Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 1, No. 10, Oktober 2017, hlm. 1193-1200 http://j-ptiik.ub.ac.id

Klasifikasi Berita Twitter Menggunakan Metode Improved Naïve Bayes

Budi Kurniawan[1], Mochammad Ali Fauzi[2], Agus Wahyu Widodo[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]kbudi.kurniawan@gmail.com, [2]moch.ali.fauzi@ub.ac.id, [3] a_wahyu_w@ub.ac.id

Abstrak

Twiter merupakan salah satu sosial media yang banyak digunakan saat ini. Selain digunakan sebagai sosial media Twitter juga digunakan untuk membaca berita. Setiap tahunnya pengguna Twitter mengalami peningkatan sehingga informasi yang ada juga semakin meningkat. Informasi yang semakin meningkat menyebabkan pengguna yang ingin mencari suatu informasi tertentu mengalami kesulitan. Untuk mengatasi masalah tersebut diperlukan pengkategorian. Pada penelitian ini menggunakan *Improved Naïve Bayes* untuk mengkategorikan *tweets* berdasarkan isi berita. Pada *Improved Naïve Bayes* akan dihitung nilai *posterior* setelah kata yang ada dilakukan pembobotan dengan menggunakan *bernoulli* atau angka 1 dan 0. Pada penelitian ini menggunakan delapan kategori berita berbahasa Indonesia yaitu: ekonomi, *entertainment*, olahraga, teknologi, kesehatan, makanan, otomotif, dan travel. Berdasarkan hasil pengujian yang telah dilakukan didapatkan hasil nilai *precision* 0.962961, *recall* 0.789164 dan *f-measure* sebesar 0.862973.

Kata Kunci: *Improved Naïve Bayes, Twitter, Bernoulli,* sosial media

Abstract

Twitter is one of the most widely used social media today. Besides being used as a social media, Twitter is also used to read news. Every year Twitter users have increased, so that information is also increasing. Increased information causes users who want to look for a certain information to experience difficulties. To solve the problem, news categorization is required. This study use Improved Naïve Bayes method to categorize tweets by news contents. In Improved Naïve Bayes posterior value will be calculated after the word is done by weighting using Bernoulli representation or by 1 and 0. This study use eight categories of news in Indonesia, which are: economy, entertainment, sports, technology, health, food, automotive, and travel. Based on the results of tests that have been done this study obtain precision value of 0.962961, recall 0.789164 and f-measure of 0.862973.

Keywords: *Improved Naïve Bayes, Twitter, Bernoulli, social media*

1. PENDAHULUAN

Perkembangan internet memberikan pengaruh pada kehidupan manusia, salah satunya tentang penyebaran informasi yang semakin pesat. Kini hampir semua informasi yang diinginkan ada di internet. Informasi di internet bisa didapatkan dari *website*, *blog*, *microblog*, sosial media, dan lain-lain.

Salah satu media bertukar informasi adalah Twitter. Twitter adalah jejaring sosial yang memungkinkan setiap pengguna berbagi informasi dalam bentuk pesan teks sejumlah 140 karakter (Phuvipadawat and Murata, 2010).

Selain digunakan sebagai media sosial, Twitter juga digunakan sebagai media membaca suatu berita. Pengguna yang ingin mendapatkan informasi dari sebuah akun harus menjadi *follower* akun tersebut terlebih dahulu (Kwak, Lee, Park and Moon, 2010).

Banyaknya informasi yang ada pada Twitter dapat dikurangi dengan melakukan pengkategorian informasi berdasarkan kategori-kategori tertentu (Weissbock, Esmin and Inkpen, 2013).

Metode-metode pengkategorian teks yang digunakan saat ini banyak macamnya antara lain algoritma klasifikasi *Bayes*, *K-Nearest Neighbor* (KNN), *Neural Network* (NN), *Support Vector Machine* (SVM), *The Decision Tree*, *Linear Least Squares Fit* (LLSF) dan lain-lain (Yuan, 2010).

