

**SENTIMENT ANALYSIS DENGAN NAIVE BAYES UNTUK MELIHAT PERSEPSI
MASYARAKAT TERHADAP BATIK
PADA JEJARING SOSIAL TWITTER**

Anindya Apriliyanti P.

Departemen Statistika FMIPA Universitas Padjadjaran

a.apriliyanti.p@gmail.com

ABSTRAK. *Sentiment analysis* atau yang disebut juga dengan *opini mining* merupakan analisis yang bertujuan untuk melihat opini masyarakat atau kelompok mengenai entitas tertentu. *Sentiment analysis* sendiri sering digunakan untuk menilai sebuah produk, apakah memberikan review yang baik dan disukai masyarakat ataupun sebaliknya. Penelitian ini dilakukan untuk mengklasifikasikan opini atau sentiment masyarakat terhadap batik. Batik merupakan salah satu identitas budaya bangsa Indonesia. Pandangan masyarakat terhadap batik, tertuang dalam sebuah persepsi, baik itu positif maupun negatif. Dengan semakin meleknya teknologi, opini dan persepsi masyarakat banyak dibagikan melalui jejaring sosial, salah satunya adalah twitter. Oleh karena itu, penelitian ini mengambil opini-opini masyarakat terhadap batik melalui twitter dengan hashtag #batik. Metode yang dipakai adalah metode *naive bayes*, yang dalam hal ini klasifikasi dikategorikan dalam *supervised learning* yang berakibat opini tersebut terbagi kedalam kutub *sentiment* positif dan negatif. Hasil penelitian menunjukkan bahwa sentiment positif terbentuk paling banyak yaitu 494 tweets dibandingkan dengan sentiment negatif yang hanya 5 tweets. Jadi secara keseluruhan, persepsi masyarakat mengenai batik yang tertuang pada jejaring sosial twitter memberikan penilaian yang positif.

Kata Kunci: *sentiment analysis; naive bayes; persepsi; batik.*

1. PENDAHULUAN

Dewasa ini, informasi sangat mudah diperoleh dan dibagikan. Adanya media sosial seperti facebook ataupun twitter memberikan akses seluas-luasnya bagi pengguna internet untuk menggali serta membagikan informasi. Media jejaring sosial kerap dimanfaatkan untuk menjadi media dalam menuangkan uneg-uneg ataupun opini terhadap entitas tertentu. Hal ini sering dimanfaatkan oleh produsen untuk melihat review terhadap produk yang dipasarkannya. Apakah memberikan penilaian yang baik atau sebaliknya dimata masyarakat.

Informasi yang tertulis pada jejaring sosial mayoritas adalah berupa data teks. Opini serta pendapat tertulis dalam kalimat-kalimat yang berupa teks. Oleh karena itu diperlukan sebuah analisis yang tepat, untuk mengolah data teks sehingga memberikan informasi yang berharga. *Text mining* merupakan salah satu bidang ilmu yang mempelajari analisis dengan inputan berupa data teks.

Sentiment analysis atau yang disebut juga dengan *opini mining* merupakan analisis yang bertujuan untuk melihat opini masyarakat atau kelompok mengenai entitas tertentu. Salah satu metode dari teks mining yang bisa digunakan untuk menyelesaikan masalah sentimen analisis adalah *Naive Bayes*, karena metode ini dapat mengklasifikasikan opini kedalam opini positif maupun negatif.

Penelitian ini dilakukan untuk mengklasifikasikan opini atau sentimen masyarakat terhadap batik. Batik merupakan salah satu identitas budaya bangsa Indonesia. Pandangan masyarakat terhadap batik, tertuang dalam sebuah persepsi, baik itu positif maupun negatif. Penelitian ini mengambil opini-opini masyarakat terhadap batik melalui twitter dengan hashtag #batik.

