Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume (2) No.2 September 2018, pp. 200-209
ISSN:2548-9771/EISSN:2549-7200
http://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Analisis Sentimen Pada Media Sosial *Twitter* Menggunakan *Naive Bayes Classifier* Dengan Ekstrasi Fitur *N-Gram*

Agung Nugroho
Sekolah Tinggi Teknologi Pelita Bangsa
Jl. Inspeksi Kalimalang, Tegal Danas, Cikarang Pusat, Kab. Bekasi
agung@pelitabangsa.ac.id

Abstract

Social media is currently an online media that is widely accessed in the world. Microblogging services such as Twitter allow users to write about various things they experience or write reviews of a product, service, public figures and so on. This can be used to take opinion or sentiment towards an entity that is being discussed on social media such as Twitter. This study utilizes these data to determine public opinion or sentiment regarding public perceptions of the issue of rising electricity tariffs. Opinion taking is based on three classes namely positive, negative and neutral. Users often use non-standard word abbreviations or spelling, this can complicate the process and accuracy of classification results. In this study the authors apply text-preprocessing in handling these problems. For feature extraction, n-gram and classification methods are used using the Naive Bayes classifier. From the results of the research that has been done, the most negative sentiments are formed in response to the issue of the increase in basic electricity tariffs. In addition, from the results of testing with the method of cross validation and confusion matrix it is known that the accuracy of the naïve Bayes method reaches 89.67% before applying n-gram, and the accuracy rate increases 2.33% after applying n-gram characters to 92.00%. It is proven that the application of the n-gram extraction feature can increase the accuracy of the naïve Bayes method.

Keywords*: Sentiment Analysis, Classification, Naive Bayes, Twitter, n-gram*

Abstrak

Media sosial saat ini menjadi media online yang banyak di akses didunia. Layanan microbloging seperti twitter memungkinkan pengguna untuk menulis tentang berbagai hal yang mereka alami atau menulis ulasan terhadap suatu produk, jasa layanan, tokoh publik dan lain sebagainya. Hal tersebut dapat dimanfaatkan untuk pengambilan opini atau sentimen terhadap suatu entitas yang sedang dibicarakan di media sosial seperti twitter. Penelitian ini memanfaatkan data tersebut untuk mengetahui opini atau sentimen publik mengenai persepsi masyarakat terhadap isu kenaikan tarif dasar listrik. Pengambilan opini berdasarkan tiga kelas yaitu positif, negatif dan netral. Pengguna sering menggunakan singkatan kata atau ejaan yang tidak baku, hal ini dapat menyulitkan proses dan ketepatan hasil klasifikasi. Pada penelitian ini penulis menerapkan text-preprocessing dalam menangani masalah tersebut. Untuk ekstrasi fitur digunakan n-gram dan metode klasifikasi menggunakan naive bayes classifier. Dari hasil penelitian yang telah dilakukan dapat diketahui bahwa sentimen negatif paling banyak terbentuk dalam menanggapi isu kenaikan tarif dasar listrik. Selain itu dari hasil pengujian dengan metode cross validation dan comfusion matrix diketahui bahwa tingkat akurasi dari metode naïve bayes mencapai 89.67% sebelum diterapkan n-gram, dan tingkat akurasi meningkat 2.33 % setelah diterapkan n-gram karakter menjadi 92.00%. Terbukti bahwa penerapan fitur ekstraksi n-gram dapat meningkatkan nilai akurasi metode naïve bayes.

Kata kunci*: Analisis Sentimen, Klasifikasi, Naive Bayes, Twitter, n-gram*

1. PENDAHULUAN

Pada era sekarang merupakan zaman modern yang menjadikan internet sebagai hal wajar, masyarakat dunia sekarang ini gemar bermain sosial media yang merupakan bagian dari internet. Penggunaan media sosial banyak dimanfaatkan oleh masyarakat umum. Masyarakat banyak menggunakan media sosial untuk mengekspresikan opini, perasaan, pengalaman maupun hal lain yang menjadi perhatian mereka [1]. Salah satu sosial media yang masih banyak digemari adalah twitter. Twitter adalah situs microblogging populer dimana pengguna membuat status yang disebut "*tweet*". *Tweet* memiliki batas maksimal 140 karakter. Orang memposting pesan singkat, menggunakan berbagai bentuk singkatan, menggunakan emoticon dan karakter lain yang mengekspresikannya arti khusus dari kalimat tersebut [2]. *Tweet* atau pesan yang dibagikan di *twitter* biasanya merupakan topik yang sedang hangat dibicarakan dan kadang menjadi *trending topic* di *twitter*.