Salah satu metode klasifikasi adalah *Naïve*

Bayes, *Naïve Bayes* merupakan metode klasifikasi sederhana yang menerapkan teorema *bayes* dengan menganggap semua fitur saling tidak berhubungan. Penggunaan algoritma ini menggunakan keseluruhan probabilitas, yaitu probabilitas dokumen terhadap kategori (*prior*). Kemudian teks akan terkategori berdasarkan probabilitas maksimumnya (*posterior*). Dengan kata lain metode ini mengasumsikan bahwa ada atau tidaknya fitur tertentu dari kelas tidak berhubungan dengan ada tidaknya fitur yang lain (Yuan, 2010).

Pada penelitian sebelumnya yang dilakukan oleh Perdana (2013), menggunakan pengkategorian pesan singkat berbahasa Indonesia pada jejaring sosial Twitter dengan metode klasifikasi *Naïve Bayes*. Penelitian tersebut menggunakan enam kategori yaitu olahraga, teknologi, hiburan, keuangan, berita,dan otomotif. Penelitian tersebut diperoleh hasil *recall* 79%, *precision* 80%, dan *f-measure* 78%.

Pada penelitian sebelumnya tentang "Klasifikasi Artikel Berita Secara Otomatis Menggunakan Metode *Naive Bayes* yang Dimodifikasi" menghasilkan peningkatan akurasi rata-rata sebesar 2,3%. Akurasi akan meningkat dengan meningkatnya data latih yang digunakan sebagai pembelajaran. Modifikasi dilakukan dengan menggunakan pembobotan berdasar posisi kata dalam berita (Widodo, 2013).

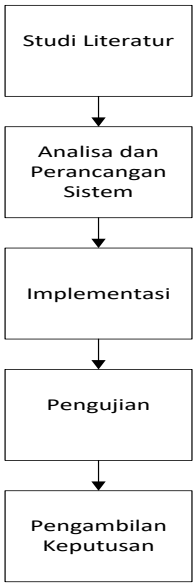
Menurut Yuan (2010), pada penelitian "*An Improved Naïve Bayes Text Classifikation Algorithm In Chinese Information Processing*". Pada saat ini pengklasifikasian teks dengan menggunakan algoritma *Naïve Bayes* banyak mengalami perbaikan seperti pengurangan dimensi dari fitur kata untuk meningkatkan algoritma itu sendiri. Selain itu untuk data latih yang kecil, fitur kata yang langka muncul secara acak. Penelitian tersebut menggunakan Sembilan kategori yaitu *economics*, IT, *health*, *sport*, *travel*, *education*, *recruitment, culture*,dan *military*. Penelitian tersebut diperoleh hasil *precision* dari 80% menjadi 85%, *recall ratio* dari 81% menjadi 83% dan nilai *f-measure* pada pengklasifikasian dari 81% menjadi 84%.

Berdasarkan pada penelitian di atas alasan penulis menggunakan metode *Improved Naïve Bayes* adalah pada penelitian yang dilakukan oleh Yuan (2010), hasil dari penerapan *Improved Naïve Bayes* dalam teks berbahasa China menghasilkan nilai *precision*, *recall*, dan *f-measure* lebih tinggi jika dibandingkan dengan menggunakan metode *Naïve Bayes*. Maka dari itu penulis mencoba mengambil kelebihan metode *Improved Naïve Bayes* untuk klasifikasi berita Twitter berbahasa Indonesia. Kelebihan dari metode *Improved Naïve Bayes* ini adalah pada pengurangan fitur kata, sehingga dapat meningkatkan nilai estimasi pada peluang dokumen terhadap kelas atau kategorinya.

Pada penelitian ini akan menggunakan metode *Improved Naïve Bayes* dan delapan kategori berita berbahasa Indonesia, yaitu: ekonomi, *entertainment*, olahraga, teknologi, kesehatan, makanan, otomotif, dan travel. Penelitian ini mengevaluasi *precision* , *recall* dan *f-measure*. Tujuan dari penelitian ini untuk mengkategorikan berita sesuai dengan kategorinya, sehingga informasi lebih mudah untuk dicari.