2. METODE PENELITIAN

Text mining merupakan proses penggalian terhadap sumber-sumber tertulis yang besar untuk menghasilkan informasi yang baru [1]. Data yang diperoleh dapat berupa data yang bersifat kualitatif maupun data yang tidak terstruktur. Secara umum teks mining bertujuan untuk menggali informasi dari sekumpulan dokumen teks dalam ukuran besar dan dalam format yang tidak terstruktur. Secara khusus, tujuan dari text mining dapat dibagi menjadi dua [5]:

1. Pengkategorisasian data teks (*text categorization*)
Dalam pengkategorisasian, *text mining* dipergunakan sebagai alat untuk menemukan kategori yang sesuai dengan kelas yang telah ditentukan (*supervised learning*)
2. Pengelompokan data teks (*text clustering*)
Pada pengelompokan, *text mining* berfungsi sebagai alat untuk mengelompokkan data teks berdasarkan kesamaan karakteristik, dan *clustering* dapat digunakan untuk memberikan label pada kelas yang belum diketahui (*unsupervised learning*)

Pada penelitian ini, pengkategorisasian data teks merupakan tujuan utama, karena penelitian ini ingin mengkategorikan sebuah tweet atau cuitan mengenai #batik masuk dalam kategori positif atau negatif. Untuk merepresentasikan dokumen teks, diperlukan sebuah tahapan awal atau *preprocessing* dokumen agar dapat diolah menggunakan *sentiment analysis*.

2.1. Representasi Dokumen Teks

Tahapan *preprocessing* terdiri atas *tokenizer*, penghilangan *stopwords*, *stemming* dan pembobotan *term* (secara gambar dapat dilihat pada Gambar 1). *Tokenizer* adalah proses pemecahan dokumen teks menjadi beberapa token atau kata, pembatasan biasanya dilakukan berdasarkan spasi atau tanda baca. Tahap selanjutnya adalah proses penghapusan kata-kata yang sesuai pada daftar *stopwords*. *Stopwords* merupakan daftar kata-kata yang dianggap

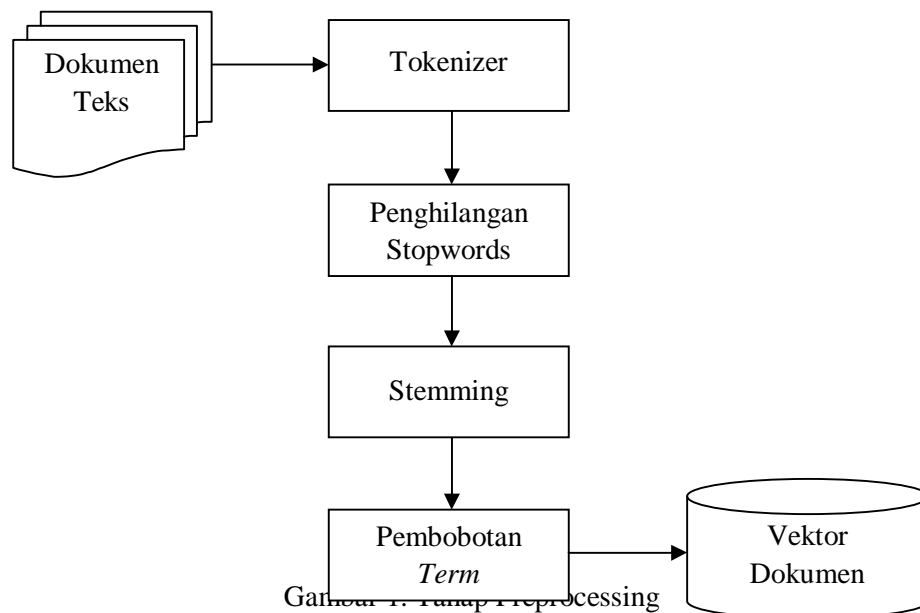
tidak merepresentasikan konten dari suatu dokumen teks. Kata-kata yang tersisa dari tahapan ini kemudian dilakukan proses *stemming*, yaitu merubah kata menjadi bentuk dasarnya [2]. Setiap kata yang dihasilkan dapat disebut sebagai *term*.

Setiap *term* yang berbeda satu dengan yang lain kemudian diberikan bobot *term* dengan menggunakan TF-IDF, rumusnya adalah

$$w_{t,d} = (1 + \log_{10} tf_{t,d})(\log_{10} n/df_t) \quad (1)$$

dengan n adalah jumlah keseluruhan dokumen, $tf_{t,d}$ merupakan jumlah kemunculan *term* pada dokumen d , dan df_t adalah jumlah dokumen yang memuat *term* pada keseluruhan dokumen.

