TUGAS STOKASTIK
REVIEW PAPER
M JALALUDDIN JABBAR
146060300111024
TEKNIK ELEKTRO UNIVERSITAS BRAWIJAYA

Judul Paper	: Data mining of automatically promotion tweet for products and services using naïve bayes algorithm to encrease twitter engagement followers at PT. Bobobobo
Jurnal	: Sciencedirect
Tahun	: 2015
Penulis	: James Luke dan Suharjito

1. PENDAHULUAN

Twitter merupakan salah satu media social populer di dunia. Indonesia menempati posisi ke 5 pengguna terbesar di dunia, setiap hari server twitter menerima data tweet dengan jumlah besar, dengan demikian kita dapat melakukan data mining yang digunakan untuk tujuan tertentu, salah satunya adalah sebagai media promosi suatu produk ataupun jasa seperti halnya penelitian yang dilakukan oleh James luke dan Suharjito ini.

Untuk data skala besar sangat dibutuhkan kecepatan dalam proses pencarian data. Sehingga dibutuhkan pengelompokan data terlebih dahulu. Naive Bayes merupakan algoritma pembelajaran untuk klasifikasi dengan efisiensi komputasi dan akurasi yang baik, khususnya untuk dimensi dan jumlah data yang besar.untuk itu dalam penelitian ini akan membuktikan kemampuan naïve bayes classifier untuk mengklasifikasikan tweet yang berisi informasi tentang suatu produk dan jasa, studi kasus penelitian ini dilaksanakan di PT Bobobobo Jakarta.

2. METODE

Data mining merupakan sebuah proses dari knowledge discovery (penemuan pengetahuan) dari data yang sangat besar. Sementara itu text mining merupakan bidang data mining yang bertujuan untuk mengumpulkan informasi yang berguna dari data teks dalam bahasa alami atau proses analisis data teks kemudian mengekstrak informasi yang berguna untuk tujuan tertentu.

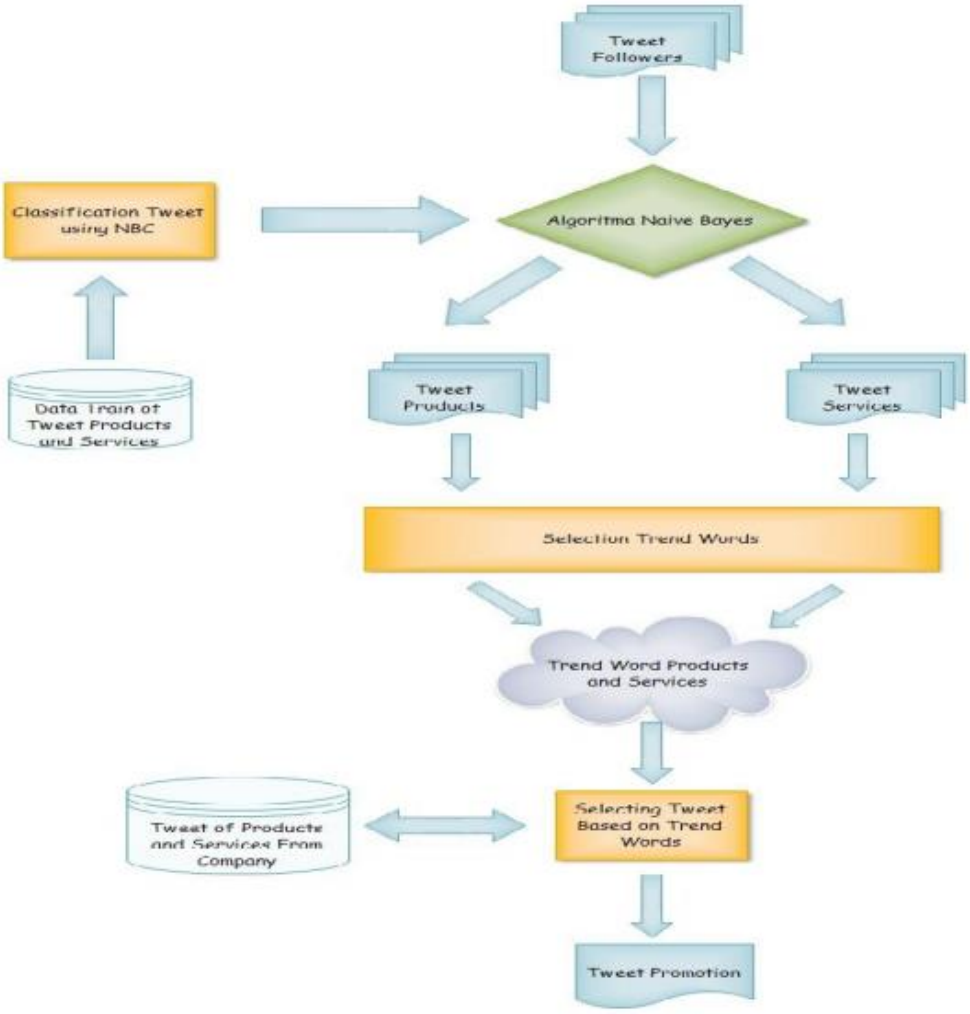
Algoritma yang digunakan pada penelitian ini adalah Naïve bayes classifier (NBC). Naïve bayes classifier merupakan salah satu metode machine learning yang