2. METODE PENELITIAN

Bab ini berisi langkah-langkah yang dilakukan dalam penelitian yang berjudul "Klasifikasi Berita Twitter Menggunakan Metode *Improved Naïve Bayes*". Tahapan-tahapan dalam penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Diagram Alir Penelitian

2.1. Data Penelitian

Penelitian ini menggunakan data yang didapatkan dari www.detik.com yang merupakan salah satu dari situs berita *online.* Data yang digunakan dalam penelitian "Klasifikasi Berita Twitter Menggunakan Metode *Improved Naïve Bayes* " telah dibedakan

menjadi delapan kategori yaitu: ekonomi, *entertainment*, olahraga, teknologi, kesehatan, makanan, otomotif, travel. Data tersebut berupa judul dari berita. Setiap kategori berita akan diambil sebanyak 50, 100, 150, 200, dan 250 data sebagai data *training* (latih). Sedangkan data *testing* yang digunakan sebanyak 300 data *testing* untuk setiap kategori berita.

2.2. Twitter

Twitter adalah jejaring sosial yang memungkinkan setiap pengguna berbagi informasi dalam bentuk pesan teks sejumlah 140 karakter (Phuvipadawat and Murata, 2010).

2.3. Teks Mining

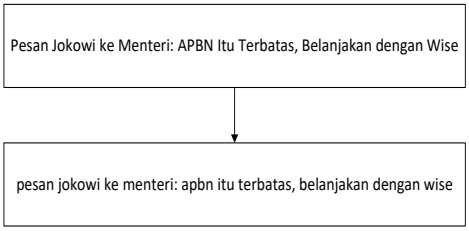
Teks mining merupakan suatu penambangan yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru, sesuatu tidak diketahui sebelumnya atau menemukan suatu informasi yang tersirat secara implisit yang berasal dari informasi yang diekstrak secara otomatis dari sumber-sumber data teks yang berbeda-beda (Feldman, 2006). Dalam hal ini teks mining digunakan untuk analisis informasi pengambilan keputusan dan tugas manajemen informasi lainnya yang berhubungan dengan data teks.

2.4. Preprocessing

Text preprocessing adalah tahap awal dari teks mining untuk merubah data yang tidak terstruktur menjadi data terstruktur (Perdana, 2013). Proses yang dilakukan pada tahap ini seperti *casefolding*, *tokenizing*, *filtering*, dan *stemming*.

1. Case Folding

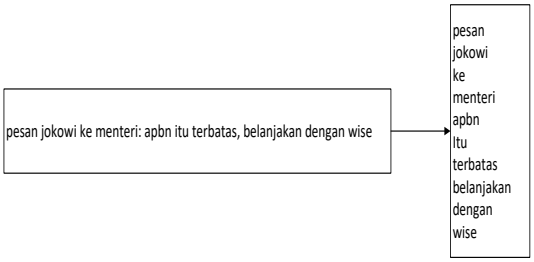
Case folding merupakan tahapan untuk mengubah semua huruf dalam dokumen menjadi huruf kecil (*lowercase*). Huruf yang dilakukan perubahan mulai ‘a’ sampai dengan ‘z’. Gambar 2. merupakan tahap *case folding.*



Gambar 2. *Case Folding*

2. *Tokenizing*

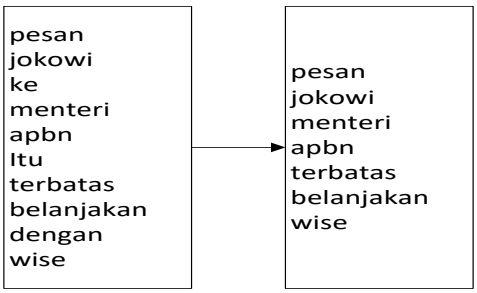
Tokenizing merupakan proses penguraian deskripsi yang semula berupa kalimat-kalimat menjadi kata-kata dan menghilangkan delimiter seperti tanda titik (.), koma (,) dan spasi serta karakter yang ada pada kata tersebut. Gambar 3 merupakan tahap *tokenizing*.