Hasil dari tahapan *preprocessing* adalah sebuah vektor yang berisi *term-term* disertai dengan bobot TF-IDF. Setelah melewati tahap *preprocessing*, akan dilakukan klasifikasi terhadap *term* yang terbentuk menggunakan *sentiment analysis* dengan algoritma *Naive Bayes*.



2.2. Sentiment Analysis dengan algoritma Naïve Bayes

Sentiment analysis atau yang sering disebut juga sebagai *opinion mining* merupakan salah satu metode dalam *text mining* yang bertujuan menganalisis opini, sentimen, sikap, penilaian, maupun emosi pada suatu entitas, seperti produk, jasa, organisasi, individu, isu, kejadian ataupun topik tertentu [3]. Secara umum, sentiment analysis terbagi dalam tiga tingkatan, yaitu pada level dokumen, level kalimat, hingga pada level entitas. Klasifikasi dalam

sentiment analisis dapat dilakukan dengan berbagai macam algoritma, salah satunya adalah dengan metode *Naïve bayes*.

Naive bayes merupakan metode klasifikasi dalam *text mining* berdasarkan probabilitas dan teorema bayes. Adapun asumsi yang harus dipenuhi yaitu adanya independensi dalam variable bebas [4]. Maksudnya adalah kemunculan sebuah kata tidak mempengaruhi kemunculan kata lainnya, ataupun sebaliknya.

Analisis *bayesian* menggali dua buah informasi tentang parameter suatu model statistik, yang mana bila digabungkan informasi ini akan membentuk informasi porterior. Peluang dari distribusi posterior dari parameter yang tidak diketahui diperoleh dari data dan juga distribusi priornya. Metode bayesian menyelidiki bentuk hubungan antara distribusi prior dan fungsi *likelihood*nya. Algoritma *Naive Bayes Classifier* dengan asumsi semua atribut independen dapat dituliskan sebagai berikut [5]:

$$\begin{aligned} P(X_1, X_2, \dots, X_n | Y) &= P(X_1 | X_2, \dots, X_n, C) P(X_2, \dots, X_n | Y) \\ &= P(X_1 | C) P(X_2, \dots, X_n | Y) \\ &= P(X_1 | C) P(X_2 | C) \dots P(X_n | C) \end{aligned} \quad (2)$$

Formula (2) dapat diterapkan pada klasifikasi dokumen teks. Misalkan diketahui bahwa peluang term ke-*i* pada dokumen *D* terjadi di suatu kelas *A* adalah $P(t_i | A)$. Maka peluang dokumen tersebut terjadi di kelas *A* adalah $P(D | A) = \prod_i P(t_i | A)$. Apabila kelas terdiri dari kelas positif (A_{pos}) dan kelas negatif (A_{neg}), maka peluang dokumen *D* masuk kelas positif adalah

$$P(A_{pos} | D) = \frac{P(A_{pos}) P(D | A_{pos})}{P(D)} = \frac{P(A_{pos}) \prod_i P(t_i | A_{pos})}{P(D)}$$

demikian juga untuk peluang *D* masuk dalam dokumen negatif

$$P(A_{neg} | D) = \frac{P(A_{neg}) P(D | A_{neg})}{P(D)} = \frac{P(A_{neg}) \prod_i P(t_i | A_{neg})}{P(D)}$$

Rasio dari keduanya diperoleh:

$$\frac{P(A_{pos} | D)}{P(A_{neg} | D)} = \frac{P(A_{pos})}{P(A_{neg})} \prod_i \frac{P(t_i | A_{pos})}{P(t_i | A_{neg})}$$

