Jurnal Dinamika Informatika
Volume 7, No 1, Februari 2018
ISSN 1978-1660 : 45-57
ISSN *online* 2549-8517

ANALISIS SENTIMEN DENGAN *PREPROCESSING* KATA (SENTIMENT ANALISYS WITH LEXICON PREPROCESSING)

Nurirwan Saputra

Program Studi Teknik Informatika, Universitas PGRI Yogyakarta
email : nurirwan@upy.ac.id

ABSTRAK

Penelitian ini berhubungan dengan politik yang mengambil data Presiden RI 2014-2019 yaitu Ir. H. Joko Widodo dari media sosial dan blog politik kemudian akan dilakukan Analisis Sentimen terhadap komentar masyarakat baik yang pro maupun kontra terhadap Ir. H. Joko Widodo.

Penelitian ini lebih ke pendekatan preprocessing kata terlebih dahulu untuk meningkatkan akurasi, yaitu dengan mengubah banyak kata menjadi sebuah kata. Metode yang digunakan adalah Naïve Bayes dan Support Vector Machine (SVM). Penelitian ini melanjutkan penelitian yang sudah dilakukan sebelumnya dengan judul "Analisis Sentimen Data Presiden Jokowi Dengan Preprocessing Normalisasi Dan Stemming Menggunakan Metode Naive Bayes Dan SVM".

Akurasi pada penelitian sebelumnya "Analisis sentimen data presiden Jokowi dengan preprocessing normalisasi dan stemming menggunakan metode naive bayes dan SVM" dengan dilakukan normalisasi dan *stemming* pada data sebesar 89,2655% menggunakan metode SVM. Kemudian pada penelitian ini, dengan melakukan preprocessing menjadi sebuah kata, terjadi peningkatan dengan menggunakan metode SVM dengan akurasi sebesar 91,5254, yaitu peningkatan sebesar 2,2599%.

Kata kunci : analisis sentimen, *svm, smo, naive bayes, preprocessing* kata.

PENDAHULUAN

Politik adalah upaya transfer kekayaan dari satu individu atau kelompok ke individu atau kelompok lain melalui kekuatan pemerintah [1]. Di Indonesia sendiri menganut trias politika [2] yang merupakan pilar dari demokrasi Indonesia yang membagi kekuasaan menjadi tiga, yaitu eksekutif, yudikatif dan legislatif [3], salah satu bentuk demokrasi adalah pemilu, pemilu adalah sarana pelaksanaan kedaulatan rakyat yang diselenggarakan secara langsung, umum, bebas, rahasia, jujur, dan adil dalam Negara Kesatuan Republik Indonesia (NKRI) berdasarkan UUD RI Tahun 1945 [2]. Pemilu dilakukan untuk memilih Presiden dan Wakil Presiden.

Jokowi merupakan tokoh yang cukup fenomenal [4], jenjang karir Jokowi sangat cepat, mulai tahun 2005 beliau menjabat sebagai walikota Solo, kemudian tahun 2012 menjadi Gubernur Ibukota RI yaitu DKI Jakarta, hingga saat ini tahun 2014 beliau sudah menjabat sebagai Presiden Republik Indonesia (RI) [5]. Selain itu, Jokowi adalah presiden pertama yang tidak memiliki kaitan dengan mantan diktator Soeharto, yang berkuasa selama 30 tahun lebih sebelum digulingkan pada 1998 [6]. Tetapi sosok Jokowi tidak lepas dari sentimen negatif masyarakat, di antaranya Jokowi munafik [7], Jokowi penipu [8] dan lain sebagainya.

Penelitian yang sudah pernah dilakukan berkaitan dengan analisis sentimen menggunakan berbagai macam metode, seperti metode Support Vector Machine (SVM) [9], Naïve Bayes [10][11][12], KNN [13], Multinomial Naïve Bayes [14], dan Enhance Naïve Bayes [15]. Selain itu menggunakan berbagai macam bahasa, seperti Bahasa China [16], Bahasa Arab [17], bahasa Indonesia [9][10][13][11][12], tiga macam bahasa yaitu Spanyol, Jerman dan Perancis [18] dan bahasa-bahasa lainnya. Datanya pun beragam ada data yang berasal dari tokoh publik [10], sosial media [9][19][20][21], dan movie review dalam bahasa inggris yang sudah populer [12][22]. Penelitian yang dilakukan sebelumnya menggunakan class attribute yang bervariasi, ada yang menggunakan dua class attribute, yaitu positif dan negatif [10][12], dan ada pula peneliti yang menggunakan tiga class attribute, yaitu positif, netral dan negatif [9]. Penelitian yang akan dilakukan berkaitan dengan analisis sentimen ini menggunakan multiclass yang dilabeli positif, netral dan negatif. Penelitian ini menangani tidak hanya lexicon tetapi juga menangani emoticon yang muncul dalam kalimat.

