

# ANALISIS SENTIMENT PADA SOSIAL MEDIA TWITTER MENGGUNAKAN NAÏVE BAYES CLASSIFIER DENGAN FEATURE SELECTION PARTICLE SWARM OPTIMIZATION DAN TERM FREQUENCY

Yono Cahyono

Teknik Informatika UNPAM

Email : dosen00843@unpam.ac.id

## ABSTRAK

Pengguna media sosial saat ini sangat besar, dimana setiap orang mengungkapkan pendapat, komentar, kritik dan lain-lain. Data tersebut memberikan informasi yang berharga untuk dapat membantu orang atau organisasi dalam pengambilan keputusan. Jumlah data yang sangat besar tidak mungkin bagi manusia untuk membaca dan menganalisis secara manual. Analisis Sentiment merupakan proses dalam menganalisis, memahami, dan mengklasifikasi pendapat, evaluasi, penilaian, sikap, dan emosi terhadap suatu entitas tertentu seperti produk, jasa, organisasi, individu, peristiwa, topik, guna mendapatkan informasi. Penelitian ini bertujuan untuk memisahkan tweets berbahasa Indonesia pada media sosial twitter kedalam kategori positif, negatif dan netral. Metode naïve bayes Classifier (NBC) dengan feature selection Particle Swarm Optimization (PSO) diterapkan pada dataset untuk mengurangi atribut yang kurang relevan pada saat proses klasifikasi. Hasil pengujian menunjukkan bahwa algoritma Naïve Bayes Classifier dengan feature selection Particle Swarm Optimization (PSO) menggunakan parameter term frequency (TF) dengan akurasi 97,48%.

Kata kunci: Analisis Sentiment, Twitter, NBC, Term frequency, PSO.

## 1. PENDAHULUAN

Pertumbuhan media *online* seperti media sosial *Twitter* mendorong munculnya informasi tekstual yang tidak terbatas, sehingga muncul kebutuhan untuk menggali nilai informasi tersebut. Informasi tekstual dikategorikan menjadi fakta dan opini. Fakta merupakan ekspresi objektif mengenai suatu entitas, kejadian atau sifat, sedangkan opini adalah ekspresi subjektif yang menggambarkan sentimen seseorang, pendapat atau perasaan tentang sebuah entitas, kejadian atau sifat. Pada *Twitter* terdapat istilah bernama *tweet* yang merupakan sebuah pesan atau status yang dibuat oleh penggunanya. Sebuah *tweet* dapat mengekspresikan sebuah perasaan atau keadaan dari pengguna *Twitter*. *Tweet* dapat mengandung sebuah opini dari penggunanya terhadap kejadian yang dialaminya. Opini tersebut dapat dimanfaatkan sebagai penilaian bagi sebuah perusahaan atau instansi.

*Sentiment Analysis* merupakan bagian dari *Natural Language Processing* (NLP) dan *Machine Learning*. Cara kerjanya adalah dengan melakukan klasifikasi terhadap opini positif, negatif dan netral. *Sentiment Analysis* dalam menganalisa pendapat orang, sentimen, evaluasi, penilaian, sikap, dan emosi terhadap entitas

seperti produk, jasa, organisasi, individu, masalah, peristiwa atau suatu topik.

Dari beberapa teknik klasifikasi yang paling sering digunakan untuk klasifikasi data adalah metode *Naïve Bayes* yang sering disebut dengan *Naïve Bayes Classifier* (NBC). Algoritma *Naïve Bayes* dipilih dikarenakan algoritma ini sangat cocok untuk *short data text*. Kelebihan NBC adalah metode sederhana tetapi memiliki akurasi serta performansi yang tinggi dalam pengklasifikasian teks (Routray, *et al*, 2013).

Seleksi fitur adalah salah satu faktor yang paling penting yang dapat mempengaruhi tingkat akurasi klasifikasi karena jika *dataset* berisi sejumlah fitur, dimensi ruang akan menjadi besar, merendahkan tingkat akurasi klasifikasi. Seleksi fitur merupakan proses optimasi untuk mengurangi satu set besar fitur besar sumber asli agar subset fitur yang relatif kecil yang signifikan untuk meningkatkan akurasi klasifikasi cepat dan efektif.