memanfaatkan perhitungan probabilitas dan statistic yang dikemukakan oleh ilmuan prancis Thomas Bayes. Yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya.

Proses klasifikasi dilakukan berdasarkan persamaan :

$$p(C|D) = \frac{p(C)}{p(D)} p(D|C)$$
$$= \frac{p(C)}{p(D)} \prod_i p(w_i|C)$$

Proses penelitian yang dilakukan dapat dilihat pada gambar 1



Gambar 1 Kerangka Penelitian

2.1 Data

Data yang digunakan pada penelitian ini adalah tweet berbahasa Indonesia pada daerah Jakarta, dari bulan Juni 2013 sampai dengan bulan Pebruari 2014 dari data tersebut kemudian dibagi menjadi dua kategori/class : produk dan jasa.berikut beberapa contoh dari dua kategori tersebut :

Kategori produk :

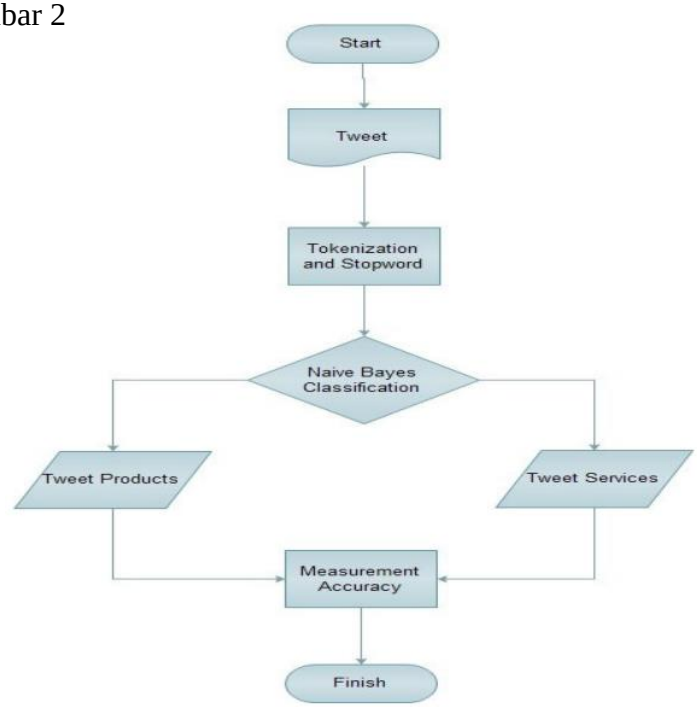
- "Liattas,sepatu,kerudung,bajubawaannya pengenbelisemua"
- "Nyari dress batik: " kalodapet yang disuka, antarmahalbingit di kantongataudressnya…sepasangsamakemeja…"

Kategori Jasa :

- "kelas yoga after hours lumayanbikinbadanlentur yah cyn…"
- "liburankeparis, traveling mahameru, study banding keausi"
- "Cari hotel murahdidekatpusat Kota Surabaya?ada saran?"

