JURNAL ELEKTRONIK SISTEM INFORMASI DAN KOMPUTER (JESIK)

ANALISIS SENTIMEN TERHADAP WACANA POLITIK PADA MEDIA MASA ONLINE MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE DAN NAIVE BAYES

AndiNurulHidayat
FakultasIlmuKomputer, StmikBinaMulia
Email: andinurulhidayat@yahoo.co.id

ABSTRAK

Analisissentimenmerupakansalahsatu domain Text Mining ataupenggalian data berupateks, yang di antaranyaterdapat proses mengolahdanmengekstrak data tekstualsecaraotomatisuntukmendapatkaninformasi. Manfaatanalisissentimendalamduniapolitikantara lain untukmembantudalammenganalisiskebijakanpublikpemerintahsertamemberikanefisiensiwaktudanefisiensikerjabagi para penyediaberitadalammengklasifikasikanberitadanmembantu para pencariberitauntukmendapatkanwacanaberitapolitikharian yang merekainginkan. Proses padaanalisissentimendiawalidengan preprocessing, dilanjutkandenganpembobotan kata, kemudianpenghitungan cosine similarity, danklasifikasi. Preprocessing terdiridaribeberapatahapyaitu cleansing, tokenizing, stopword removal, dan stemming. Metodepembobotan kata yang digunakanadalah Term Frequecny Inverse Document Frequency (TF-IDF) danmenggunakan Support Vector Machine (SVM). Naive Bayes Classifier (NBC), sebagaimetodeklasifikasinya. Adalahsuatumetodepengklasifikasianberdasarkanmayoritasdari polarity documensubjectifity yang di hasilkandari crawling. Metodeinibertujuanuntukmengklasifikasiobjekbaruberdasarkanatributdan training sample. PengujianakurasidariAnalisisSentimenTerhadapWacanaPolitikPada Media Masa Online BerbahasaInggrisdenganmetode NB menghasilkan rata-rata akurasisebesar 59,98 % dannilaitertinggiakurasisebesar SVM 90,50%.

Kata Kunci: AnalisisSentimenWacanaPolitikPada Media Masa Online, Text Mining, SVM (Support Vector Machine), NBC (Naive Bayes Classifier).

1. Pendahuluan

Analisis sentimen atau *opinion* mining mulai populer pada tahun 2002 mempublikasikan ide di balik penelitiannya yang di lakukan analisis sentimen adalah proses menyajikan informasi dengan membangun sebuah sistem yang dapat mengklasifikasikan dokumen teks ke dalamdua katehgori, yakni nilai positif dan negatif yang sesuai dengan keseluruhan sentimen yang dinyatakan di dalamsetipa dokumen tersebut. Dalam sebuah metode klasifikasi dalam analisis sentimen menggunakan metode-metode klasifikasi yang biasa di gunakan untuk kategorisasi teks antara lain metode *supervised learningmachine*(SVM) *Based Approach* maupun *Maximum Entropy*.

Website adalah tempat yangbaik bagi orang-orang untuk mengekspresikan pendapat mereka pada berbagai topik salah satunya adalah memanfaatkan situs jejaring sosial misalnya facebook, twitter, bahkan pemberi opini secara profesional, seperti reviewer berita politik dan film, pemilik *blog* dimana publik dapat mengomentari dan merespon apa yang mereka pikirkan. Kemampuan untuk merangkak dari website serta mengekstrak pendapat dari baris-baris teks dapat menjadi sangat berguna bidang ini adalah area studi yang banyak dikaji karena kemungkinan nilai komersialnya.

Kebanyakan informasi disimpan sebagai teks, sehingga *text mining* memiliki potensi nilai komersial*Sentiment analysis* atau *opinion mining*

JURNAL ELEKTRONIK SISTIM INFORMASI DAN KOMPUTER (JESIK) VOL.1 No.1 Januari-Juni 2015
ISSN

adalah studi komputisional dari opini-opini orang, sentimen dan emosi melalui entitas dan atribut yang dimiliki yang diekspresikan dalam bentuk teks.Analisis sentimen akan mengelompokkan polaritas dari teks yang ada dalam kalimat atau dokumen untuk mengetahui pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif, negatif atau netral.