Normalisasi pada penelitian ini menggunakan kamus KBBA (Kamus Besar Bahasa Alay) yang didapat dari Nurfalah Adiyasa karena kamus yang menjadi penelitiannya di-*share* untuk kepentingan penelitian selanjutnya. *Stemmer* yang dipakai menggunakan Sastrawi Master karena merupakan *Library PHP* untuk *stemming* bahasa Indonesia, mudah diintegrasikan dengan *framework* atau *package* lainnya, mempunyai *API* yang sederhana dan mudah digunakan .

Jurnal Dinamika Informatika
Volume 7, No 1, Februari 2018
ISSN 1978-1660 : 45-57
ISSN *online* 2549-8517

Tujuan dan manfaat dilakukannya penelitian ini adalah melihat seberapa besar preprocessing kata sebelum dilakukan klasifikasi dengan menggunakan metode Naïve Bayes dan SVM

Data Mining

Data sangat banyak jumlahnya dan sangat beragam, termasuk data yang ada dalam politik Indonesia, banyak sekali data yang dapat diolah, di antaranya dalam penelitian ini diambil data mengenai komentar masyarakat tentang tokoh Jokowi baik itu positif, netral maupun negatif. Dengan tersedianya data dalam kualitas dan ukuran yang memadai, teknologi data mining memiliki kemampuan-kemampuan di antaranya sebagai berikut [23].

1. Melakukan Prediksi Tren

   Data mining secara otomatis dapat melakukan proses yang pencarian informasi di dalam basis data.

2. Mengotomatisasi penemuan pola yang tidak diketahui sebelumnya

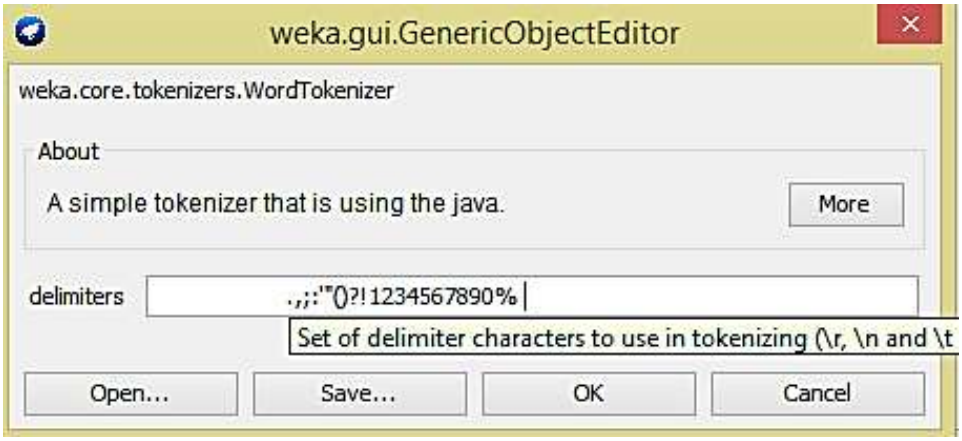
   Data mining dapat menemukan pola-pola dari data yang dimasukkan.

Analisis Sentimen

Analisis sentimen atau *opinion mining* mencakup berbagai macam tugas yang berbeda-beda seperti analisis sentimen, *opinion mining*, *opinion extraction*, *sentiment mining*, *subjectivity analysis*, *affect analysis*, *emotion analysis*, dan *review mining* [10]. Preprocessing yang dilakukan pada penelitian ini adalah sebagai berikut.

1. Cleansing

   Pada proses cleansing ini, sudah difasilitasi oleh WEKA, untuk itu perlu dilakukan *delimiter* atau penghapusan karakter atau tanda baca pada WEKA, maupun penghapusan secara manual khususnya link situs dan link gambar, *delimiter* pada weka ditambahkan karakter dan angka “.,;:'"()?!1234567890%” (Gambar 1).