2.2 Klasifikasi tweet dengan NBC

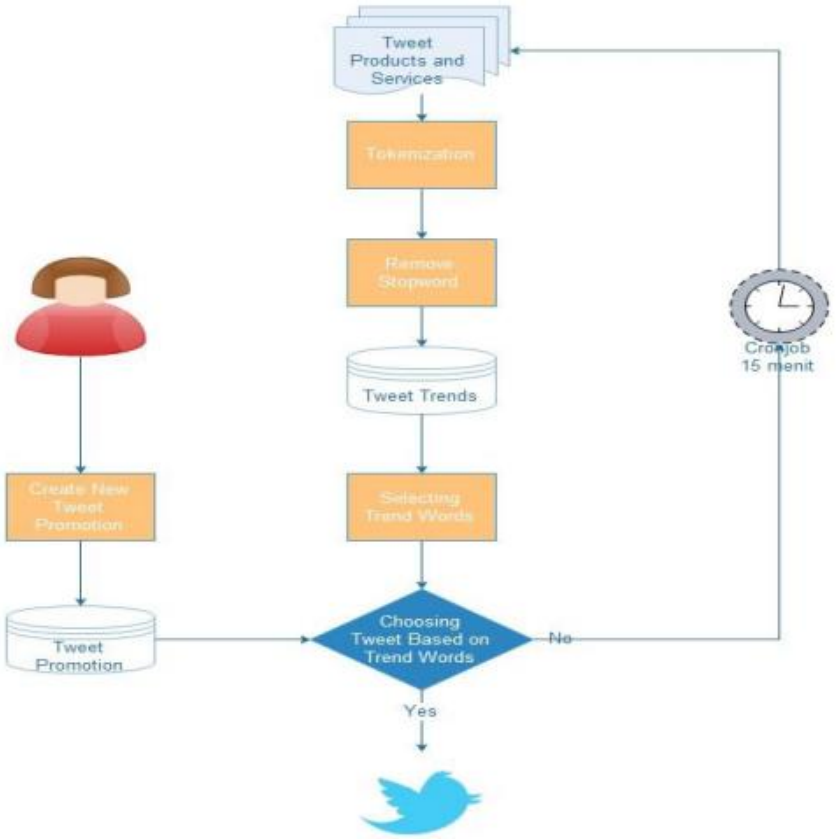
Setelah data tweet dikumpulkan, kemudian data digunakan sebagai dataset training dan pengelompokan berdasarkan class/atribut produk dan jasa. Proses selanjutnya adalah klasifikasi menggunakan algoritma Naïve Bayes Classifier (NBC) untuk mengukur tingkat akurasinya. Proses klasifikasi tweet digambar dengan flowchart pada gambar 2



Gambar 2 Flowchart klasifikasi menggunakan NBC

2.3 Seleksi Trend Word Dan Tweet

Setelah tweet diklasifikasi kedalam class produk dan jasa. Tahapan selanjutnya adalah memilih kata popular/trend words dan membandingkannya dengan tweet yang akan dijadikan promosi secara otomatis.berikut alur dari proses seleksi dan promosi tweet. Proses selesi dan promosi ditunjukkan pada Gambar 3



Gambar 3 flowchart seleksi dan promosi otomatis

Langkah seleksi kata popular/trend words dan tweet yang akan dijadikan bahan promosi adalah sebagai berikut :

1. Tokenization

   Tahap pemotongan string input berdasarkan kata yang menyusunnya.

   kata dipisahkan dari tweet yang sebagai tanda. kata itu dianggap valid apabila terdiri atas 3-25 huruf dan bukan link atau url
2. Stopword

Merupakan kata-kata yang tidak berpengaruh terhadap proses klasifikasi. Hasil dari proses ini disimpan dalam database

3. Trend words

   Dengan menggunakan query dari database untuk mendapatkan kembali 5 kata dari kategori produk dan jasa

4. Promoting tweets

   Promosi tweet ditulis oleh admin twitter, setiap tweet dispesifikasikan dengan tenggang waktu promosi, keyword, dan score kecocokan dengan trend word. Kemudian tweet disimpan dalam database

5. Tweet automatically

   Proses query ke database untuk mendapatkan kecocokan/match promosi tweet dengan trend words. Apabila setiap kriteria cocok, maka secara otomatis system akan mentweet.