Pemilihan fitur sekaligus penyetingan parameter di NBC secara signifikan mempengaruhi hasil akurasi klasifikasi. Maka dari itu dalam penelitian ini menggunakan penggabungan metode pemilihan fitur dan NBC. *Feature selection* yang digunakan pada penelitian

ini adalah *Particle Swarm Optimization* (PSO) dengan parameter *term frequency* (TF).

## 2. PENELITIAN TERKAIT

*Sentiment Analysis* telah dilakukan oleh beberapa peneliti. Analisis sentimen menggunakan teknik pembelajaran mesin *Naïve Bayes classifier* dengan seleksi fitur *chi-square*, untuk mengurangi gangguan (*noise*) dalam klasifikasi. Hasil penelitian menunjukkan bahwa frekuensi kejadian fitur yang diharapkan dalam kategori yang benar dan dalam kategori palsu memiliki peran penting dalam pemilihan fitur *chi-square*. Kemudian dengan klasifikasi memperoleh akurasi 83% dan rata-rata harmonik 90,713% (Juen Ling, dkk, 2014).

Penelitiannya menggabungkan metode SVM dan PSO untuk mengklasifikasikan pendapat positif dan negatif. Metode *hybrid* (SVM - PSO) telah meningkatkan akurasi, dibandingkan dengan menggunakan metode SVM saja (K.Umamaheswari, Ph.D, et al, 2015).

Analisis Sentiment Pada Sosial Media *Twitter* Menggunakan *Naive Bayes Classifier* Terhadap Kata Kunci “Kurikulum 2013” (Dyarsa Singgih Pamungkas, dkk, 2015). Dalam penelitian ini menggunakan *Twitter Search API* untuk mengambil data dari *twitter*, penulis menerapkan proses *n-gram* karakter untuk seleksi fitur serta menggunakan algoritma *Naive Bayes Classifier* untuk mengklasifikasi sentimen secara otomatis. Penulis menggunakan 3300 data *tweet* tentang sentimen kepada kata kunci “kurikulum 2013”. Data tersebut diklasifikasi secara manual dan dibagi kedalam masing-masing 1000 data untuk sentimen positif, negatif dan netral. Untuk proses latih di gunakan 3000 data *tweet* dan 1000 *tweet* tiap kategori sentimentnya. Hasil penelitian ini menghasilkan sebuah sistem yang dapat mengklasifikasi sentimen secara otomatis dengan hasil pengujian 3000 data latih dan 100 *tweet* data ujicoba mencapai 91 %.

Analisis Sentimen *Twitter* untuk Teks Berbahasa Indonesia dengan *Maximum Entropy* dan *Support Vector Machine* (Noviah Dwi Putranti & Edi Winarko, 2014). Analisis sentimen dalam penelitian ini merupakan proses klasifikasi dokumen tekstual ke dalam dua kelas, yaitu kelas sentimen positif dan negatif. Data opini diperoleh dari jejaring sosial *Twitter* berdasarkan query dalam Bahasa Indonesia. Penelitian ini bertujuan untuk menentukan sentimen publik terhadap objek tertentu yang disampaikan di *Twitter* dalam bahasa Indonesia, sehingga membantu usaha untuk melakukan riset

pasar atas opini publik. Data yang sudah terkumpul dilakukan proses *preprocessing* dan *POS tagger* untuk menghasilkan model klasifikasi melalui proses pelatihan. Teknik pengumpulan kata yang memiliki sentimen dilakukan dengan pendekatan berdasarkan kamus, yang dihasilkan dalam penelitian ini berjumlah 18.069 kata. Algoritma *Maximum Entropy* digunakan untuk *POS tagger* dan algoritma yang digunakan untuk membangun model klasifikasi atas data pelatihan dalam penelitian ini adalah *Support Vector Machine*. Fitur yang digunakan adalah *unigram* dengan fitur pembobotan TFIDF. Implementasi klasifikasi diperoleh akurasi 86,81 % pada pengujian *7 fold cross validation* untuk tipe *kernel Sigmoid*. Pelabelan kelas secara manual dengan *POS tagger* menghasilkan akurasi 81,67%.

## 3. METODE YANG DIUSULAKN

Pemilihan fitur sekaligus penyetingan parameter di NBC secara signifikan mempengaruhi hasil akurasi klasifikasi. Maka dari itu dalam penelitian ini menggunakan penggabungan metode pemilihan fitur dan NBC. *Feature selection* yang digunakan pada penelitian ini adalah *Particle Swarm Optimization* (PSO) dengan parameter *term frequency* (TF).