Gambar 1. Penambahan *Delimiter* Pada WEKA

2. Normalisasi Kata

   Komentar yang diberikan seseorang tidak semuanya menggunakan bahasa baku. Untuk normalisasi ini menggunakan bantuan kamus KBBA yang didapat dari Nurfalah Adiyasa [18], contoh dari kamus normalisasi ini dapat dilihat pada Tabel 1.

Tabel 1. Tabel Normalisasi

No	**Bahasa Tidak Baku**	**Bahasa Baku**
1.	7an	Tujuan
2.	Adlh	Adalah
3.	Gue	Saya
4.	Loe	Kamu
5.	Gawe	Kerja

3. *Stemming*

   Stemming adalah proses mengubah kata menjadi kata dasarnya dengan menghilangkan imbuhan-imbuhan pada kata dalam dokumen atau mengubah kata kerja menjadi kata benda [28]. Contoh : kata "dihilangkan" setelah imbuhan di- dan –kan berubah menjadi "hilang".

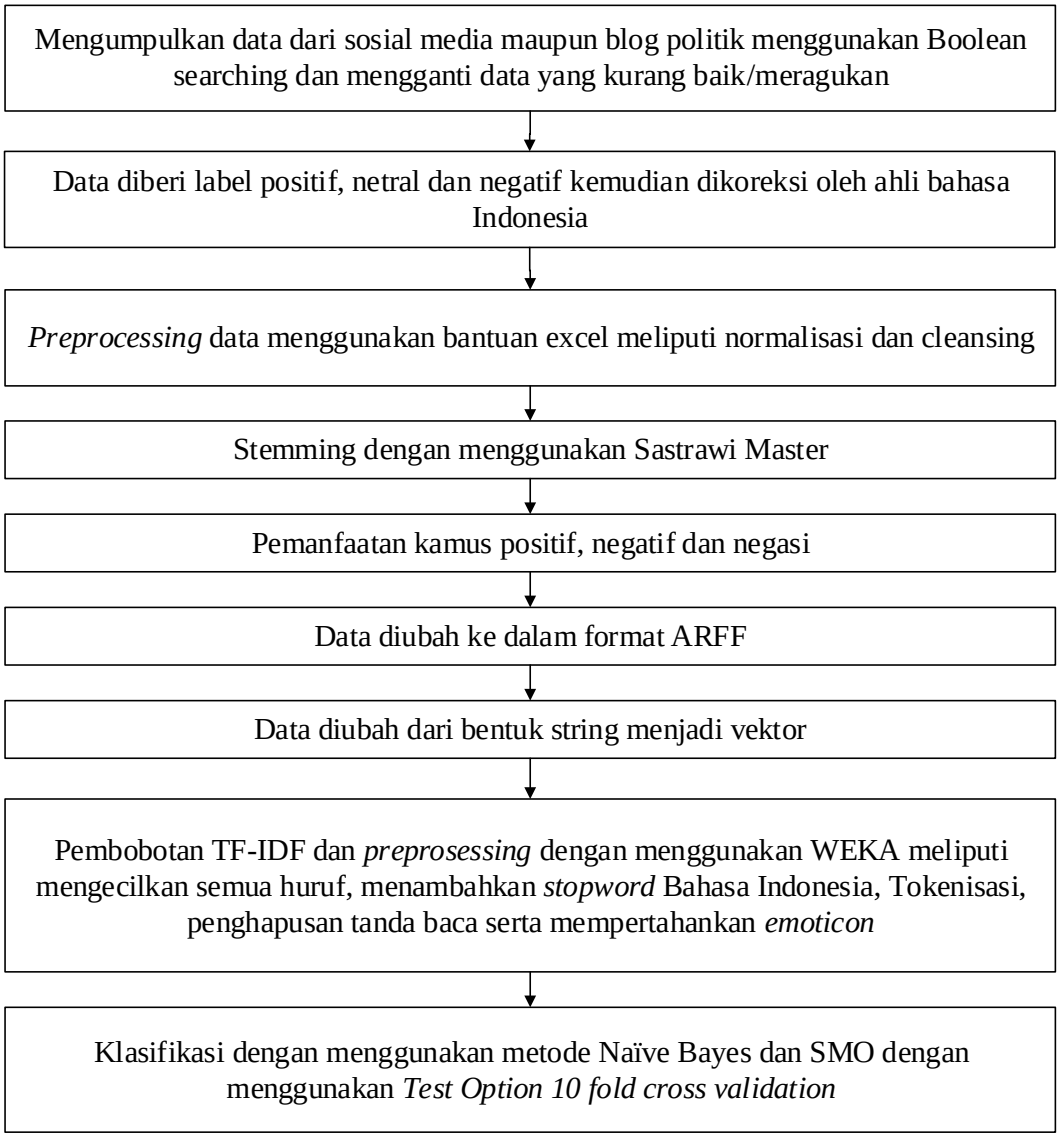
4. Mengubah menjadi sebuah kata

   Pada penelitian ini khususnya, dilakukan pengubahab kata yang sama maknanya menjadi sebuah kata. Pengubahan ini menggunakan kamus Nurfalah Adiyasa, semua kata yang mengandung makna positif akan diubah menjadi kata "positif", semua kata yang mengandung makna negatif akan diubah menjadi kata "negatif", dan semua kata yang mengandung makna netral akan diubah menjadi kata "netral".

Jurnal Dinamika Informatika
Volume 7, No 1, Februari 2018
ISSN 1978-1660 : 45-57
ISSN *online* 2549-8517

METODE PENELITIAN