Seperti akhir–akhir ini yang menjadi trending topic yaitu isu mengenai kenaikan tarif dasar listrik untuk golongan 900VA. Memang semua yang berhubungan dengan listrik termasuk tarif dasar listrik akan berpengaruh langsung terhadap perekonomian masyarakat. Sehingga bermunculan pendapat positif maupun negatif dari masyarakat terhadap isu tersebut. Pendapat-pendapat tersebut bisa diolah dan dianalisa agar menghasilkan data atau informasi yang sehingga informasi yang dihasilkan dapat membantu banyak pihak untuk mendukung suatu keputusan atau pilihan. Salah satu teknik pemrosesan teks yang tepat yaitu analisa sentimen. Analisis sentimen adalah studi komputasi mengenai pendapat, perilaku dan emosi seseorang terhadap entitas. Entitas tersebut dapat menggambarkan individu, kejadian atau topik [3].

Sedangkan untuk mengklasifikasikan kumpulan data *tweet* tersebut bisa menggunakan salah satu metode klasifikasi yaitu *naive bayes classifier*. *naive bayes classifier* merupakan algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasi data uji pada kategori yang paling tepat [4]. Selain itu penulis juga menerapkan *n-gram* pada penelitian ini sebagai ekstraksi fitur. Ekstraksi fitur *n-gram* digunakan untuk pengambilan fitur pada suatu *tweet* sebelum diklasifikasikan menggunakan *naive bayes* dengan harapan bisa meningkatkan akurasi dari algoritma *naïve bayes* dalam mengklasifikasikan data *tweet*.

Dalam penelitian ini mencoba melakukan analisa sentimen untuk melihat persepsi masyarakat terhadap isu kenaikan tarif dasar listrik pada media sosial *twitter* menggunakan metode *naïve bayes*, dengan mengklasifikasikan sentimen menjadi positif, negatif dan netral. Selain itu juga untuk melihat keakurasian metode yang digunakan sebelum dan setelah diterapkan fitur ekstrasi *n-gram* karakter. Dan data dari hasil penelitian ini bisa dimanfaatkan untuk lembaga terkait sebagai bahan pertimbangan kembali dalam mengambil keputusan atau membuat kebijakan yang menyangkut hajat orang banyak.

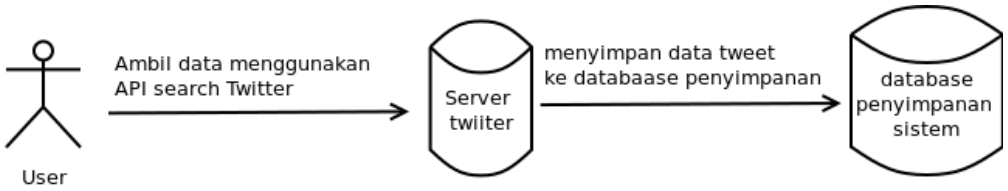
2. METODOLOGI PENELITIAN

2.1. Pengumpulan Data

Pada penelitan ini menggunakan 4 macam data yaitu data *tweet*, data kata *stopword*, data kata dasar dan data pengetahuan.

1. Data *Twitter*

Data *tweet* yang digunakan pada penelitian ini yaitu tweet pada twitter terhadap isu kenaikan tarif dasar listrik yang berupa pendapat positif, negatif dan netral dari masyarakat di twitter. Data tweet didapat dengan memanfaatkan API search twitter dengan mengetikan keyword yang berhubungan dengan tarif dasar listrik.



Gambar 1. Skema Pengumpulan data dari *Twitter*

Data yang sudah terkumpul tersebut nantinya akan dibagi 2 bagian yaitu untuk data training dan data testing untuk menguji keakuratan sistem dalam mengklasifikasikan teks.

2. Data *Stopword*

Data awal *stopword* diperoleh https://github.com/nolimitid/nolimit-kamus. Dimana datanya berjumlah 758 kata dan di simpan di dalam database. Daftar kata *stopword* ini digunakan ketika proses *stopword removal* dalam proses *text mining*. Berikut contoh kata *stopword*.