4. EVALUASI

Penelitian ini memiliki dua tahapan, sebagai berikut:

1. Analisis akurasi dari algoritma Naïve Bayes Classier

Pada pengukuran akurasi algoritma naïve bayes classifier, tweet/data dibentuk kedalam tiga variasi data pelatihan dan pengujian. Setiap tweet produk dan jasa diuji . distribusi dari data pelatihan dan data pengujian ditunjukkan pada table 1

Table 1 komposisi dan variasi data pelatihan

Data Test (tweet)	**Data Train (tweet)**		
	Product = Service	**Product > Service**	**Product < Service**
2000 P/S	500 P & 500 S	700 P & 300 S	300 P & 700 S
1500 P/S	1500 P & 1500 S	2000 P & 1000 S	1000 P & 2000 S
1000 P/S	2500 P & 2500 S	3500 P & 1500 S	1500 P & 3500 S
500 P/S	3500 P & 3500 S	4900 P & 2100 S	2100 P & 4900 S
50 P/S	4500 P & 4500 S	6300 P & 2700 S	2700 P & 6300 S

2. Analisis peningkatan keterlibatan follower

Untuk mengukur Performa keterlibatan follower, dihitung dengan persamaan berikut :

$$\text{tweet engagement rate} = \frac{\frac{\text{replies+retweets}}{\text{jumlah tweet}}}{\text{total followers}} \times 100$$

3. HASIL DAN DISKUSI

3.1 Hasil klasifikasi tweet menggunakan NBC

Digunakan komposisi variasi dari data pelatihan/training pada semua kategori produk, dan menghasilkan seperti yang ditunjukkan pada table 2 berikut

Tabel 2 hasil NBC menggunakan data tweet produk

Data Test (tweet)	Data Train		
	Product = Service	Product > Service	Product < Service
2000	77.65%	18.85%	96.95%
1500	87.20%	20.13%	98.20%
1000	91.90%	11.20%	98.90%
500	96.80%	12.80%	99.20%
50	98.00%	3.20%	99.60%
Average	**90.31%**	**13.24%**	**98.57%**

Penggunaan komposisi variasi dari data training untuk tweet kategori jasa, hasilnya seperti yang ditunjukkan pada table 3

Table 3 hasil NBC menggunakan data tweet servis/jasa

Data Test (tweet)	Data Train		
	Product = Jasa	Product > Jasa	Product < Jasa
2000	70.65%	97.05%	6.45%
1500	73.50%	98.86%	10.73%
1000	79.60%	99.00%	4.40%
500	85.80%	99.10%	3.00%
50	95.00%	99.60%	4.30%
Average	**80.91%**	**98.72%**	**5.77%**

Komposisi variasi data training terhadap kombinasi dari tweet kategori produk dan jasa.

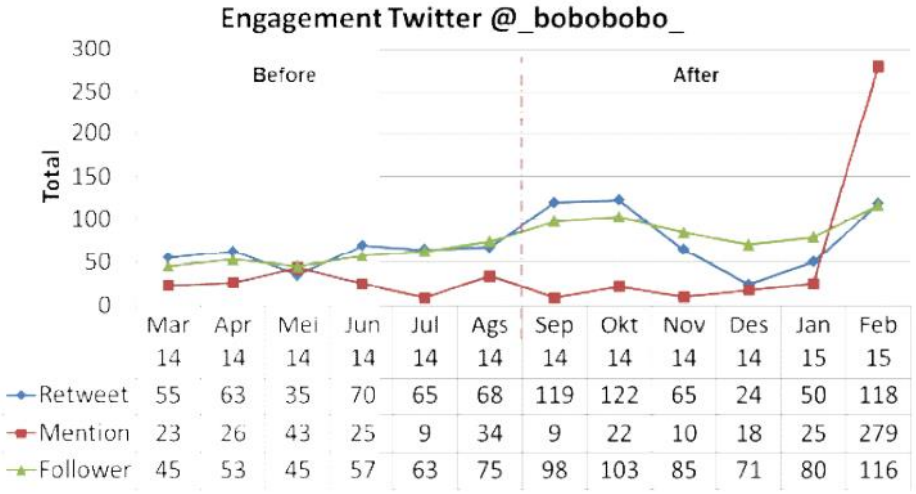
Tabel 4 hasil NBC (kombinasi data tweet produk dan jasa)

Data Test (tweet)	Data Train (tweet)	Accuration
2000 P & 2000 S	500 P & 500 S	**74.50%**
1500 P & 1500 S	1500 P & 1500 S	**78.63%**
1000 P & 1000 S	2500 P & 2500 S	**83.55%**
500 P & 500 S	3500 P & 3500 S	**88.80%**
50 P & 50 S	4500 P & 4500 S	**92.10%**
Average		**83.51%**

Dari hasil data pengujian diatas, algoritma NBC memiliki tingkat akurasi yang cukup baik.