Metode yang dilakukan pada penelitian ini dapat dilihat pada Gambar 2.



Gambar 2. Metode Penelitian

HASIL

Penelitian ini menggunakan teknik *lexicon* dengan mengubah berbagai macam kata menjadi sebuah kata, teknik ini menggunakan bantuan kamus yang terdiri dari kamus yang berisi kata positif, kamus yang berisi kata negatif, dan kamus yang berisi kata negasi. Kata-kata dalam data apabila terdapat dalam kamus diubah menjadi sebuah kata "positif" untuk seluruh kata positif, "negatif" untuk seluruh kata negatif, dan "negasi" untuk seluruh kata negasi. Selain itu penelitian ini memberikan hasil akurasi yang didapat dengan data yang dilakukan *preprocessing* menjadi sebuah kata yang memiliki kesamaan makna. Hasil penelitian ini disajikan dalam bentuk gambar dan table baik menggunakan metode SVM maupun Naïve Bayes serta berdasarkan tokenisasi yang dilakukan.

Berikut ini merupakan singkatan yang dipakai pada penamaan kolom table.

1. LC = Lowecase.
2. MTF = minTermFreq.
3. N = *normalize.*
4. SW = *Stopwords.*
5. Met = Metode yang digunakan.
6. T = Tokenizer.
7. TPre = Teknik *Preprocessing.*
8. Emo = *Emoticon.*
9. NB = Naive Bayes.
10. TTB = *Time taken to build model* (Waktu yang dibutuhkan untuk membangun model.

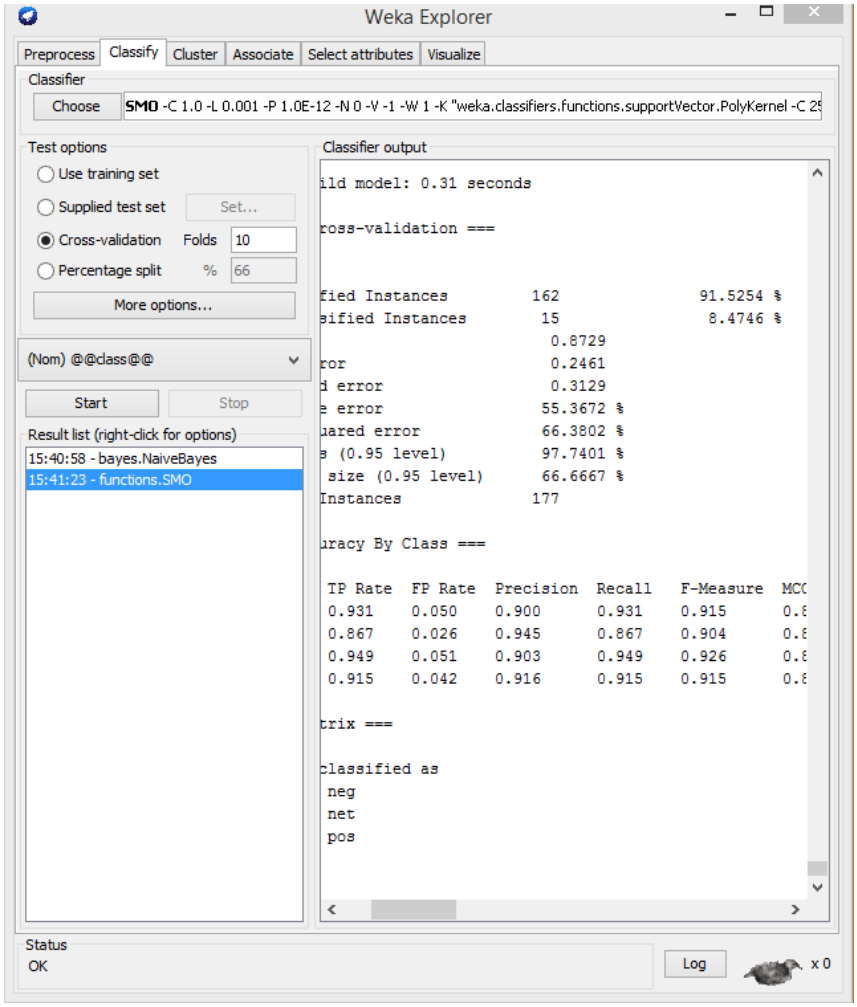
Tabel 2. Data Yang Dinormalisasi, *Stemming* dan Ubah *Lexicon*