Gambar 3. *Tokenizing*

3. Filtering

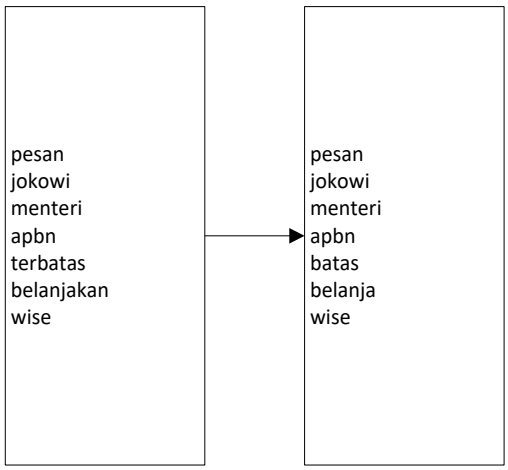
Filtering merupakan tahap mengambil kata-kata yang memiliki arti dari hasil *tokenizing* dengan menggunakan algoritma *stopword removal*. Gambar 4 merupakan tahap *filtering*.



Gambar 4. *Filtering*

4. *Stemming*

Stemming merupakan proses normalisasi dalam sistem *information retrival* yang digunakan untuk mencari kata dasar yang ada dalam suatu dokumen dengan mengacu aturan-aturan tertentu (Agusta, Kristen and Wacana, 2009). Gambar 5 merupakan tahap *stemming.*



Gambar 5. *Stemming*

2.5. Lucene

Lucene adalah sebuah *library* pencarian berbasis Java yang dengan mudahnya dalam menambahkan fitur pencarian pada suatu aplikasi. Saat ini sebagian besar Lucene digunakan sebagai *library Information Retrival* (IR) yang menjadi kekuatan fitur pencarian dibalik *website* dan aplikasi *desktop* (Phuvipadawat and Murata, 2010).

2.6. Naïve Bayes

Naïve Bayes merupakan klasifikasi sederhana yang menerapkan teorema *bayes* dengan menganggap semua fitur saling tidak berhubungan. Pengguna algoritma *bayes* ini menggunakan keseluruhan probabilitas, yaitu probabilitas dokumen terhadap kategori (*prior*). Kemudian teks akan terkategori berdasarkan probabilitas maksimum (*posterior*). Dengan kata lain metode ini mengasumsikan bahwa ada atau tidaknya fitur tertentu dari kelas tidak berhubungan dengan ada atau tidaknya fitur yang lain (Yuan, 2010).

$$p(c_j \mid w_i) = \frac{p(w_i \mid c_j)\, p(c_j)}{p(w_i)} \quad (1)$$

$p(c_j \mid w_i)$ Peluang kategori j ketika terdapat kemunculan kata i

$p(w_i \mid c_j)$ Peluang sebuah kata i masuk ke dalam kategori j

$p(c_j)$ Peluang kemunculan sebuah kategori j

$p(w_i)$ Peluang kemunculan sebuah kata

Ada banyak cara untuk menghitung $p(w_i \mid C_j)$, cara paling sederhananya adalah dengan menggunakan Persamaan 2.

$$p(w_i \mid C_j) = \frac{N_{ic} + 1}{N_c + M + V} \quad (2)$$

N_{ic} Jumlah dokumen latih dengan fitur atribut w_i dengan kategori C_j

N_c Jumlah dokumen dari kategori C_j

V Jumlah kategori

M Untuk menghindari masalah yang disebabkan terlalu kecilnya nilai N_{ic}

Pada saat ini pengklasifikasian teks mengalami banyak perbaikan. Pada algoritma Klasifikasi *Naïve Bayes* dilakukan pengurangan dimensi dari fitur kata-kata untuk meningkatkan algoritma klasifikasi itu sendiri, perbaikan ini sudah meningkat tetapi penerapan dari *Naïve Bayes* klasifikasi hanya bisa dilakukan untuk kategori tertentu. Selain itu untuk data latih yang kecil, fitur kata-kata langka yang muncul secara acak. Sehingga probabilitasnya dihitung dengan menggunakan Persamaan 3.

$$p(w_i \mid C_j) = \frac{1 + 1}{N_c + M + V} \quad (3)$$

Sedangkan probabilitas untuk kategori yang lain dihitung dengan menggunakan Persamaan 4.

$$p(w_i \mid C_j) = \frac{0 + 1}{N_c + M + V} \quad (4)$$

Keduanya memiliki nilai kurang lebih sama dengan 0, yang akan mempengaruhi hasil dari Persamaan 5.

$$\max_{C_j \in C} p(C_j \mid d_i) = \max_{C_j \in C} p(C_j) \prod_{i=1}^{m} p(w_i \mid C_j) \quad (5)$$