Text mining adalah proses mengambilinformasiberkualitastinggidariteksinformasiberkualitastinggibiasanyadiperolehmelaluipermalanpoladankecenderunganmelaluisaranasepertipembelajaranpolastatistik. Secaraumum proses *text mining* dapatmeliputikategorisasiteks*text clustering*, ekstraksikonsep/entitas, produksitaksonomi granular, *sentiment analysis*, penyimpulandokumen, danpemodelanrelasientitas.Salah satumetodeklasifikasi yang dapatdigunakanadalahmetode Naive Bayes yang seringdisebutdengan*Naive Bayes Classifier* (NBC). Kelebihan NBC adalahsederhanatetapimemilikiakurasi yang tinggi. NBC terbuktidapatdigunakansecaraefektifuntukmengklasifikasikanberitasecaraotomatis. Algoritma NBC yang sederhanadankecepatannya yang tinggidalam proses pelatihandanklasifikasimembuatalgoritmainimenarikuntukdigunakansebagaisalahsatumetodeklasifikasiKeduaalgoritmaklasifikasitersebutbanyakdigunakandalamkategorisasiteks. PadahasileksperimenUntukkategorisasiteksberbahasaindonesia didapatkanbahwa SVM menunjukkanperformansi yang sedikitlebihbaikdibandingkanmetode NBC. Metode NBC adalahmetode yang jauhlebihsederhanadanmudahdiaplikasikan. Sehinggapadapenelitianiniingindiketahuimetode yang manamemilikiperformansi yang lebihbaikuntuk di implementasikandalam*sentiment analysis* terhadapwacanapolitik media masaindonesiaberbahasaInggrisDalampengklasifikasiansentimenanalisisdaritahun2004 sampaihadirdenganmenggunakan google trend halinidapat di amatibahwadenganpertumbuhan yang tersedia data di internet, kebutuhanuntukanalisissentimenjugameningkat, berbagaistatistikdan linear metodeguistictelah di kembangkanuntukanalisissentimentekswacanapolitikberbahasinggris.

2. Penelitian Terkait

Penelitian mengenai klasifikasi sentimen analisis pada wacana politik salah satu permasalahannya adalahtingkat pengklasifikasian sebuah teks kalimat berbahasa inggris yang dimana dimensi sebuah teks opini wacana politik ini yang di analisa terdapat ambigu dalam penggunaan kata, tidak adanya intonasi dalam sebuah teks sehingga menyebabkan banyak atribut yang kurang spesifik dan relevan sehingga menurunkan kinerja dan performa klasifikasi teks opini sentimen analisis opini wacana politik.Untuk mendapatkan accurasy yang baik, atribut yang ada harus di pilih dengan algoritma yang tepat. Bagian yang penting untuk mengoptimalkan sebuahklasifikasi dokumen teks adalah menggunakan *Feature Selection* salah satunya, yaitu unigram, unigram + bigram, unigram + Part of Speech (POS), adjective, danu nigram dan dikombinasikan dengan unigram.

Berdasarkan beberapa penelitian sentimen analisis menkombinasikan beberapa algoritma *Feature Selection* dan algoritma untuk mendapatkan hasil yang baik dan *performance* yang baik. Penelitian *sentiment analysis* yang dilakukanoleh AhmedAbbasi, Hsinchun Chen, & Arab Salem berjudul*Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums*di gabungkanmetodehybridisasialgoritmagenetika EWGA mendapatkanhasil yang lebihbaik.Sedangkan*Bayesian Opinion Mining*dilakukan oleh Ian Barberpada data review filmberbahasainggrisdandiujikanuntuk 5000 record opini negative dan5000record opinipositifsebagai data latihdan 333 record opininegatifsebagai data