Dilakukan *text processing* pada dataset terlebih dahulu. *Text preprocessing* berfungsi untuk mengubah data teks yang tidak terstruktur atau sembarang menjadi data yang terstruktur.

Proses yang dilakukan dalam tahapan *preprocessing* adalah sebagai berikut:

### a. Case Folding

*Case folding* adalah proses penyamaan *case* dalam sebuah dokumen. Hal ini dilakukan untuk mempermudah pencarian. Tidak semua dokumen teks konsisten dalam penggunaan huruf kapital. Oleh karena itu peran *case folding* dibutuhkan dalam mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (dalam hal ini huruf kecil atau *lowercase*).

### b. Tokenizing

*Tokenizing* adalah proses pemotongan sebuah dokumen menjadi bagian-bagian, yang disebut dengan token. Pada saat bersamaan *tokenizing* juga berfungsi untuk membuang beberapa karakter tertentu yang dianggap sebagai tanda baca.

c. *Stopword Removal*

*Stopword removal* adalah proses penghilangan kata-kata yang tidak berkontribusi banyak pada isi dokumen. Kata-kata yang termasuk ke dalam *stopword* dihilangkan karena memberikan pengaruh yang tidak baik dalam proses text mining seperti kata-kata “bagaimana”, “juga”, “agar”, “jadi” dan lain-lain.

Tahapan berikutnya adalah *feature selection*. Tahapan ini merupakan tahapan penting dalam text mining. Salah satu fungsi penting yang disediakan oleh proses ini adalah untuk dapat memilih *term* atau kata apa saja yang dapat dijadikan sebagai wakil penting untuk kumpulan dokumen yang akan kita analisis. Pada penelitian ini akan menggunakan *feature selection* yaitu *Particle Swarm Optimization* (PSO) dengan parameter *term frequency* (TF).

*Particle Swarm Optimization* (PSO) dimulai dengan suatu populasi yang terdiri dari sejumlah *term* (yang menyatakan solusi) yang dibangkitkan secara acak dan selanjutnya melakukan pencarian solusi optimum melalui perbaikan *term* untuk sejumlah kategori tertentu. Pembangkitan posisi ( $x_i, d$ ) dan kecepatan ( $v_i, m$ ) dari kumpulan partikel dibangkitkan secara *random* menggunakan batas bawah ( $X_{\min}$ ) dan batas atas ( $X_{\max}$ ). Untuk *update velocity* (kecepatan) untuk semua *term* dari nilai *fitness* dapat ditentukan *term* mana yang memiliki nilai *global* terbaik (*global best*) dan juga dapat ditentukan posisi terbaik (*local best*) dari tiap *term* pada semua waktu sekarang dan sebelumnya. Akan melakukan pengulangan sampai kriteria terpenuhi.

TF-IDF (*Term Frequency - Inverse Document Frequency*) adalah statistik numerik yang menunjukkan pentingnya kata pada dokumen (Rajaraman, Leskovec, & Ullman, 2011). Umumnya TF-IDF digunakan sebagai faktor untuk menghitung bobot pada pengambilan informasi. Nilai TF-IDF meningkat setiap banyaknya kata muncul pada dokumen, namun turun apabila frekuensi kata sering muncul, hal ini untuk menangani kata-kata yang sering muncul. Oleh karena itu, bobot yang dihasilkan dari TF-IDF dapat dijadikan salah satu fitur untuk melakukan *Clustering*/pengelompokan dari kata.

TF dapat ditunjukkan dengan 3 cara (Salton & Buckley, 1988), yaitu :

1. Menggunakan nilai biner, yaitu diberi nilai 1 untuk kata-kata yang terdapat pada dokumen dan nilai 0 terhadap kata-kata yang tidak muncul pada dokumen. Pada konsep ini, frekuensi kemunculan kata tidak dimasukkan ke dalam perhitungan.
2. Menggunakan nilai frekuensi kemunculan kata secara langsung untuk menjadi TF.
3. Menggunakan nilai pecahan *term* yang telah dilakukan normalisasi, rumus tersebut adalah (Salton & Buckley, 1988):

$$TF(t, d) = 0,5 + 0,5 \frac{f(t,d)}{\max\{f(w,d):w \in d\}}$$

dimana ( $t, d$ ) merupakan frekuensi kata  $t$  muncul pada dokumen  $d$  dan  $\max\{f(w,d):w \in d\}$  merupakan frekuensi maksimum dari term lain pada dokumen  $d$ .