Untuk menghitung nilai peluang kemunculan suatu kategori atau Prior pada suatu dokumen dapat dilakukan dengan menggunakan Persamaan 6.

$$p(C_j) = \frac{N_c}{V} \quad (6)$$

Beberapa bentuk representasi dari metode naïve bayes antara lain:

1. *Gaussian Naive Bayes*

Gaussian Bayes biasanya digunakan untuk merepresentasikan probabilitas bersyarat dari fitur *continue* pada sebuah kelas $P(X_i \mid Y)$, dan dikarakteristikkan dengan dua parameter : *mean* dan *varian.*

2. Bernaulli Naïve Bayes

Pada *Bernaulli Naïve Bayes*, pembobotan dilakukan dengan menggunakan *binary* (0 dan 1) dalam pembobotan tiap *term*, hal ini berbeda dengan perhitungan term frekuensi yang melakukan pembobotan pada setiap *term*.

3. Multinomial Naïve Bayes

Multinomial Naïve bayes mengasumsikan indepedensi diantara kemunculan kata-kata dalam dokumen, tanpa memperhitungkan urutan kata dan konteks informasi dalam kalimat atau dokumen secara umum. Selain itu metode ini memperhitungkan jumlah kemunculan kata dalam dokumen (Destuardi and Sumpeno, 2009).

2.7. Improved Naïve Bayes

Berdasarkan kekurangan yang ada pada metode *Naïve Bayes* biasa, maka dilakukan penambahan agar metode *Naïve Bayes* menghasilkan hasil yang maksimal. Penambahan-penambahan tersebut antara lain: *stemming*.

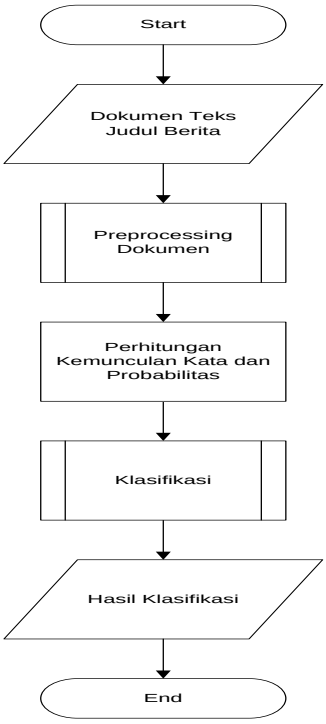
1. Untuk meningkatkan nilai estimasi dari $p(d_i \mid C_j)$ dengan menggunakan *Multivariate Bernoulli Event Model*. Alasan menggunakan model ini tidak menggunakan banyak fitur. Fitur dikurangi dengan cara menggunakan bobot dengan nilai *Boolean* (0 atau 1). Sehingga jika terdapat fitur kata yang muncul dua kali maka hanya satu fitur yang digunakan. Pada *Multivariate Bernouli Event Model* Bobot bernilai 1 jika kata itu muncul dan bernilai 0 bila kata tersebut tidak muncul pada dokumen (He, Xie and Xu, 2011).
2. Mengurangi dimensi $d_i = (w_i, \dots, w_j \dots, w_m)$ melalui segmentasi kata dokumen l mungkin memiliki duplikat fitur kata, sebagai contoh ada $w_i = w_j$, dimana $1 <= l$, $j <= m$, karena karakteristik dari *Multivariate Bernoulli Event Model* yaitu fiturnya tidak mengingat banyaknya kejadian kata dalam teks untuk menghitung $p(w_i \mid C_j)$ maka ketika $w_i = w_j$ dapat menghapus kata yang sama $d_i = (w_i, \dots, w_j \dots, w_m)$, dimana $t <= m$, hal tersebut menyebabkan dimensi dari sebuah dokumen akan semakin kecil karena kata yang terduplikasi tidak digunakan kembali (Yuan, 2010).