ujisertamenghasilkanakurasisebesar 80% menggunakan metode naive bayes classifier sedangkan KlasifikasiBeritaBerbahasa Indonesia Menggunakan*Naive Bayes Classifier,dilakukan oleh*yudiWibisono. Sedangkan penelitian yang berjudul *text mining* dengan metode *naive bayes classifier support vector machine*untuk analisis oleh ni wayan sumartini saraswati dalam penelitiannya menguji data sebanyak 3000 record positif dan negatif menggunakan metode *Naive Bayes Classifier*80.18% dan *support vector machine* 80.15%.sedangkan pada penelitian yang dilakukan Fatimah WulandinidanAntoSatriyoNugroho dengan judul*Text Classification Using SupportVector MachineforWebmining BasedSpation Temporal Analysis of theSpread of Tropical Diseases,* mennggunakan 4 metode klasifikasi dan mengcomparasikan dengan metode klasifikasi yang digunakan yakni algoritma SVM, NBC, KNN dan C45 hasilnya pada data 3713feature dan360 instance.360 instance sebagaidata latihdan 120 instance.sedangkan pada penelitian Fabrice Colas &PavelBrazdil yang berjudul c*omparison ofSVM and Some OlderClassification Algorithms inTextClassification Tasks,*melakukan pengujian 3 metode dengan menggunakan algoritmaSVM, KKN dan NBC.

Sedangkan pada penelitian Blitzer, J., Dredze, M. & Pereira yang berjudul *Biographies, Bollywood,Boom-boxesand Blenders: Domain Adaptation forSentiment Classification,*menguji dokumen bahwadengan menggunakan metodeStructuralcorrespondenceelearning(SCL),baseline,SCLMI SCLMImenunjukkanperformansiyang lebihbaikuntukadaptasidomain.hasil eksperimen menunjukkan bahwa menguji data sentimen analisis wacana politik media masa online mengunakan algoritma *naive bayes*hasilnya mendapatkanakurasisebesar 59,98 % dan membandingkan dengan algoritma *support vector machine*mendapatkan hasil yang lebih baik mendapatkan akurasi90,50%.

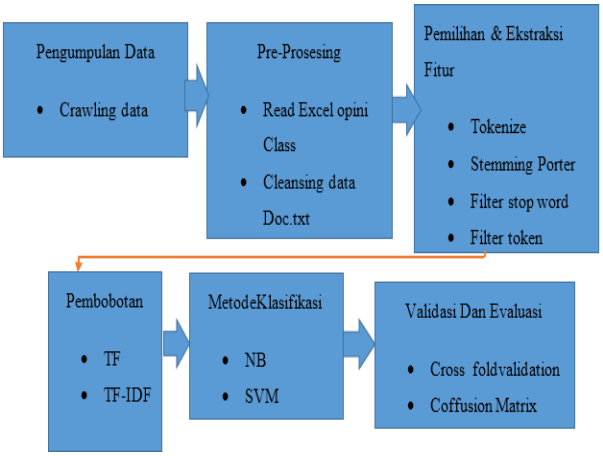
Tabel 1Hasil Akurasi algoritma klasifikasi

Algorithma	Accurasy
NB	59,98 %
SVM	90,50%

3. METODE YANG DI USULKAN

Peneliti mengusulkan untuk mengkomparasikan 2 algoritma klasifikasi yakni (naive bayes classifier dan support vector machine)dan menggunakan fitur ekstraksi dokumen (tokenize, stemming porter, filter token dan stop word). Peneliti menguji data tiap-tiap dokumen berita yang di kumpulkan secara online melalui teknik web mining yakni tools crawling data melalui situs portal berita.Data yang sudah di bersihkan yang berisi tag-tag, doc, html, dan untuk menghitung pembobotan teks / term adalah menggunakan *space vector* pembobotan TF-IDF serta menganalisa hasil performa suatu metode dalam klasifikasi dokume harian politik dan meghasilkan masing-masing tingkat akurasi algoritma *Naive Bayes*Dan *support Vector Machine*.Dalam pemilihan fitur ekstrasi dokumen teks wacana harian politik media masa online pada situs portal berita ini. Sebelum melakukan komparasi/kombinasi dataset di lakukan *textprocesing*terlebih dahulu*text procesing*bertujuan untuk mempersiapkan dokumen teks yang tidak terstruktur menjadi data terstruktur yang siap di gunakan untuk procesing data selanjutnya.