3.2 Hasil otomatisasi promosi tweet

Dari eksperimen yang dilakukan pada bulan September 2014 hingga pebruari 2015, menghasilkan



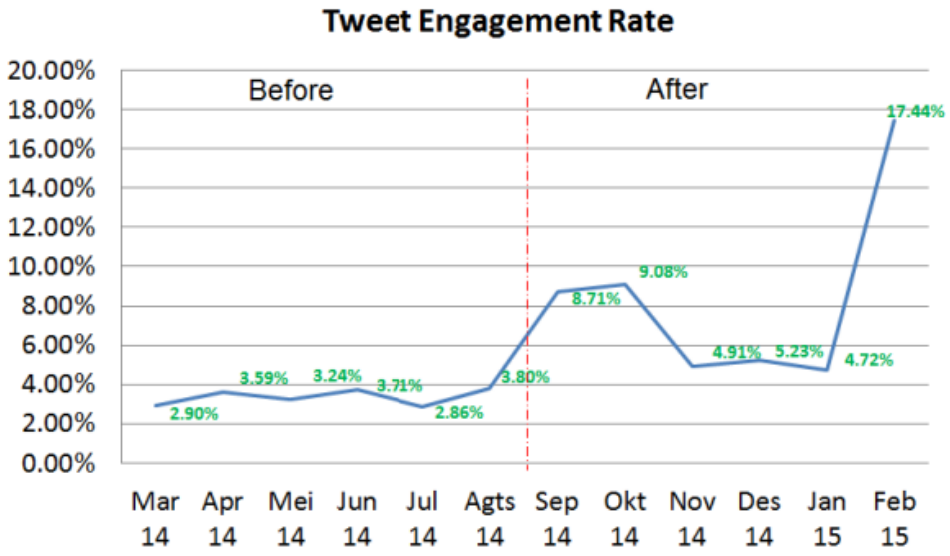
Gambar 4 Engagement twitter @_bobobobo_

Dari gambar diatas, kalkulasi rata-rata telah dilakukan dengan mengkombinasikan hasil dari retweet, mentions dan penambahan follower sebelum dan setelahnya. Ditunjukkan pada table 5

Tabel 5 Peningkatan keterlibatan follower

	Before	After	Results
Retweet	59.33	83	**39%**
Mention	26.67	60.5	**120%**
Follower	56.33	92.16	**63%**

Tabel diatas menunjukkan bahwa peningkatan dari retweet dan mention akan sangat berpengaruh terhadap peningkatan keterlibatan dari follower. Hasil perbandingan keterlibatan follower sebelum dan setelah implementasi ditunjukkan pada gambar 5



Gambar 5 tingkat keterlibatan sebelum dan setelah implementasi

Penelitian ini memberikan hasil yang positif pada peningkatan keterlibatan follower.

4. KESIMPULAN

Kesimpulan dari penelitian ini adalah :

1. Algortma NBC memiliki tingkat akurasi yang tinggi dalam proses klasifikasi ditunjukkan dengan tingkat akurasi mencapai 90,31% menggunakan data uji kategori produk, dan 80,91% menggunakan data uji kategori jasa. Dan kombinasi dari keduanya menghasilkan akurasi 83.51%

2. Banyak dari stopword bisa menentukan trendword dari koleksi tweet kategori produk dan jasa
3. Peningkatan aktivitas terjadi di twitter oleh penelitian ini, untuk tweet mencapai 39%, mention 120% dan follower baru 69%. me-retweet dan mention memberikan dampak terhadap hasil keterlibatan follower.
4. Jumlah tingkat keterlibatan tweet, setelah penelitian ini memberikan hasil yang cukup tinggi 17.44% dan terendah 4.72%. jika dibandingkan dengan studi sebelumnya tertinggi 3.80% dan terendah 2.90%
5. Dengan menggunakan twitter sebagai media promosi, memberikan hasil yang cukup memuaskan, sebelum me-ngetweet kita bisa menganalisa trend word/tranding topic dari follower, dimana memberikan respon yang baik dari follower.

Catatan :

Ini hanyalah sebuah tugas, mungkin masih banyak kesalahan. Kritik dan saran sangat diharapkan

Link download Journal

http://www.sciencedirect.com/science/article/pii/S1877050915020797