2.8. Perancangan Sistem

Perancangan sistem digunakan untuk memperlihatkan alur dari jalannya sistem secara umum sehingga mempermudah dalam proses implementasi. Proses perancangan sistem antara lain:

1. *Dokumen judul berita berupa file dengan format txt.*
   File dokumen judul berita berisi beberapa jenis judul berita. Satu baris judul berita merepresentasikan satu dokumen. Kemudian dokumen tersebut akan dilakukan *preprosesing*. Tahap-tahap dari *preprosesing* terdiri dari *case folding*, *tokenizing*, *filtering* dan *stemming*.
2. *Perhitungan kemunculan kata dan probabilitas nilai kata terhadap kategori.*
3. *Klasifikasi dengan menggunakan metode Improved Naïve Bayes.*

Alur kerja sistem secara umum dapat dilihat pada Gambar 6.



Gambar 6. Alur Kerja Sistem

Tahap dari alur kerja sistem terdiri dari beberapa bagian utama:

1. *Dokumen Judul Berita*
   Dokumen judul berita berupa file dengan format.txt isi file berupa beberapa jenis judul berita. Satu baris judul berita merepresentasikan satu dokumen. Kemudian, dokumen tersebut akan dilakukan *preprocessing*.

2. *Preprocessing*

   Pada *preprocessing* akan dilakukan *case folding*, *tokenizing*, *filtering*, dan *stemming*. Pada penelitian ini stemming dengan menggunakan Lucene.

3. *Perhitungan nilai probabilitas kata terhadap kategori.*

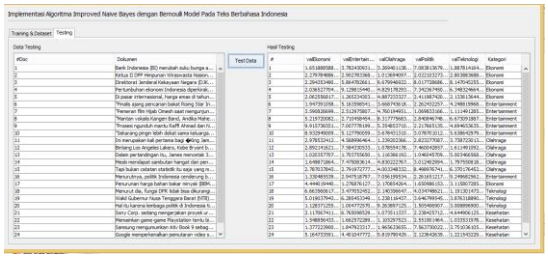
   Setelah *preprocessing* selesai, dilakukan pengurangan fitur dengan menggunakan nilai *boolean* (0 dan 1), nilai 0 apabila kata tidak muncul pada dokumen, sedangkan nilai 1 apabila kata muncul dalam dokumen. Setelah itu dilakukan perhitungan prior lalu perhitungan nilai peluang kata terhadap kategori.

4. Klasifikasi Dokumen

   Pada tahap ini, dilakukan perhitungan nilai posterior, dokumen akan akan masuk ke dalam suatu kelas yang memiliki nilai posterior yang tertinggi.

2.9. Implementasi

Implementasi aplikasi klasifikasi teks dengan menggunakan *Improved Naïve Bayes* dilakukan dengan mengacu pada perancangan sistem yang telah dilakukan. Implementasi perangkat lunak dilakukan menggunakan Bahasa pemprograman Java. Untuk hasil dari implementasi sistem dapat dilihat pada Gambar 7.



Gambar 7. Implementasi Sistem

3 PENGUJIAN

Bab ini membahas pengujian yang dilakukan terhadap sistem yang telah dibuat. Selain itu, bab ini juga menjelaskan analisis terhadap hasil dari implementasi dan pengujian yang telah dilakukan.

3.1. Skenario Pengujian

Pengujian ini menjelaskan tentang pengujian dengan menggunakan metode *Improved Naïve Bayes*. Pengujian variasi merupakan pengujian pengaruh dari jumlah data latih yang bervariasi dan data uji yang tetap. Komposisi data latih dan data uji yang digunakan dalam pengujian ini seperti tertera pada Tabel 1.

Tabel 1. Komposisi Data Pengujian dan Data Latih Per Kategori

Jumlah Data Latih	Jumlah data Uji
50	300
100	300
150	300
200	300
250	300

3.2. Hasil dan Analisa Pengujian

Hasil pengujian dengan menggunakan metode *Improved Naïve Bayes* dengan menggunakan jumlah data latih yang berbeda dan data uji yang tetap ditampilkan pada Tabel 2 hingga Tabel 6.