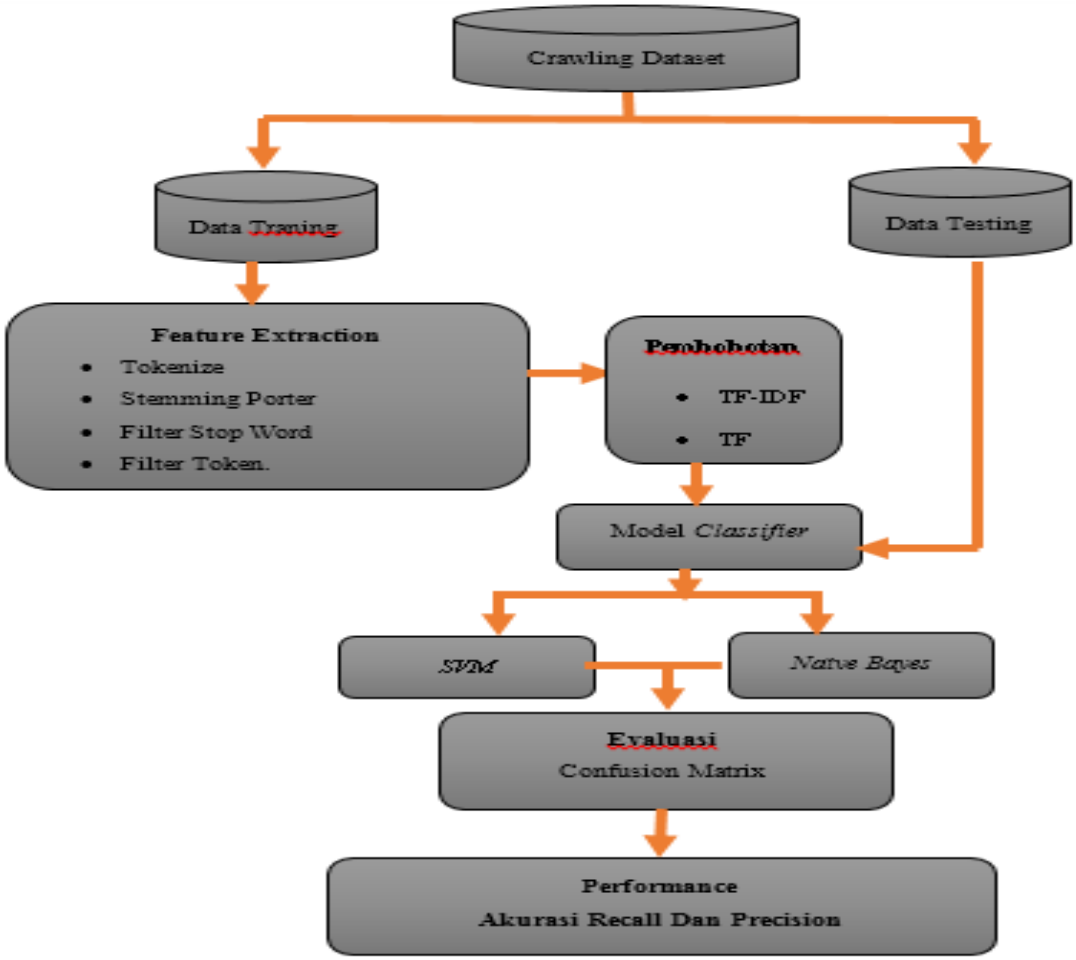
Adapun beberapa tahap-tahap dan implementasi text procesing yangmeliputi:
1.Tokenize merupakan proses untuk memisah-misahkan kata. Pemotongan kata tersebut yang sering disebut token term.
2.Filter Token merupakan pengambilan/menyaring sebuah kata yang berkarakter misal di input nilai karakter 3 maka panjang dlam sebuah karakter kata akan di filter menjadi panjang 3 karakter sesuai panjang karakter yang diinputkan.
4. Stemming yaitu proses menghilangkan