Tabel 1. Contoh data *stopword*

id_stopword	stopword
1	ada
2	adapun
3	agar

3. Data Kata Dasar

Data kata dasar diperoleh dari https://github.com/nolimitid/nolimit - kamus, dimana data kata dasar berjumlah 28526 kata kemudian data kata dasar tersebut disimpan pada database. Kata dasar ini digunakan dalam proses stemming. Berikut contoh kata dasar.

Tabel 2. Contoh data kata dasar

id_kata_dasar	kata_dasar
1	abu
2	absen
3	acak

4. Data Pengetahuan

Data pengetahuan adalah data hasil dari proses training yang telah dilakukan. Data berbentuk n-gram karakter kata. Data *n-gram* ini dijadikan acuan atau model dalam proses testing untuk menentukan klasifikasi pada *tweet*.

Tabel 3. Contoh data kata dasar

n-gram	sentimen	frekuensi	probabilitas
_l	Positif	8	0.0346154
li	Positif	6	0.0269231
is	Positif	5	0.0230769
st	Positif	5	0.0230769
tr	Positif	5	0.0230769
ri	Positif	5	0.0230769
ik	Positif	5	0.0230769
k_	Positif	5	0.0230769

2.2. Pengolahan Data

Sebelum diproses, perlu dilakukan pengolahan data yang diperoleh dari twitter. Proses pengolahan data atau *text prepocessing* berfungsi untuk merubah data teks yang tidak terstruktur menjadi data yang terstruktur. Proses yang dilakukan adalah sebagai berukut:

1. Case folding

Proses *case folding* untuk menyeragamkan bentuk huruf menjadi huruf kecil. Hal ini dilakukan untuk mempermudah pencarian. Tidak semua dokumen teks konsisten dalam penggunaan huruf kapital.

2. Tokenizing

Pada proses *tokenization* ini, semua kata yang ada di dalam tiap dokumen dipisahkan dan dihilangkan tanda bacanya, serta dihilangkan jika terdapat simbol atau apapun yang bukan huruf.

3. Stopword removal

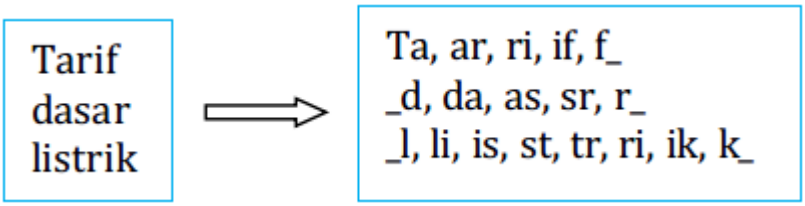
Pada tahap ini, kata-kata yang tidak relevan akan dihapus, kata-kata yang tidak mempunyai makna tersendiri jika dipisahkan dengan kata yang lain dan tidak terkait dengan kata sifat yang berhubungan dengan sentimen.

4. Stemming

Stemming adalah proses pencarian kata dasar dengan menghilangkan imbuhan. Dalam proses ini kata-kata akan dikelompokkan ke dalam beberapa kelompok yang memiliki kata dasar yang sama, seperti lantik, melantik, dan pelantikan di mana kata dasar dari semuanya adalah kata lantik. Pada penelitian ini *stemming* yang digunakan yaitu dari *library sastrawi stemmer* yang dibangun berdasarkan algoritma Nazief & Andriani [5].

2.3. Ekstraksi Fitur n-gram

Setelah melalui tahap *preprocessing* akan dilakukan proses *n-gram* pada dokumen teks. Pada tahap ini sistem akan mengambil sejumlah n karakter sebagai suatu term dan menghitung berapa banyak kata itu muncul dan probabilitas dari *n-gram* karakter tersebut. Dalam penelitian ini menggunakan bigram. Berikut contoh tweet yang mengalami proses *n-gram*:



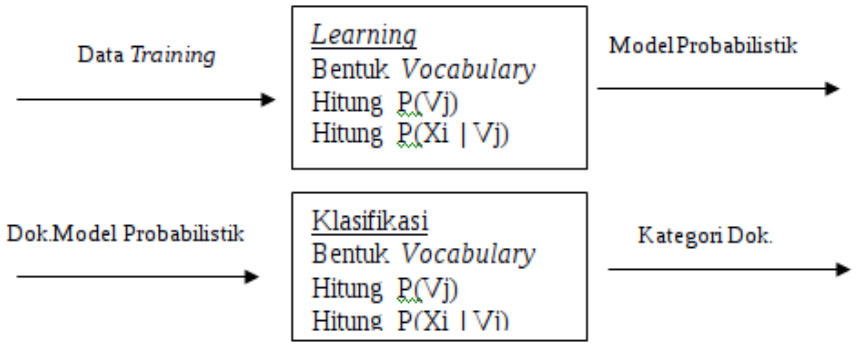
Gambar 2. Contoh proses *n-gram*

n-gram merupakan potongan dari sejumlah n karakter dari sebuah string [6]. Metode *n-gram* digunakan untuk mengambil potongan-potongan karakter dari kata atau kalimat sebanyak jumlah karakter pada kata tersebut. Salah satu keunggulan *n-gram* adalah bahwa n-gram tidak terlalu sensitif terhadap kesalahan dalam penulisan kata [6]. *N-gram* memiliki karakteristik sebagai berikut [7] diantaranya:

1. Dapat berfungsi dengan baik walaupun terdapat kesalahan tekstual
2. Dapat berjalan secara efisien, membutuhkan penyimpanan yang sederhana
3. dan waktu proses yang cepat.

2.4. Algoritma Naïve Bayes Classifier

Algoritma *Naive Bayes Classifier* merupakan pengklasifikasian dengan metode probabilitas, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes [8].



Gambar 3. Algoritma Naïve Bayes Classifier

Pendekatan *naïve bayes* membuat asumsi sederhana bahwa semua atribut bersifat independen. Hal ini menyebabkan penggolongan yang jauh lebih sederhana, ini membuat efektif dalam praktiknya.

Pada tahap ini data diklasifikasikan dengan menggunakan metode *naive bayes classification* yang dikombinasikan dengan ekstraksi fitur *n-gram* untuk mendapatkan hasil analisis sentimen. Metode evaluasi dilakukan dengan menggunakan *confusion matrix*.

Comfusion Matrix adalah tool yang berguna untuk menganalisis seberapa baik *classifier* mengenali kelas yang berbeda. TP dan TN menjelaskan ketika pengklasifikasi mendapatkan sesuatu dengan benar, sementara FP dan FN

menjelaskan ketika pengklasifikasi mendapatkan hal yang salah [9]. Dibawah ini adalah rumus *confusion matrix* untuk menghitung nilai tingkat akurasi.

$$Acuracy = \frac{TP+FN}{TP+FP+TN+FN} \quad (1)$$

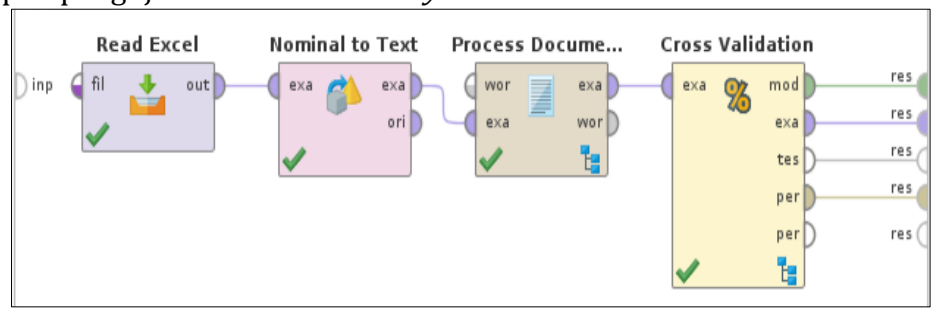
TP = True Positif
TN = True Negatif
FP = False Negatif
FN = False Positif

3. HASIL DAN PEMBAHASAN

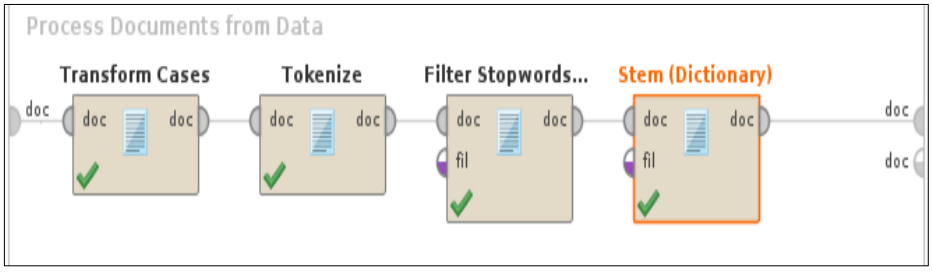
Hasil dari peneltian ini dapat dibagi menjadi tiga bagian. Yaitu hasil pengujian akurasi metode *naïve bayes classifier*, hasil pengujian akurasi metode *naïve bayes* berbasis *n-gram*, hasil analisis sentimen terhadap isu kenaikan tarif dasar listrik.