Sedangkan untuk menghitung IDF digunakan rumus sebagai berikut (Salton & Buckley, 1988):

$$IDF(t, d) = \log \frac{N}{Df(t,d)}$$

dimana  $N$  merupakan jumlah dokumen dan  $Df(t, D)$  sebagai banyak dokumen dalam kumpulan dokumen  $D$  yang mengandung *term*  $t$ . Namun bila *term* tidak muncul, maka akan terdapat nilai 0 pada pembagian, sehingga perlu penanganan untuk menggantinya menjadi  $1 + Df(t, D)$ .

Untuk menghitung TF-IDF dari kata, digunakan rumus :

$$TF - IDF(t, d, D) = tf(t, d) \times idf(t, D)$$

Dari rumus diatas akan didapatkan nilai yang dapat dijadikan sebagai pembobotan kata pada saat dilakukan pengelompokan kata.

Tahap *Cross Validation*, setiap *record* digunakan beberapa kali untuk *training* dan untuk testing. Untuk mengilustrasikan metode ini, anggaplah mempartisi data ke dalam dua *subset*. Pertama, dipilih satu dari kedua *subset* tersebut untuk *training* dan satu lagi untuk *testing*. Kemudian dilakukan pertukaran fungsi dari *subset* sedemikian sehingga *subset* yang sebelumnya sebagai *training set* menjadi *testing set* demikian sebaliknya.

Metode ini merupakan evaluasi standard yaitu *stratified 10-fold cross-validation* karena menunjukkan bahwa *10-fold cross-validation* adalah pilihan terbaik untuk mendapatkan hasil validasi yang akurat, *10-fold cross-validation* akan mengulang pengujian sebanyak 10 kali dan hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian.

Tahap Klasifikasi Algoritma *naive bayes classifier* merupakan algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasikan data uji pada kategori yang paling tepat (Feldman & Sanger, 2007).

Dalam algoritma *naive bayes classifier* setiap dokumen direpresentasikan dengan pasangan atribut " $x_1, x_2, x_3, \dots, x_n$ " dimana  $x_1$  adalah kata pertama,  $x_2$  adalah kata kedua dan seterusnya. Sedangkan  $V$  adalah himpunan kategori *Tweet*. Pada saat klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan ( $V_{MAP}$ ), dimana persamaannya adalah sebagai berikut :

$$V_{MAP} = \underset{V_j \in V}{\operatorname{argmax}} \frac{P(X_1, X_2, X_3, \dots, X_n | V_j) P(V_j)}{P(X_1, X_2, X_3, \dots, X_n)}$$

Untuk  $P(x_1, x_2, x_3, \dots, x_n)$  nilainya konstan untuk semua kategori ( $V_j$ ) sehingga persamaan dapat ditulis sebagai berikut :

$$\underset{V_j \in V}{\operatorname{argmax}} P(x_1, x_2, x_3, \dots, x_n | V_j) P(V_j)$$

Persamaan diatas dapat disederhanakan menjadi sebagai berikut :

$$V_{MAP} = \underset{V_j \in V}{\operatorname{argmax}} \prod_{i=1}^n P(x_i | V_j) P(V_j)$$

Keterangan :

$V_j$  = Kategori *tweet*  $j = 1, 2, 3, \dots, n$ . Dimana dalam penelitian ini  $j_1$  = kategori *tweet* sentimen negatif,  $j_2$  = kategori *tweet* sentimen positif, dan  $j_3$  = kategori *tweet* sentiment netral

$P(x_i | V_j)$  = Probabilitas  $x_i$  pada kategori  $V_j$

$P(V_j)$  = Probabilitas dari  $V_j$

Untuk  $P(V_j)$  dan  $P(x_i | V_j)$  dihitung pada saat pelatihan dimana persamaannya adalah sebagai berikut :

$$P(V_j) = \frac{|docs\ j|}{|contoh|}$$

$$P(X_i | V_j) = \frac{nk+1}{n+|kosakata|}$$

Keterangan :

$|docs\ j|$  = jumlah dokumen setiap kategori  $j$

$|contoh|$  = jumlah dokumen dari semua kategori

$nk$  = jumlah frekuensi kemunculan setiap kata

$n$  = jumlah frekuensi kemunculan kata dari setiap kategori

$|kosakata|$  = jumlah semua kata dari semua kategori

#### 4. HASIL PENELITIAN

*Particle Swarm Optimization* (PSO) digunakan sebagai seleksi fitur dan *Naive Bayes Classifier* untuk mengevaluasi subset fitur.