TPre	Met	TF-IDF	LC	MTF	N	SW	T	Emo	TTB(s)	Hasil (%)
A	NB	Yes	Yes	1	1	Yes	N-Gram	Yes	0,97	84,7458
	SVM	Yes	Yes	1	1	Yes	N-Gram	Yes	0,42	89,2655
B	NB	Yes	Yes	1	1	Yes	N-Gram	No	1,08	80,791
	SVM	Yes	Yes	1	1	Yes	N-Gram	No	0,49	88,1356
C	NB	Yes	Yes	1	1	No	N-Gram	Yes	0,94	87,5706
	SVM	Yes	Yes	1	1	No	N-Gram	Yes	0,33	90,3955
D	NB	Yes	Yes	1	1	No	Unigram	Yes	0,22	83,6158
	SVM	**Yes**	**Yes**	**1**	**1**	**No**	**Unigram**	**Yes**	**0,31**	**91,5254**

Jurnal Dinamika Informatika
Volume 7, No 1, Februari 2018
ISSN 1978-1660 : 45-57
ISSN *online* 2549-8517

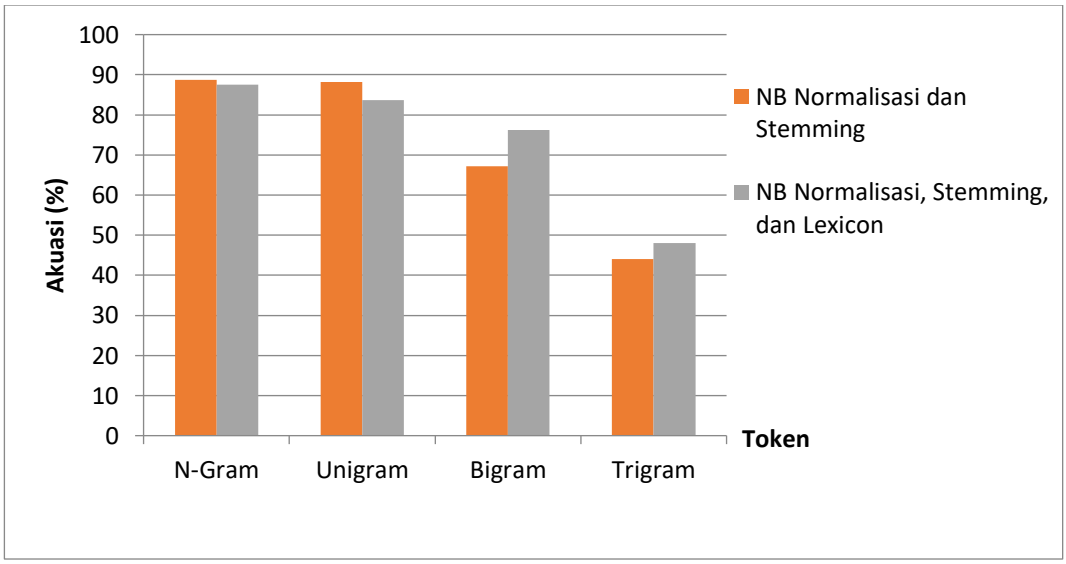
TPre	Met	TF-IDF	LC	MTF	N	SW	T	Emo	TTB(s)	Hasil (%)
E	NB	Yes	Yes	1	1	No	Bigram	Yes	0,44	76,2712
	SVM	Yes	Yes	1	1	No	Bigram	Yes	0,31	71,7514
F	NB	Yes	Yes	1	1	No	Trigram	Yes	0,47	48,0226
	SVM	Yes	Yes	1	1	No	Trigram	Yes	0,31	46,8927