kata-kata yang tidak penting dalam teks namun sering meuncul yang tidak memilki pengaruh apapun dalam proses ekstraksi

JURNAL ELEKTRONIK SISTIM INFORMASI DAN KOMPUTER (JESIK) VOL.1 No.1 Januari-Juni 2015
ISSN

sentimen suatu preview. Misalnya kata yang termasuk kata penunjuk waktu dan kata



tanya.

Gambar. 2 Diagram Alur Klasifikasi

Gambar.1 Desain Eksperimen

4. Coffusion Matrix

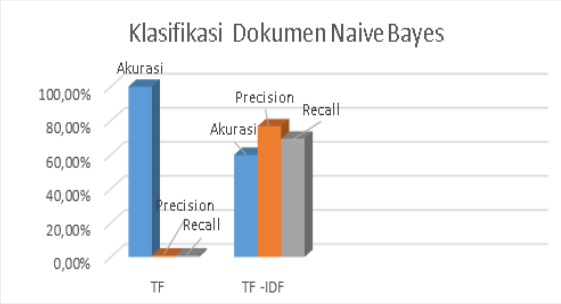
Precision dan *Recall* adalahduaperhitungan yangbanyakdigunakanuntukmengukurkinerjadarisistem/ metode yang digunakan. *Precision* adalahtingkatketepatanantarainformasiyang dimintaolehpenggunadenganjawaban yang diberikanolehsistem.Sedangkan Recall adalahtingkatkeberhasilansistemdalammenemukankembalisebuahinformasi. Sedangkan di dunia lain sepertiduniastatistikadikenaljugaistilah *accurasy*. Akurasidi definisikansebagaitingkatkedekatanantaranilaipre

Tabel2.Cuffusion Matrix NB

Dokumenaktual	true negatif	true positif	class precision
aktual. negatif	54	214	20.15%
aktual. positif	147	487	76.81%
class recall	26.87%	69.47%	

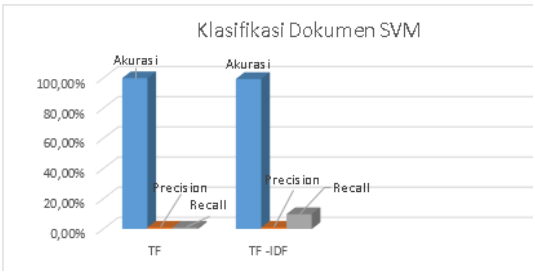
Total Akurasi: 59.98% + /-4.42% (mikro: 59.98%)

JURNAL ELEKTRONIK SISTIM INFORMASI DAN KOMPUTER (JESIK) VOL.1 No.1 Januari-Juni 2015
ISSN



diksidengannilaiaktual.Untuk proses evaluasiinidimanaakanmemberikanakurasiatauper formance teks sentimen yang telahdilakukan. Untuk proses evaluasiinimenggunakan*confusion matrix* yang dimanaakanmemudahkanpenelitiuntukmendapatk antingkatakurasidariklasifikasi dokumen teks sentimen analisis berita politik.

