan k-Means dan Farthest-First ditunjukkan pada Gambar 3.



Gambar 3 Grafik Perbandingan Purity antara k-Means dan Farthest-First

Purity didapatkan dengan cara membandingkan nilai presisi maksimal setiap kelasnya. Nilai purity 0.9292 ini hampir mencapai nilai maksimal dari purity yaitu 1. Ini menandakan cluster yang dihasilkan oleh k-Means dengan pre-proses tanpa stem memiliki kualitas yang sangat baik atau bisa diartikan bahwa pembeda antar cluster sangat jelas.Tabel 1 menunjukkan bahwa pada seed 10, 20 dan 30, stemming memang menghasilkan perbaikan purity walaupun sangat kecil. Tapi kemudian di seed berikutnya dan pada Tabel 2, stemming sama sekali tidak membawa perbaikan. Hal ini menunjukkan bahwa stemming hanya membawa pengaruh yang kecil baik terhadap kualitas cluster yang dihasilkan.

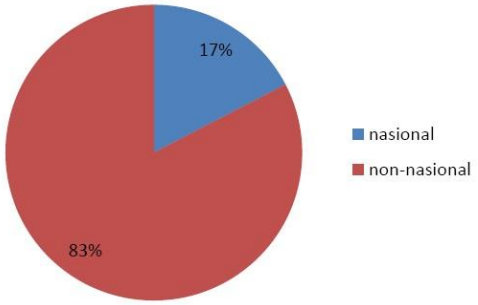
Tabel 3 Matching Matrix k-Means k=2

C0	C1	Label
30	78	Nasional
355	3	Non-Nasional

Tabel 3 menunjukkan komposisi jumlah label di masing-masing cluster. Matching Matrixyang diperlihatkan ini adalah pada purity tertinggi yaitu seperti yang dihasilkan oleh k-Meansdi atas. Dengan melihat jumlah label terbanyak di setiap clusternya, dapat diinterpretasikanbahwa Cluster 0 adalah kelompok yang terdiri dari wall yang memunyai topik "non-nasional".Sedangkan Cluster 1 adalah kelompok wall yang bertopik "nasional". Akurasi dariinterpretasi ini adalah sebesar 92.92%. Tabel 4 menunjukkan hasil interpretasi dan nilaiprecision recall masing-masing cluster.

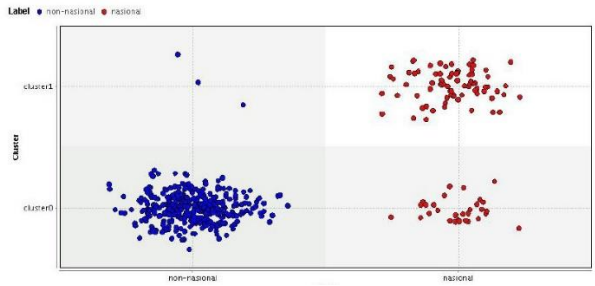
Tabel 4 Hasil Interpretasi Cluster

CLUSTER	LABEL	PRECISION	RECALL
0	Non-Nasional	0,9221	0,9916
1	Nasional	0,9630	0,7222



Gambar 4 Grafik Perbandingan Label "nasional" dan "non-nasional"

Hasil *Clustering* Kondisi purity yang tinggi yang dihasilkan oleh k-Means ditunjukkan pada Gambar 4.5. Walaupun sangat tinggi, namun masih belum mencapai nilai sempurna sehingga terlihat masih ada label yang menyeberang cluster.



Gambar 5 Grafik Hubungan Label Terhadap *Cluster* Pada k-Means k=2

Tabel 5 Centroid Per*Cluster*

CLUSTER	CENTROID	LABEL	FEATURE CENTORID
0	0,0196	Non-nasional	Ekonomi, emirates,enam, esdm,fpi, garuda, jokowi,juli,jurnal,kalinya, kanada, kandidat,merusak, meter,mikro, nomor,nigroho,obat oktober,olimpiade,opsi,organisasi,oscar, otak,pabrik,pakai,pakar,palu,paul,pegawai,pecan,pelatihan,telekomunikasi
1	0,0526	nasional	Ode,partainya, pdip,pks,pramono

Cluster 0 adalah kumpulan wall berlabel "non-nasional" dengan centroid cluster 0.0196. Jarak terdekat dengan centroid tersebut dicapai oleh fitur-fitur seperti yang ditunjukkan oleh Tabel 5. Sedangkan Cluster 1 adalah kumpulan wall yang berlabel "nasional" dengan centroid 0.0536. Cluster "nasianal" pada page KompasCom lebih banyak menuliskan kata-kata yang berhubungan dengan partai. Sedangkan cluster "non-nasional" lebih banyak menggunakan kata-kata yang bersifat umum.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil percobaan dapat disimpulkan bahwa :

1. k-Means dan Farthest-First dapat digunakan untuk melakukan pengelompokkan otomatis topik pada *wall* Facebook ber-Bahasa Indonesia.
2. Pre-proses *stemming* dalam percobaan ini mampu memberikan pengaruh perbaikan kualitas *cluster* sebesar 5%.
3. Untuk label "nasional" dan "non-nasional", hasil percobaan terbaik didapatkan dari hasil k-Means dengan nilai *Purity* = 0.0.9262 dan akurasi sebesar 92.92%. Hasil ini didapatkan dari *wall* yang tidak melewati *pre-proses stemming*. *Purity* dan akurasi yang didapatkan dari percobaan menggunakan Farthest-First masih lebih kecil dari k-Means dengan ratarata akurasi sebesar 75.69%.