Dilihat dari Tabel 2 terdapat peningkatan akurasi dibandingkan penelitian sebelumnya [29] dengan akurasi tertinggi yaitu 91,5254% (Gambar 3) dengan menggunakan metode SVM *token* Unigram serta menggunakan *emoticon*, peningkatan juga terjadi pada *token* Bigram dengan menggunakan kombinasi *lexicon* dan *emoticon* ini memperoleh akurasi sebesar 76,2712% untuk Naive Bayes dan 71,7514% untuk SVM.



Gambar 3. Akurasi tertinggi dengan menggunakan metode SVM *token* Unigram

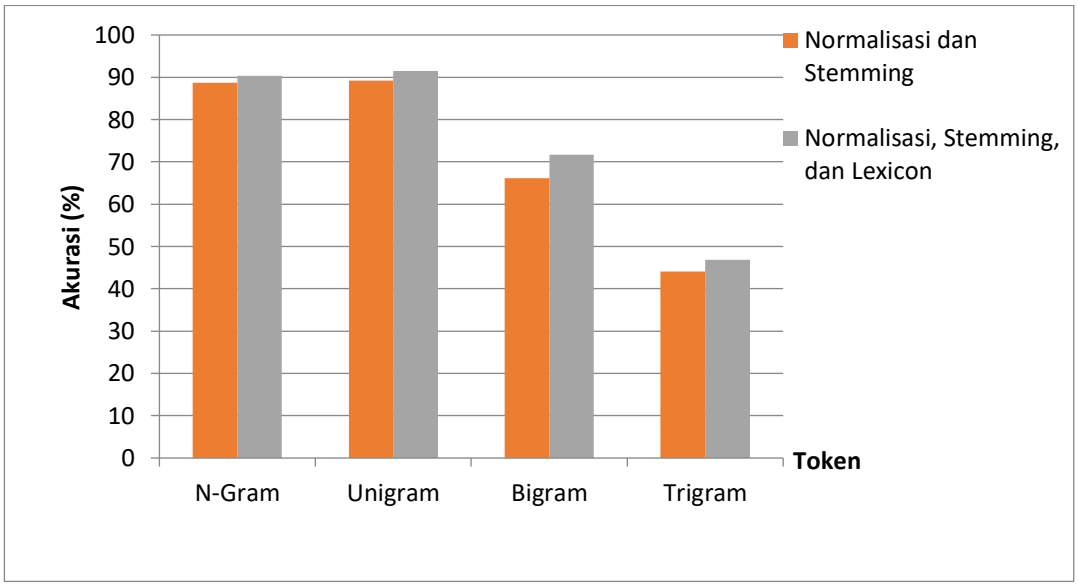
Berikut ini merupakan grafik perbandingan token pada metode naive bayes dibandingkan penelitian sebelumnya [29] Gambar 4. Dari gambar tersebut terlihat perbandingan masing-masing token. Baik token Unigram maupun N-Gram dengan menggunakan metode Naïve Bayes lebih baik dibandingkan Bigram maupun Bigram. Tetapi dibandingkan penelitian sebelumnya[29], ada penurunan akurasi pada token N-Gram dan Unigram dan kenaikan pada token Bigram dan Trigram.



Gambar 4. Grafik Perbandingan Token Pada Metode Naive Bayes dibandingkan penelitian sebelumnya.

Berikut ini grafik perbandingan token pada metode SVM berdasarkan penelitian sebelumnya [29] (Gambar 5). Dari gambar tersebut terlihat perbanding token antara N-Gram, Unigram, Bigram dan Trigram. Sama seperti metode Naïve Bayes, token N-Gram dan Unigram memberikan hasil akurasi yang lebih baik dibandingkan Bigram dan Trigram. Tetapi dibandingkan dengan penelitian sebelumnya [29], metode SVM mengalami kenaikan untuk seluruh token, baik N-Gram, Unigram, Bigram dan Trigram.