Gambar2. Klasifikasi dokumen NB



Gambar3. Klasifikasi dokumen SVM

Untukmencariakurasi, digunakan alat ukur coffusion matrix precision dan recall dapat di gunakanrumussebagaiberikut :

$$Accuracy = \left(\frac{a+d}{a+b+c+d}\right) = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

Dokumen Aktual	true negatif	true positif	Class precision
aktual. Negative	100	19	84.0%
aktual. Positif	0	81	100.00%
class recall	100.00%	81.00%	

$$Recall = \frac{TP}{TP+FN}$$

Keterangan :
TP = True positif yang positif
FN = False positif yang negatif
FP = False negatif yang positif
TN = True negatif yang negatif

Tabel 3.Coffusion Matrix SVM

Total Akurasi: 90.50% +/- 6.87%(mikro: 90.50%)

Data yang di gunakan dalam penelitian ini juga berasal website thejakartapost.com tahun 2014 dengan menggunakan metode crawling dan menggunakan pendekatan text mining. Dataujicobaadalahdokumen 700 positifdan dokumen 700negatifhasilpenambahandari data yang sebelumya 200 example menjadi 700 data di antaranyanegatif danpositif total data 1400 data example data berita.

5. Hasil Penelitian

Hasilpenelitianyang dilakukan menggunakan spesifikasi komputer AMD E- 450 APU dan sistem operasi windows 7-Ultimate 32–bit.Aplikasi yang di gunakan adalah rapidminer 5.3. Data penelitian ini di ambil dari situs portal di khususkan pada dokumen teks berbahasa inggris situs berita yang digunakan pada penelitian ini dariwww.thejakartapost.com/channel/opinion.yait u kumpulan teks opini-opini berita politik.Pada pengolahan data dari beberapa halaman tag web dan dilakukankan crawling data pada website www.thejakartapost.com/channel/opinion dengan menggunakan bantuan tools dokumen subjectivity collection yaitu data kumpulan hasil crawling website dalam satu folder dan di kumpulkan dalam bentuk ekstensi.txt sebanyak 700 record data positif negatif yang terdiri dari beberapa politik harian tersebut dan di tentukan beberapa analisis sentimen di antaranya positif, netral, dan negatif dan melakukan proses klasifikasi dan ekstraksi sebuah data. Analisis sentimen wacana politik harian tersebutujicobamengklasifikasikanopiniwacanapol itikharianberbahasainggrisdengan data opini yang di dapatkannyasecara online portal beritatersebutkemudian di proses data yang berupateksopiniberitaberbahasainggrismelakukan preprocessing teksantara lain (pemecah kata), (menghilangkan kata tidakpentingdalamteks), (menyaringkata menjadikarakterpanjang),(mengurangi kata-kata dasaratauinduknya) dilanjutkanmencariperformaakurasidenganmengg

unakanmetodealgoritma Naive Bayes dan Support Vector Machine hasilnyaakanberupaefektifitaskalimatopinipositif, netral, negatifterhadapwacanapolitiktersebut. Berdasarkanpengetahuanseorangahliuntukmenentukan class/label positifnetraldannegatif di lakukansecara manual berdasarkanseorangahliwartawanmisbah (redaksikoranharianumummercusuar), yaitupenentuansecara manual kalimatsentimenanalisisdenganmenggunakan 3 klasifikasi class/ label berikut:

- pos: Politik yang membawa sentimen positif terhadap topik.
- neg: Politik yang membawa sentimen negatif terhadap topik.
- campuran: Politik yang membawa kedua positif dan negatif sentimen terhadap topik.

6. Kesimpulan

Dari hasilpenelitianinimenunjukkanbahwapenggunaan metode algoritma SVM dapatmengimplementasikanhasildaripengujian data beritawacanapolitikdandaripenggunaan Naive Bayes dapat di hasilkanperbedaandenganalgoritma SVM padamasing - masing data yang telah di analisisdandarihasilpengukuranmenggunakanteknikpengukuran Precision, Recall dan di dapatitingkatakurasiklasifikasiberitapolitikmasing - masing. Sehinggapadahasilanalisatersebutdapat di ambilkesimpulanbahwapenerapanalgoritma SVM menunjukkanakurasilebihbesar 90,50% SedangkanNaive Bayes di uji data sebanyak 700 data 59,98% dalamhasilklasifikasiopiniwacanapolitiksertadenganadanyahasilanalisistersebutdapatmemperolehklasifikasiopini yang baik.