5.2 Saran

Agar diperoleh hasil yang lebih maksimal, maka beberapa saran untuk pengembangan lebih lanjut adalah sebagai berikut:

1. Perlu diuji coba menggunakan teknik *clustering* yang lain sebagai perbandingan teknik yang menghasilkan *cluster* lebih baik lagi.
2. Komentar dari *wall* perlu diujicoba untuk di*cluster* juga untuk menemukan sentiment terhadap *wall* yang dikomentari.

6. Ucapan Terima Kasih

Ucapan Terima Kasih Terima kasih disampaikan kepada Universitas Pendidikan Nasional Denpasar dan Fakultas Teknik Dan Informatika Undiknas Juga Program Profesi Insinyur Undiknas yang telah memberikan masukan juga mendanai keberlangsungan jurnal ini.

DAFTAR PUSTAKA

[1] J. Informatika, W. Mega, and P. Dhuhita, "CLUSTERING MENGGUNAKAN METODE K-MEANS UNTUK," vol. 15, no. 2, 2015.
[2] D. P. Langgeni, Z. K. A. Baizal, and Y. F. A. W, "CLUSTERING ARTIKEL BERITA BERBAHASA INDONESIA," vol. 2010, no. semnasIF, pp. 1–10, 2010.
[3] P. Ke, "K-MEANS CLUSTERING," pp. 1–16.
[4] M. Kumar, "An Optimized Farthest First Clustering Algorithm," no. November 2013, 2019, doi: 10.1109/NUiCONE.2013.6780070.
[5] S. Godara, "A Comparative Performance Analysis of Clustering Algorithms," vol. 1, no. 3, pp. 441–445.
[6] M. H. Adiya and Y. Desnelita, "Jurnal Nasional Teknologi dan Sistem Informasi Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada RSUD Pekanbaru," vol. 01, pp. 17–24, 2019.
[7] N. Valarmathy and S. Krishnaveni, "Performance Evaluation and Comparison of Clustering Algorithms used in Educational Data Mining," no. 6, pp. 103–113, 2019.
[8] G. Sehgal and K. Garg, "Comparison of Various Clustering Algorithms," vol. 5, no. 3, pp. 3074–3076, 2014.
[9] A. V. D. Sano and H. Nindito, "APPLICATION OF K-MEANS ALGORITHM FOR CLUSTER ANALYSIS ON POVERTY OF PROVINCES IN INDONESIA," no. 6, pp. 141–150, 2011.
[10] S. Gnanapriya, "Evaluation of Clustering Capability Using Weka Tool," vol. 8, no. 1, pp. 181–187.
[11] B. K. Teknomo, "K-Means Clustering Tutorial," pp. 1–12, 2007.
[12] D. A. Vadeyar and H. K. Yogish, "Farthest First Clustering in Links Reorganization," vol. 5, no. 3, pp. 17–24, 2014.
[13] J. O. Ong, "IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING UNTUK MENENTUKAN STRATEGI MARKETING," no. April, pp. 10–20, 2013.
[14] M. M. C. K-means, "VOL . 9 NO . 1 April 2016," vol. 9, no. 1, pp. 94–101, 2016.
[15] M. H. Zafar and M. Ilyas, "A Clustering Based Study of Classification Algorithms," vol. 8, no. 1, pp. 11–22, 2015.
[16] M. M. Hassan, "ENHANCING CLUSTERING-BASED CLASSIFICATION ALGORITHMS IN E-COMMERCE APPLICATIONS," vol. 96, no. 18, pp. 6095–6105, 2018.
[17] S. Dasgupta and P. M. Long, "Performance guarantees for hierarchical clustering," no. July 2010.
[18] R. Pebria, B. H. I, and I. Sugihartono, "GFK-12 : IDENTIFIKASI PENYEBARAN GEMPA DI INDONESIA DENGAN METODE CLUSTERING," pp. 366–370.
[19] L. Mutawalli, M. T. A. Zaen, and W. Bagye, "KLASIFIKASI TEKS SOSIAL MEDIA TWITTER MENGGUNAKAN SUPPORT VECTOR MACHINE (Studi Kasus Penusukan Wiranto)," J. Inform. dan Rekayasa Elektron., vol. 2, no. 2, p. 43, 2019, doi: 10.36595/jire.v2i2.117.