Jurnal Dinamika Informatika
Volume 7, No 1, Februari 2018
ISSN 1978-1660 : 45-57
ISSN *online* 2549-8517



Gambar 5. Grafik Perbandingan Token Pada Metode SVM

KESIMPULAN

Berdasarkan penelitian yang dilakukan, dapat ditarik kesimpulan sebagai berikut.

1. Dengan mengubah kata yang memiliki kesamaan makna menjadi sebuah kata, terjadi peningkatan sebesar 1,32% untuk metode Naive Bayes dan peningkatan yang cukup drastis dihasilkan oleh metode SVM yaitu sebesar 4,43%.
2. Akurasi yang dihasilkan metode SVM tidak selalu unggul dibandingkan metode Naive Bayes, begitu pula sebaliknya. Untuk metode yang paling tinggi pada penelitian sebelumnya [29] mendapat akurasi 89,2655% untuk teknik dan pada penelitian ini mendapatkan akurasi dengan menggunakan metode SVM sebesar 91,5254%, yaitu terjadi kenaikan akurasi sebesar 2,599%.

SARAN

Berdasarkan penelitian yang sudah dilakukan, terdapat beberapa poin saran yang dapat dilakukan, diantaranya sebagai berikut.

1. Menggunakan Big data yang jumlahnya banyak
2. Memberikan stopword yang baik untuk Analisis Sentimen, khususnya Bahasa Indonesia.

DAFTAR PUSTAKA

[1] A. Na'im and J. Hartono, "The Effect Of Antitrust Investigations On The Management Of Earnings: A Further Empirical Test Of Political Cost Hypothesis," *Kelola*, vol. 5, no. 1996, 1996.

[2] E. Elisabeth Sinaga, K. Titiek, and Y. Mirza, "GOLONGAN PUTIH (GOLPUT) DALAM PEMILU LEGISLATIF 2009 DI KOTA BENGKULU," ut, Fakultas Ilmu Sosial Dan Ilmu Politik UNIB, 2010.

[3] M. Yahya, "SEJARAH PERKEMBANGAN DEMOKRASI," *KARYA Ilm. Mhs. S1 Sist. Inf.*, vol. 0, no. 0, Nov. 2011.

[4] "Jokowi, Sang Pemimpin Fenomenal," *nasional.inilah.com*. [Online]. Available: http://nasional.inilah.com/read/detail/1815215/jokowi-sang-pemimpin-fenomenal. [Accessed: 24-Nov-2014].

[5] P. J. | R. B. Tips and T. says, "Profil Jokowi I Biodata Lengkap Joko Widodo | Dunia Baca dot Com." .

[6] "Berita Dunia: Jokowi, Presiden Terpilih Pertama yang Tak Terkait Soeharto," *beritasatu.com*. [Online]. Available: http://www.beritasatu.com/nasional/198621-berita-dunia-jokowi-presiden-terpilih-pertama-yang-tak-terkait-soeharto.html. [Accessed: 24-Nov-2014].

[7] "Indonesia Baru yang Munafik (Jokowi-JK)," *KOMPASIANA.com*. [Online]. Available: http://politik.kompasiana.com/2014/09/21/indonesia-baru-yang-munafik-jokowi-jk-680297.html. [Accessed: 24-Nov-2014].

[8] Kandunk, "Jadi Penipu, Jokowi Belajar Sama Siapa Ya?," *Silontong.com: Berita dan Ulasan Menarik*. .

[9] J. K. Wibisono and M. S. Drs. Edi Winarko, "OPINION MINING PADA TWITTER UNTUK BAHASA INDONESIA DENGAN METODE SUPPORT VECTOR MACHINE DAN METODE BERBASIS LEXICON," Universitas Gadjah Mada, 2013.

[10] A. F. Hidayatullah and M. T. Dr. Azhari SN, "ANALISIS SENTIMEN DAN KLASIFIKASI KATEGORI TERHADAP TOKOH PUBLIK PADA DATA TWITTER MENGGUNAKAN NAIVE BAYES CLASSIFIER," Universitas Gadjah Mada, 2014.

[11] "metode naive bayes sentiment analysis - Google Cendekia." [Online]. Available: http://scholar.google.co.id/scholar?q=metode+naive+bayes+sentiment+analysis&btnG=&hl=id&as_sdt=0%2C5. [Accessed: 17-Nov-2014].