3.1. Pengujian Akurasi Metode *Naïve Bayes Classifier*

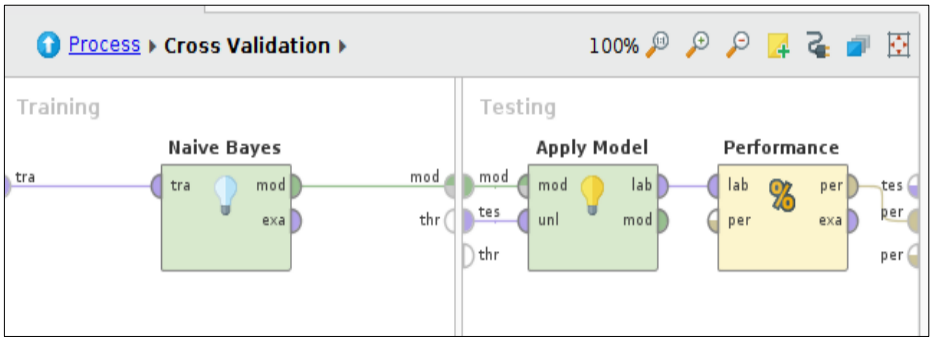
Pengujian akurasi ini dilakukan dengan menggunakan tool bantuan yaitu *rapidminer*. Pengujian dilakukan dengan 300 data *tweet*. Berikut design dari tahapan pengujian metode *naïve bayes*.



Gambar 4. Design Pengujian Metode Naïve Bayes



Gambar 5. Tahapan Process Document From Data



Gambar 6. Validasi Pengujian Naïve Bayes

accuracy: 89.67% +/- 2.33% (mikro: 89.67%)

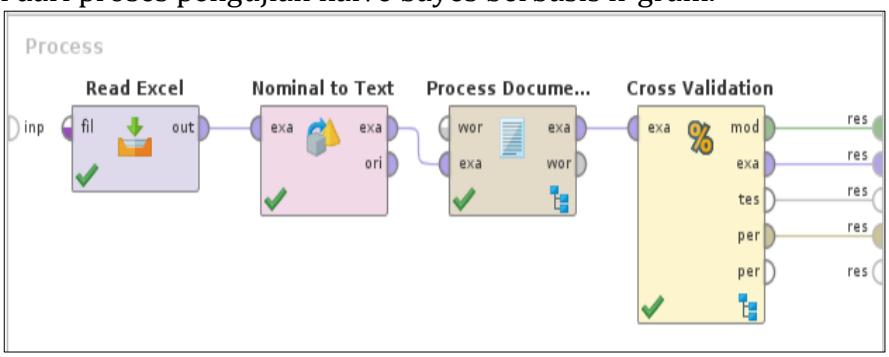
	true netral	true positif	true negatif	class precision
pred. netral	93	0	19	83.04%
pred. positif	1	100	5	94.34%
pred. negatif	6	0	76	92.68%
class recall	93.00%	100.00%	76.00%	