Reference

[1] Barber, I. 2009. *Support Vector Machine* In PHP.

[2] Brige, C.2011. Unstructured Data and the 80 Percent Rule**.**

[3] Liu B. Sentiment analysis and opinion mining, Synth Lect Human Lang Technol (2012).

[4] Pang, B. & Lee, L. 2008. *Subjectivity Detection and Opinion Identification*. *Opinion Mining and Sentiment Analysis***.**

[5] BoPangandLillianLee*Foundations and Trends in Information Retrieval* 2(1-2), pp.1–135, 2008.Alsoavailableasabook or e-book.

[6] Wibisono,Y. 2005. Klasifikasi Berita Berbahasa Indonesia menggunakan Naïve Bayes Classifier. diases 29 September 2012).

[7] Saraswati, N.W.S., 2011, *Text Mining denganMetode Naive Bayes Classifier dan SupportVectorMachine untukSentimen Analysis*.

[8] Wulandini, F. &Nugroho, A. N. 2009. *Text Classification Using Support Vector MachineforWebmining Based Spation Temporal Analysis of the Spread of Tropical Diseases*.International

[9] Tan, P. N., Steinbach, M. & Kumar, V. 2006. Introduction to Data Mining. Boston : Pearson Addison Wesley.

[10] Pang, B. & Lee, L. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proceedings of the Association for Computational Linguistics (ACL).pp.271–278.

[11] Parikh, R., &Movassate, M. (2009). Sentiment Analysis of User Generated Twitter Updates using Various Classification Retrieved November 12, 2011, from CS224N Final Projects 2008-9**.**

[12] Prasad,S.(n.d.).Microblogging SentimentAnalysis Using Bayesian Classification Methods. Retrieved November 29, 2011, from The Stanford NLP (Natural Language Processing).

[13] Read, J. (2005). Using Emoticons to Reduce Dependency in MachineLearning Techniques for Sentiment Classification.ACLstudent '05 Proceedings of the ACL Student Research Workshop(pp. 43-48). Stroudsburg: Association for Computational Linguistic.

[14] Naradipha, A. R., &Purwarianti, A. (2011). Sentiment Classificationfor Indonesian Message in Social Media.InternationalConference on Electrical Engineering and Informatics, (pp.14).

[15] Bo Pang and Lilian Lee. 2008. Opinion Miningand SentimentAnalysis, Foundations and Trends in Information Retrieval, vol. V olume 2, no. Issue 1-2, pp. 1-135.

[16] Blitzer, J., Dredze, M. & Pereira, F. 2006. *Biographies, Bollywood, Boom boxesand Blenders: Domain Adaptation for Sentiment Classification*.

[17] Feldman, R & Sanger, J. 2007. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press: New York.

[18] Berry, M.W. &Kogan, J. 2010. Text Mining Aplication and theory. WILEY: United Kingdom

[19] Indah Tri, R. 2010. PembuatanJudulOtomatisDokumenBeritaBerbahasa Indonesia MenggunakanMetode K-Nearest Neighbor. Prodi IlmuKomputer, UniversitasBrawijaya.

[20] Dehaff, M. 2010. Sentiment Analysis, Hard But Worth It.

[21] Pang, B. & Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. Proceedings of the Association for Computational Linguistic.

[22] Snyder B. &Barzilay R. 2007 Multiple Aspect Ranking using the Good Grief Algorithm. Proceedings of the Joint Human Language Technology/North American Chapter of the ACL.

[23] Zhang, H. 2004. The Optimality of Naive Bayes. FLAIRS2004 conference.

[24] Caruana, R. &Niculescu-Mizil, A. 2006. Anempirical comparison of supervised learning algorithms. Proceedings of the 23 rd international conference on Machine learning,