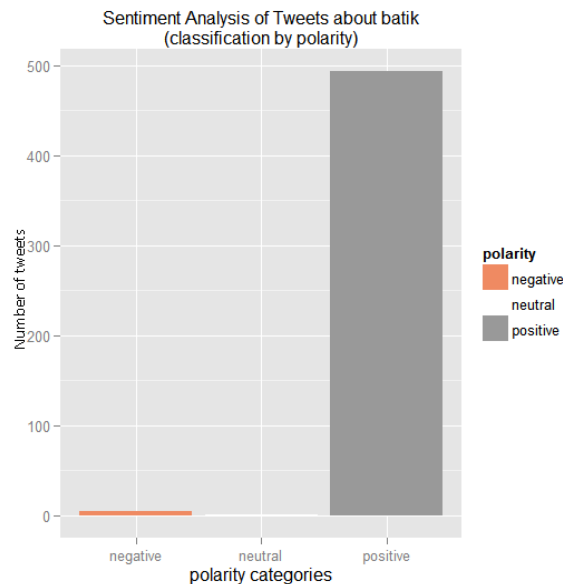
dengan mengambil logaritma di setiap ratio diatas diperoleh

$$\ln \frac{P(A_{pos} | D)}{P(A_{neg} | D)} = \ln \frac{P(A_{pos})}{P(A_{neg})} + \sum_i \ln \frac{P(t_i | A_{pos})}{P(t_i | A_{neg})} \quad (3)$$

Dari persamaan (3), maka klasifikasi dokumen akan terbentuk. Dokumen dikategorikan positif jika $P(A_{pos}|D) > P(A_{neg}|D)$ atau $\ln \frac{P(A_{pos}|D)}{P(A_{neg}|D)} > 0$, demikian pula sebaliknya. Sedangkan dokumen akan masuk kelas netral jika $\ln \frac{P(A_{pos}|D)}{P(A_{neg}|D)} = 0$.

3. HASIL PENELITIAN DAN PEMBAHASAN

Data pengamatan merupakan 500 dokumen yang diambil dari twitter dengan kata kunci #batik. Setelah melalui tahapan *preprocessing*, diperoleh vector dengan *64term* yang dilengkapi pembobotan TF-IDF. Dengan analisis lanjutan diperoleh berbagai hasil.

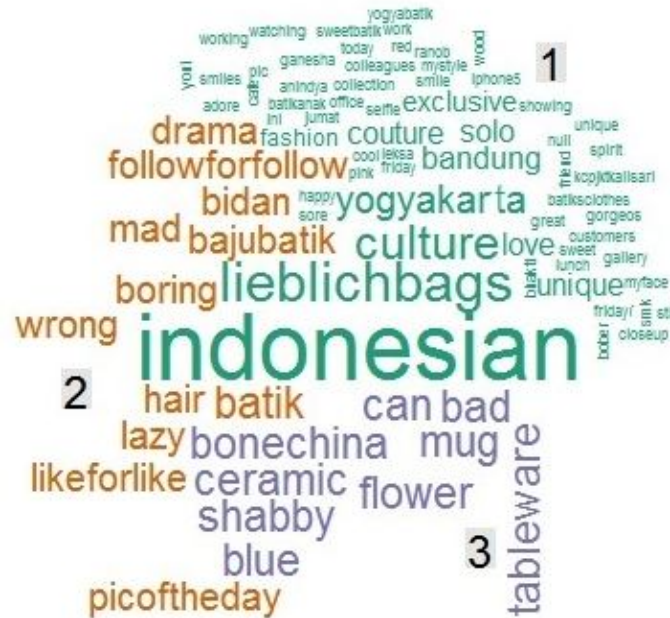


Gambar 2. Hasil klasifikasi tweets berdasarkan polaritas

Dari 500 dokumen, yang terklasifikasi kedalam sentimen kelas positif adalah sebanyak 494 dokumen, kelas negatif sebanyak 5 dokumen dan kelas netral hanya ada 1 dokumen. Gambar 2, memberikan plot polaritas dari sentimen kelas positif, netral dan negatif. Dari Gambar 2 dapat dilihat bahwa respon atau pandangan masyarakat terhadap batik tergolong sangat baik. Dengan perhitungan lebih lanjut, diperoleh peluang kelas positif adalah sebesar 0.988, peluang kelas negatif sebesar 0.01 dan kelas netral 0.002. kelas netral terbentuk apabila tweet tidak merepresentasikan opini yang positif atau negatif. Kelas netral merupakan bentuk representasi tweet yang objektif.