Gambar 7. Comfusion Matrix Pengujian Naïve Bayes

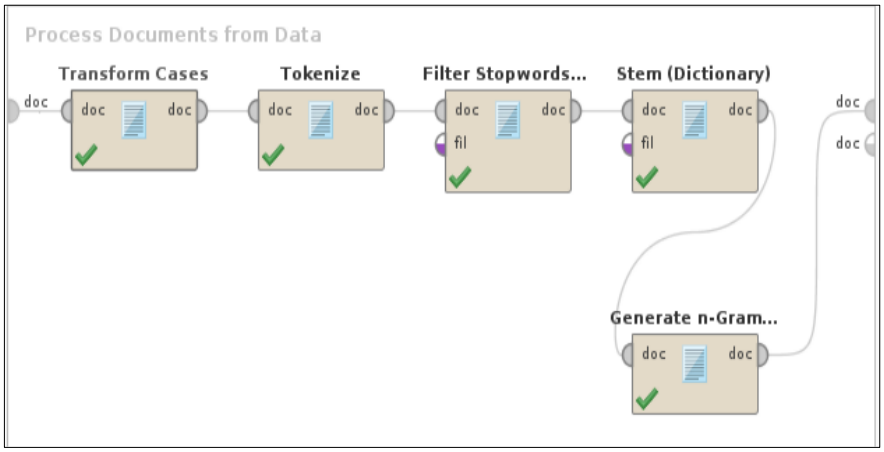
Dari gambar diatas diketahui hasil akurasi dari pengujian metode *naïve bayes* sebesar 89.67%. Nilai ini termasuk *Good classification*.

3.2. Pengujian Akurasi Metode *Naïve Bayes Classifier* Berbasis n-*Gram*

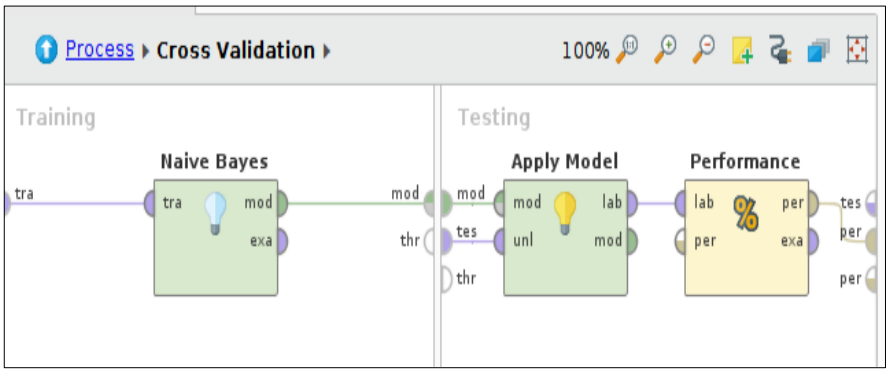
Design dari proses pengujian naïve bayes berbasis n-gram.



Gambar 8. Design Pengujian Naïve Bayes Berbasis N-gram



Gambar 9. Tahapan Proses Text Prepocessing dengan N-gram



Gambar 10. Validasi Penegujian Metode *Naïve Bayes* berbasis *n-gram*

accuracy: 92.00% +/- 3.40% (mikro: 92.00%)

	true netral	true positif	true negatif	class precision
pred. netral	89	0	13	87.25%
pred. positif	2	100	0	98.04%
pred. negatif	9	0	87	90.62%
class recall	89.00%	100.00%	87.00%	

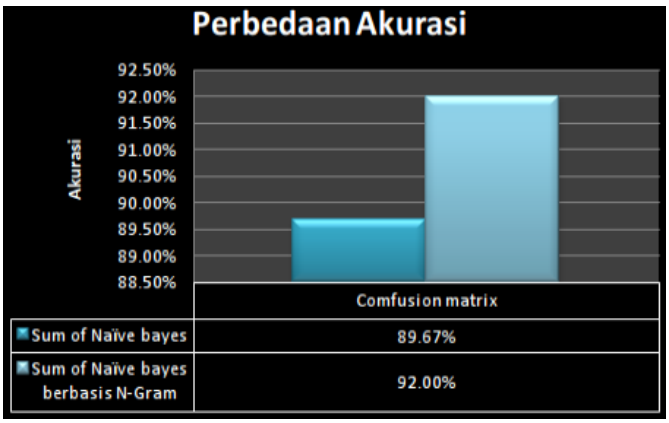
Gambar 11. *Comfusion Matrix* Pengujian *Naïve Bayes* Berbasis *n-gram*

Dari hasil pengujian diketahu bahwa penerapan *n-gram* dapat meningkatkat akurasi dari metode *naïve bayes.* Hasil akurasi pengujian menunjukan angka 92.00% setelah diterapkan *n-gram*. Setelah melakukan pemodelan dan perhitungan berdasar kedua algoritma diatas, kemudian dilakukan perbandingan hasil yang berupa nilai akurasi. Maka diperoleh data perbandingan sebagai berikut:

Tabel 4. Perbandingan hasil pengujian

Perbandingan	Naïve bayes	Naïve bayes berbasis N-Gram
Acuracy	89.67%	92.00%

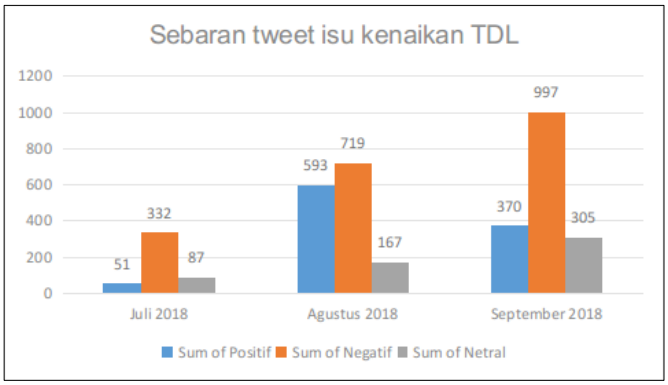
Tabel diatas merupakan hasil akhir dari pengujian yang menujukan perbandingan tingkat akurasi dari metode *naïve bayes* tanpa ektraksi fitur *n-gram* dan metode *naïve bayes* dengan ditambah ektraksi fitur *n-gram*. Berdasarkan table diatas diketahui bahwa penambahan ektraksi fitur *n-gram* memberikan peningkatan nilai akurasi terhadap metode naïve bayes dalam mengklasifiksikan teks sekitar 2.33%.



Gambar 12. Grafik Perbedaan Tingkat Akurasi

3.3. Sebaran *Tweet* Terhadap Isu Kenaikan Tarif Dasar Listrik

Dari data yang dikumpulkan dari periode Juli – September 2018 sekitar 4000 data tweet dari twitter yang sudah terklasifiksikan divisualisasikan dalam bentuk grafik seperti berikut:



Gambar 13. Grafik Sebaran *Tweet* Periode Juli - September 2018

4. SIMPULAN

Berdasarkan dari hasil penelitian terhadap model algoritma *naïve bayes* yang di kombinasikan dengan ekstraksi fitur *n-gram* terhadap analisa perseberan *tweet* terhadap isu kenaikan tarif dasar listrik, maka dapat ditarik kesimpulan

bahwa algoritma *naïve bayes* mampu mencapai tingkat akurasi 89,67%, dan pengaruh ekstraksi fitur *n-gram* yang diterapkan dapat meningkatkan nilai akurasi pada algoritma *naïve bayes* sekitar 2,33%, yaitu menjadi 92,00% dalam pengklasifikasian data *tweet*.

DAFTAR PUSTAKA

[1] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno and J. Caro, "Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning," IISA 2013, Piraeus, 2013, pp. 1-6.

[2] Kotwal, Aishwarya et al, 2016, Improvement in Sentiment Analysis of Twitter Data using Hadoop, International Conference on "Computing for Sustainable Global Development", 16th – 18th March, 2016, BVICAM, New Delhi (INDIA).

[3] Medhat, Walaa, Hassan, Ahmed, & Korashy, Hoda, 2014, Sentiment Analysis Algorithms And Applications: A Survey, Ain Shams Engineering Journal (2014) 5, 1093–1113.

[4] Feldman, R and Sanger, J. 2007. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press:NewYork.

[5] Afuan, L. (2013). Stemming Dokumen Teks Bahasa Indonesia. Jurnal Telematika, Vol. 6, No. 2, Hal. 34–40

[6] Nurfalah, A., & Adiwijaya, A. A. S. (2017). Analisis Sentimen Berbahasa Indonesia dengan Pendekatan Lexicon-Based pada Media Sosial. Jurnal Masyarakat Informatika Indonesia, 2(1), 1-8.

[7] Sadida, Rizqon dkk, 2017, Perancangan Sistem Analisis Sentimen Masyarakat Pada Sosial Media Dan Portal Berita, Seminar Nasional Teknologi Informasi dan Multimedia 2017 STMIK AMIKOM Yogyakarta, 4 Februari 2017.

[8] Kusrini dan Luthfi, Emha Taufiq, 2010, Algoritma Data Mining, Penerbit Andi: Yogyakarta

[9] Han, Jiawei, Kamber, Micheline and Pei, Jian, 2012, Data Mining Concepts and Techniques Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevier. 225 Wyman Street, Waltham, MA 02451, USA, ISBN 978-0-12-381479-1.