Tabel 2. Hasil Skenario Pengujian dengan Jumlah Data Latih 50

Jenis Berita	TP	FP	FN	Precision	Recall	F-Measure
Ekonomi	185	23	64	0.889423	0.742972	0.809628
Entertainment	108	0	187	1	0.366102	0.53598
Kesehatan	163	2	124	0.987879	0.567944	0.721239
Makanan	213	3	76	0.986111	0.737024	0.843564
Olahraga	188	1	106	0.994709	0.639456	0.778468
Otomotif	204	3	86	0.985507	0.703448	0.820926
Teknologi	107	13	147	0.891667	0.42126	0.572193
Travel	205	10	76	0.953488	0.729537	0.826613
Rata-rata				0.961098	0.613468	0.738576

Tabel 3. Hasil Skenario Pengujian dengan Jumlah Data Latih 100

Jenis Berita	TP	FP	FN	Precision	Recall	F-Measure
Ekonomi	207	30	42	0.873418	0.831325	0.851852
Entertainment	139	0	156	1	0.471186	0.640553
Kesehatan	195	5	89	0.975	0.68662	0.805785
Makanan	244	5	47	0.97992	0.838488	0.903704
Olahraga	213	1	79	0.995327	0.729452	0.841897
Otomotif	223	3	66	0.986726	0.771626	0.866019
Teknologi	140	11	112	0.927152	0.555556	0.694789
Travel	213	6	68	0.972603	0.758007	0.852
Rata-rata				0.963768	0.705283	0.807075

Tabel 4. Hasil Skenario Pengujian dengan Jumlah Data Latih 150

Jenis Berita	TP	FP	FN	Precision	Recall	F-Measure
Ekonomi	216	32	33	0.870968	0.86747	0.869215
Entertainment	174	0	121	1	0.589831	0.742004
Kesehatan	202	9	81	0.957346	0.713781	0.817814
Makanan	247	5	44	0.980159	0.848797	0.909761
Olahraga	228	3	62	0.987013	0.786207	0.87524
Otomotif	228	3	62	0.987013	0.786207	0.87524
Teknologi	151	10	103	0.937888	0.594488	0.727711
Travel	222	9	58	0.961039	0.792857	0.868885
Rata-rata				0.960178	0.747455	0.835734

Tabel 5. Hasil Skenario Pengujian dengan Jumlah Data Latih 200

Jenis Berita	TP	FP	FN	Precision	Recall	F-Measure
Ekonomi	211	38	36	0.84739	0.854251	0.850806
Entertainment	180	0	115	1	0.610169	0.757895
Kesehatan	217	9	66	0.960177	0.766784	0.852652
Makanan	248	3	43	0.988048	0.852234	0.915129
Olahraga	230	3	60	0.987124	0.793103	0.879541
Otomotif	230	3	60	0.987124	0.793103	0.879541
Teknologi	148	12	106	0.925	0.582677	0.714976
Travel	212	8	68	0.963636	0.757143	0.848
Rata-rata				0.957312	0.751183	0.837318

Tabel 6. Hasil Skenario Pengujian dengan Jumlah Data Latih 250

Jenis Berita	TP	FP	FN	Precision	Recall	F-Measure
Ekonomi	220	37	29	0.856031	0.883534	0.869565
Entertainment	188	0	107	1	0.637288	0.778468
Kesehatan	230	10	53	0.958333	0.812721	0.879541
Makanan	258	3	33	0.988506	0.886598	0.934783
Olahraga	247	2	45	0.991968	0.84589	0.913124
Otomotif	238	2	52	0.991667	0.82069	0.898113
Teknologi	157	8	95	0.951515	0.623016	0.752998
Travel	225	8	55	0.965665	0.803571	0.877193
Rata-rata				0.962961	0.789164	0.862973

Berdasarkan hasil pengujian pada Tabel 2 sampai Tabel 6 diatas, rata-rata *precision*, *recall* dan *f-measure* ditampilkan pada Tabel 7.

Tabel 7. Rata-rata Nilai Precision, Recall, dan F-Measure

Jumlah data latih	Precision	Recall	F-Measure
50	0.961098	0.613468	0.738576
100	0.963768	0.705283	0.807075
150	0.960178	0.747455	0.835734
200	0.957312	0.751183	0.837318
250	0.962961	0.789164	0.862973

Berdasarkan pada Tabel 7 hasil pengujian secara umum meningkat. Pada pengujian *recall,* nilai tertinggi diperoleh pada penggunaan data latih sebesar 250 data sedangkan nilai terendahnya ketika menggunakan data latih sebesar 50 data.

Hasil pada pengujian *f-measure*, nilai tertinggi diperoleh dari penggunaan data latih sebesar 250 data dan nilai terendahnya ketika menggunakan data latih sebesar 50 data. Semakin banyak data, semakin bagus.