Gambar 3 memberikan *wordcloud* dari *term-term*. Terdapat tiga kelompok yang secara berturut-turut merepresentasikan emosi yang diwakili oleh “joy”, “anger” dan

“sadness”. Kelompok 1, yang merepresentasikan “joy” merupakan kumpulan *term* pada tweet kelas positif dan netral. Sedangkan, kelompok 2 dan 3 yang merepresentasikan “anger” dan “sadness” merupakan *term-term* pada kelas negatif.



Gambar 3. Wordcloud untuk *term* positif dan negatif.

Dari Gambar 3, terlihat bahwa untuk kelompok 1 yang merepresentasikan kelas sentiment positif, batik dijelaskan oleh *term-term* “indonesian”, “lieblichbags”, “culture”, “yogyakarta”, “bandung”, “solo”, “couture”, “exclusive”, dan “unique”. Hal ini berarti batik sangat lekat dengan *culture* atau kebudayaan Indonesia yang eksklusif dan unik. Adapun kota-kota di Indonesia yang banyak disebut dalam kaitannya dengan batik adalah Yogyakarta, Bandung dan Solo. Peneliti menemukan hal menarik terkait dengan *term* “lieblichbags”, dengan mesin pencarian, peneliti menemukan bahwa lieblich merupakan nama produk tas yang menggunakan batik sebagai bahan utama (<http://lieblichbags.com/>). Dengan banyaknya tweet mengenai lieblichbags, maka dapat disimpulkan bahwa saat ini di dunia maya sedang gandrung dengan tas batik merk lieblich.

Hal lain yang dapat menarik perhatian adalah mengenai tweet bernaga negatif yang ditemukan. Berikut adalah tweet dengan *term-term* yang telah melalui *preprocessing*:

- batik mug can bad shabby flower blue tableware ceramic bonechina
- bad hair day save batik
- lazy wrong bidan batik
- boring picoftheday likeforlike followforfollow bajubatik batik
- mad summer nights bad dream drama batik indonesian style

dari berbagai tweets tersebut, terlihat bahwa bentuk persepsi negatif tidak ditujukan kepada batik. Tweets tersebut dimasukkan ke kelas sentimen negatif, kemungkinan karena mengandung *term-term* seperti “bad”, “shabby”, “wrong”, “boring” ataupun “mad”. Jadi secara keseluruhan, opini masyarakat mengenai batik yang tertuang pada jejaring sosial twitter memberikan penilaian yang positif.

4. SIMPULAN

Dari berbagai analisis dapat diberikan hasil bahwasespon atau pandangan masyarakat terhadap batik tergolong sangat baik, dengan klasifikasi tweets kedalam sentiment kelas positif adalah sebanyak 494 dokumen, kelasnegatif sebanyak 5 dokumen dan kelas netral hanya ada 1 dokumen. Berdasarkan wordcloud, dapat disimpulkan bahwa batik sangat lekat dengan *culture* atau kebudayaan Indonesia yang eksklusif dan unik. Adapun kota-kota di Indonesia yang banyak disebut kaitannya dengan batik adalah Yogyakarta, Bandung dan Solo. Tweets yang tergolong pada kelas negatif tidak mengarah pada opini negatif tentang batik, jadi secara keseluruhan, opini masyarakat mengenai batik yang tertuang pada jejaring sosial twitter memberikan penilaian yang positif.

DAFTAR PUSTAKA

- [1] Mitra, Mandar. 2014. *Introduction to Text Mining*. Indian Statistical Institute.
- [2] Lama, Prabin. 2013. “Clustering System Based On Text Mining Using The K-Means Algorithm”. Thesis: Turku University Of Applied Sciences, Southwest Finland.
- [3] Liu, Bing. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publisher.
- [4] Destuardi dan Sumpeno, Surya. 2009. “Klasifikasi Emosi untuk Teks Bahasa Indonesia menggunakan Metode Naive Bayes”. Surabaya: Institut Teknologi Sepuluh Nopember.
- [5] Sihombing, Eirene. 2014. “ Penerapan Analisis Sentimen dengan Metode Naïve Bayes pada Klasifikasi Data Teks”. Bandung: Universitas Padjadjaran.