Pada penelitian ini dipilih menggunakan seleksi fitur *Particle Swarm Optimization* (PSO) menggunakan parameter *Term Frequency* (TF) dengan *Naive Bayes Classifier* karena hasil seleksi fitur didapatkan *terms* lebih banyak yaitu sebanyak 776 kata yang terseleksi, sedangkan seleksi fitur *Particle Swarm Optimization* (PSO) menggunakan parameter *TF-IDF* dengan *Naive Bayes Classifier* didapatkan *terms* sebanyak 774 kata yang terseleksi

Jumlah seleksi fitur dengan *terms* lebih banyak dijadikan sebagai wakil penting untuk kumpulan dokumen yang akan dianalisis.

Hasil pengujian seleksi fitur *Particle Swarm Optimization* (PSO) terbukti dapat meningkatkan akurasi algoritma *Naive Bayes Classifier*. Pengujian menggunakan seleksi fitur *Particle Swarm Optimization* (PSO) menggunakan parameter *term frequency* (TF) dengan *Naive Bayes Classifier* sebesar 97,48%.

#### 5. KESIMPULAN

Berdasarkan hasil analisis yang dilakukan dalam penelitian, dapat disimpulkan hal-hal berikut ini.

- Dengan seleksi fitur pada penelitian ini terbukti metode seleksi fitur *Particle Swarm Optimization* pada *dataset* terseleksi jumlah *terms* sebanyak 776 kata, dapat membantu proses klasifikasi *Naive Bayes* menjadi lebih efektif dan akurat.
- Hasil pengujian menunjukkan bahwa algoritma *Naive Bayes Classifier* dengan *feature selection* menggunakan *Particle Swarm Optimization* (PSO) dengan parameter *term frequency* (TF) mendapatkan akurasi 97,48%.

#### DAFTAR PUSTAKA

- Berry, M.W. & Kogan, J. 2010. *Text Mining Application and theory*. WILEY : United Kingdom.
- Barber, I. 2010. *Bayesian Opinion Mining*.
- Church, K. W., & Hanks, P. (1990, mar). *Word Association Norms, Mutual Information, and*