[12] M. Merina, "Klasifikasi Dokumen Beropini Me nggunakan Metode Naive Bayes dan Metode Categorical Pr oportional Difference," *Klasifikasi Dok. Beropini Me Nggunakan Metode Naive Bayes Dan Metode Categ. Pr Oportional Differ.*, 2013.

[13] "metode knn sentiment analysis - Google Cendekia." [Online]. Available: http://scholar.google.co.id/scholar?q=metode+knn+sentiment+analysis&btnG=&hl=id&as_sdt=0%2C5. [Accessed: 17-Nov-2014].

[14] "metode multinomial naive bayes sentiment analysis - Google Cendekia." [Online]. Available: http://scholar.google.co.id/scholar?q=metode+multinomial+naive+bayes+sentiment+analysis&btnG=&hl=id&as_sdt=0%2C5. [Accessed: 17-Nov-2014].

[15] V. Narayanan, I. Arora, and A. Bhatia, "Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model," in *Intelligent Data Engineering and Automated Learning – IDEAL 2013*, H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise, B. Li, and X. Yao, Eds. Springer Berlin Heidelberg, 2013, pp. 194–201.

[16] W. Zheng and Q. Ye, "Sentiment Classification of Chinese Traveler Reviews by Support Vector Machine Algorithm," in *Third International Symposium on Intelligent Information Technology Application, 2009. IITA 2009*, 2009, vol. 3, pp. 335–338.

[17] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in *2012 International Conference on Collaboration Technologies and Systems (CTS)*, 2012, pp. 546–550.

[18] A. Balahur and M. Turchi, "Comparative Experiments for Multilingual Sentiment Analysis Using Machine Translation." [Online]. Available: ceur-ws.org/Vol-917/SDAD2012_8_Balahur.pdf. [Accessed: 07-Dec-2014].

[19] M. Choy, M. L. F. Cheong, M. N. Laik, and K. P. Shung, "A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction," *ArXiv11085520 Cs Stat*, Aug. 2011.

[20] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France," *New Media Soc.*, p. 1461444813480466, Apr. 2013.

[21] S. Rill, D. Reinel, J. Scheidt, and R. V. Zicari, "PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis," *Knowl.-Based Syst.*, vol. 69, pp. 24–33, Oct. 2014.

[22] V. K. Singh, R. Piryani, A. Uddin, P. Waila, and Marisha, "Sentiment analysis of textual reviews; Evaluating machine learning, unsupervised and SentiWordNet approaches," in *2013 5th International Conference on Knowledge and Smart Technology (KST)*, 2013, pp. 122–127.

[23] V. Moertini, "Data Mining Sebagai Solusi Bisnis," *Data Mining Sebagai Solusi Bisnis*. [Online]. Available: http://mfile.narotama.ac.id/files/Tubagus%20Purworusmiadi/Kumpulan%20File%20PDF/idatamining_ok.pdf. [Accessed: 02-Dec-2014].

[24] N. M. Huda, "Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa (Studi Kasus di Fakultas MIPA Universitas Diponegoro)," other, FACULTY OF MATHEMATICS AND NATURAL SCIENCES, 2010.

[25] N. S. YUDA, "Data Mining Menggunakan Algoritma Naïve Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro. (Studi Kasus: Fakultas Ilmu Komputer Angkatan 2009).," *SkripsiFakultas Ilmu Komput.*, 2014.

[26] B. Warsito, D. Ispriyanti, and H. Widayanti, "CLUSTERING DATA PENCEMARAN UDARA SEKTOR INDUSTRI DI JAWA TENGAH DENGAN KOHONEN NEURAL NETWORK," *J. PRESIPITASI*, vol. 4, no. 1, pp. 1–6, Mar. 2008.

[27] "Quadratic Programming – MATLAB." [Online]. Available: http://www.mathworks.com/discovery/quadratic-programming.html. [Accessed: 07-Dec-2014].

[28] D. Kerami and H. Murfi, "Kajian Kemampuan Generalisasi Support Vector Machine dalam Pengenalan Jenis Splice Sites Pada Barisan DNA," 03-Dec-2004. [Online]. Available: http://repository.ui.ac.id/dokumen/lihat/246.pdf. [Accessed: 08-Mar-2015].

[29] Saputra, Nurirwan, Teguh Bharata Adji, and Adhistya Erna Permanasari. "Analisis sentimen data presiden Jokowi dengan preprocessing normalisasi dan stemming menggunakan metode naive bayes dan SVM." Jurnal Dinamika Informatika 5, no. 1 (2015).