Sedangkan pada pengujian *precision* pada penggunaan data latih 150 dan 200 terjadi penurunan, tetapi penurunan tersebut tidak signifikan. Nilai *precision* tertinggi pada penelitian ini ketika menggunakan data latih 100 dan nilai *precision* terendah ketika menggunakan data latih 200.

Dari hasil percobaaan diatas penggunaan metode *Improved Naïve Bayes* dapat digunakan dalam mengklasifikasikan teks yang berukuran pendek karena algoritma ini nilai pembobotan kata (*Term*) tidak berdasarkan frekuensi kata sehingga bisa memperbesar nilai peluang kata terhadap kelasnya.

4 KESIMPULAN

Berdasarkan pada pengujian yang telah dilakukan pada Klasifikasi Berita Twitter Menggunakan Metode *Improved Naïve Bayes* dapat diambil kesimpulan. Pada penggunaan Improved Naïve Bayes tahapan klasifikasi dimulai dari proses *preprocessing*. Pada proses *preprocessing* dilakukan proses *case folding*, tokenizing, *filtering* dan *stemming*. Setelah proses *preprocessing* selesai dilakukan perhitungan kemunculan kata dan probabilitas, kata yang muncul akan 1 sedangkan yang tidak akan diberi nilai 0. Setelah itu baru dilakukan klasifikasi. Klasifikasi *Improved Naïve Bayes* menghasilkan *precision* atau tingkat kesesuaian

antara informasi yang diperoleh dari sistem dengan informasi yang diinginkan oleh pengguna, *recall* atau pengukuran dari jumlah dokumen benar yang yang berhasil diklasifikasikan oleh sistem, dan *f-measure* atau pengukuran untuk mengetahui tingkat keseimbangan antara *precision* dan *recall* terbaik pada data latih 250 dengan nilai *precision* 0.962961, *recall* 0.789164 dan *f-measure* sebesar 0.862973. Jumlah data latih mempengaruhi performa. semakin banyak data latih yang digunakan, semakin bagus.

DAFTAR PUSTAKA

Agusta, L., Kristen, U. and Wacana, S., 2009. Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia. *Konferensi Nasional Sistem dan Informatika 2009*, (KNS&I09-036), pp.196–201.

Chen, X., Jianfang, X. and Youquan, H., 2011. An improved Naive Bayesian algorithm for Web page text classification - *Proceedings - 2011 8th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2011*, pp. 2765-2768.

Destuardi, I. and Sumpeno, S., 2009. Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naive Bayes. *Seminar Nasional Pascasarjana Institut Teknologi Sepuluh Nopember*, (c).

Feldman, R., & James, S. 2006. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Dara. Cambridge: Cambridge University Press.

Phuvipadawat, S. and Murata, T., 2010. Breaking news detection and tracking in Twitter. *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2010*, pp.120–123.

Kwak, H., Lee, C., Park, H. and Moon, S., 2010. What is Twitter, a Social Network or a News Media ? Categories and Subject Descriptors. *Www 2010*, pp.591–600.

Weissbock, J., Esmin, A., Inkpen, D. 2013. Using external information for classifying tweets. *Proceedings - 2013 Brazilian Conference on Intelligent Systems, BRACIS 2013*, pp. 1-5.

www.twitter.com [diakses 1 2017]

Yuan, L., 2010. *An Improved Naive Bayes Text Classification Algorithm In Chinese Information Processing*. Jiaozuo, s.n.

Perdana, R. & Rekyan M. 2013. Pengkategorian Pesan Singkat Berbahasa Indonesia pada Jejaring Sosial Twitter dengan Metode Klasifikasi Naïve Bayes, pp. 1-12.

Torunoğlu, D., Çakirman, E., Ganiz, M.C., Akyokuş, S. and Gürbüz, M.Z., 2011. Analysis of preprocessing methods on classification of Turkish texts. *INISTA 2011 - 2011 International Symposium on INnovations in Intelligent SysTems and Applications*, pp.112–117.

Widodo, A.W., 2013. Klasifikasi Artikel Berita Menggunakan Naive Bayes Classifier yang Dimodifikasi.