- Lexicography. Comput. Linguist.*, 16(March 1990), 22-29.
- Cortes, C. & Vapnik, V. 1995. *Support-Vector Networks. Machine Learning*, 20.
- Cawley, G. C. 2006. *Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs*. Proceedings of the IEEE/INNS International Joint Conference on Neural Networks (IJCNN-06).
- Darujati, Cahyo. 2010. Perbandingan Klasifikasi Dokumen Teks Menggunakan Metode *Naïve Bayes* Dengan *K-Nearest Neighbor*. *Jurnal Link* (Information Retrieval).
- Dehaff, M. 2010. *Sentiment Analysis, Hard But Worth It!*.
- Dennis Aprilla C, Donny Aji Baskoro, Lia Ambarwati, I Wayan Simri Wicaksana. 2013. Belajar Data Mining dengan RapidMiner.
- Dyarsa Singgih Pamungkas, Noor Ageng Setiyanto, Erlin Dolphina. 2015. Analisis *Sentiment* Pada Sosial Media *Twitter* Menggunakan *Naïve Bayes Classifier* Terhadap Kata Kunci Kurikulum 2013, Universitas Dian Nuswantoro.
- Feldman, R & Sanger, J. 2007. *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press : New York.
- Feldman, R & Sanger, J. 2007. *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press : New York.
- Frits Gerit John Rupilele. 2013. Danny Manongga, Wiranto Herry Utomo, *Sentiment Analysis of National Exam Public Policy with Naïve Bayes Classifier Method (NBC)* , Satya Wacana Christian University.
- Guo, Gongde. Wang, Hui. Bell, David. Bi, Yaxin. & Greer, Kieran. 2001. *Using K-NN Model-based Approach for Automatic Text Categorization*. Northern Ireland : School of Computing and Mathematics
- Gorunescu, F. 2011. *Data Mining Concepts, Model and Techniques*. Berlin: Springer.
- Han, J & Kamber, M. 2006. *Data Mining: Concepts and Techniques Second Edition*. Morgan Kaufmann publisher : San Francisco.
- Ian, H,Witten. Frank, Eibe & Mark A, Hall. 2011. *Data mining: Practical Machine Learning Tools and Techniques 3rd Edition*. Elsevier.
- Imam Fahrur Rozi, Sholeh Hadi Pramono dan Erfan Achmad Dahlan. 2012. Implementasi *Opinion Mining* (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi, Universitas Brawijaya.
- Imelda A.Muis1, Muhammad Affandes, M.T. 2015. Penerapan Metode *Support Vector Machine (SVM)* Menggunakan *Kernel Radial Basis Function (RBF)* Pada Klasifikasi *Tweet*, UIN Sultan Syarif Kasim Riau.
- Jenkins, M. C. 2011. *How Sentiment Analysis works in machines*.
- Juen Ling, I Putu Eka N. Kencana, Tjokorda Bagus Oka. 2014. Analisis Sentimen Menggunakan Metode *Naïve Bayes Classifier* Dengan Seleksi Fitur *Chi Square*, Universitas Udayana.
- Kumar, M. A. & Gopal, M. 2009. *Least Squares Twin Support Vector Machines for Pattern Classification*.
- Karatzoglou, Alexandros. Smola, Alexandros. Hornik, Kurtand & Zeileis, Achim. 2004. *kernlab - An S4 Package for Kernel Methods in R*. Journal of Statistical Software.
- K.Umamaheswari, Ph.D ; S.P.Rajamohana ; G.Aishwaryalakshmi. 2015. *Opinion Mining using Hybrid Methods*”, Department of Information Technology PSG College of Technology Coimbatore-4.
- Mitchell, T. M. 1997. *Machine Learning*. McGraw-Hill Science /Engineering /Math;.
- Mihalcea, R. , Banea, C. & Wiebe, J. 2007. *Learning Multilingual Subjective Language via Cross-Lingual Projections*. Proceedings of the Association for Computational Linguistics (ACL). pp. 976–983.
- Ni Wayan Sumartini Saraswati. 2011. *Text Mining Dengan Metode Naïve Bayes Classifier Dan Support Vector Machines Untuk Sentiment Analysis*, Universitas Udayana.
- Nugroho, A.S. 2012. *Pemanfaatan Data Mining untuk Mengekstrak Pengetahuan Data Medis*. Workshop on Data Mining, Seminar Nasional Teknologi Informasi (SNATI 2012), Universitas Islam Indonesia, Yogyakarta.
- Noviah Dwi Putranti, Edi Winarko. 2014. Analisis Sentimen *Twitter* untuk Teks Berbahasa Indonesia dengan *Maximum Entropy* dan *Support Vector Machine*, FMIPA UGM.
- Nurirwan Saputra, Teguh Bharata Adji, Adhistya Erna Permanasari. 2015. Analisis Sentimen Data Presiden Jokowi Dengan *Preprocessing Normalisasi Dan Stemming* Menggunakan Metode *Naïve Bayes* Dan *SVM*”, Universitas PGRI Yogyakarta, Universitas Gajah Mada.
- 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.
- Pang, B. & Lee, L. 2005. *Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales*. Proceedings of the Association for Computational Linguistics (ACL). pp. 115–124.
- Pang, B. & Lee, L. 2008. *Subjectivity Detection and Opinion Identification. Opinion Mining and Sentiment Analysis*. Now Publishers Inc.
- Rajaraman, A., Leskovec, J., & Ullman, J. D. 2011. *Mining of Massive Datasets* (hal. 1–17).

- Routray, P., Swain, C. K. & Mishra, S.P., 2013. A Survey on Sentiment Analysis. *International Journal of Computer Applications*, Agustus, 70(10), pp. 1-8.
- Salton, G., & Buckley, C. (1988). *Term-weighting approaches in automatic text retrieval*. *Information Processing & Management*, 24, 513 - 523.
- 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 Conference (HLT-NAACL). pp. 300–307.
- Su, F. & Markert, K. 2008. *From Words to Senses: a Case Study in Subjectivity Recognition*. Proceedings of Coling 2008, Manchester, UK.