Perbandingan Algoritma Data Mining pada Analisis Sentimen Twitter Menggunakan framework CRISP-DM

DATA MINING – Kelas B

Alfian Ardiansyah

**Pendahuluan**

Data mining merupakan proses eksplorasi data agar memahami dan memprediksi menggunakan algoritma machine learning agar dapat memprediksi potensi kejadian yang akan terjadi melalui data yang telah diolah. Hal ini membantu dalam pengambilan keputusan untuk masa depan dengan cara menggambarkan. Proses data mining memiliki tiga hal utama, yang pertama adalah satistik yang artinya penelitian numerik terhadap relasi data, kecerdasan buatan yang meniru kecerdasan manusia di dalam sebuah software atau mesin, dan machine learning yang tujuannya memahami guna data dan membuatnya bermanfaat bagi masa depan.

Dengan perkembangan teknologi, proses pengolahan data lebih mudah saat ini dan lebih mudah untuk diimplementasikan dibandingkan dengan cara manual yang memakan waktu proses yang lebih lama. Sektor yang mengurus bisnis seperti bank, manufaktur, dan pedagang retail telah menggunakan teknologi maju dari data mining agar dapat memahami potensi di masa depan dari berbagai hal seperti prediksi kenaikan harga, memahami trend konsumen, dan lain-lain.

Twitter merupakan media sosial dan berita dari amerika yang tiap orang mengungkapkan perasaan dan emosi dalam sebuah tulisan yang bernama sebuah ‘tweets’. Tweets ini bisa dilihat oleh publik tetapi pengirim dapat membatasinya untuk teman atau beberapa orang saja. Tweets hanya dibatasi 280 karakter, memiliki banyak guna dan fungsi seperti review film, untuk memahami trend yang sedang berlangsung, dan lain-lain. Tujuan penelitian ini adalah membuat model yang dapat menjelaskan dan mensimpulkan sentimen menggunakan tweets dari suatu topik. Ada beberapa alasan mengapa twitter dapat digunakan untuk memprediksi sesuatu di masa depan, yaitu :

* Twitter mudah digunakan, dengan cara membuat akun baru lalu seseorang dapat melakukan tweets setiap saat tergantung pengguna.
* Sebuah tweets menerbitkan pesan pendek yang disebarkan dan diketahui oleh orang yang memfollow anda.
* Twitter juga bisa jadi sarana microblogging yang artinya membuat konten terhadap suatu topik dengan format tulisan seperti di blog tetapi pendek.
* Dengan dibatasnya karakter di twitter, justru membuat twitter sangat populer.
* Seseorang dapat mengirim tweets dikarenakan keangkuhan, meminta perhatian, promosi diri sendiri, atau sekadar kebosanan, dan masih banyak alasan lainnya.
* Banyak orang menggunakan twitter sebagai sarana rekruit, konsultasi bisnis, dan toko menggunakan twitter, dan dapat berhasil.

**Business Understanding**

Analisis sentimen menggunakan twitter merupakan model yang dibuat agar mengetahui sentimen pengguna dari teks opini atau tweets yang ada. Didefinisikan sebagai proses teknik text mining yang diekstrak lalu selannjutnya dipelajari untuk menentukan polaritas pendapat dari dokumen teks yang diberikan. Dalam prinsip implementasi ini, kita berfokus terhadap seberapa akurat pemahamanan respon dengan penilaian manusia itu sendiri dan berbagai kesimpulan lain seperti pemahaman kemiripan teks dan fokus topik.Pemahaman dan perbandingan ini terutama dihitung dengan pengukuran yang berbeda tergantung dengan presisi dan keakuratan dengan membagi menjadi tiga variabel dari reaksi negatif, netral, dan positif.

Meskipun kebanyakan hasil akhir sentimen analisis ini digunakan untuk urusan bisnis, hasil akhirnya bisa juga dapat digunakan untuk menggambarkan kondisi politik, dalam penelitian ini kita berfokus pada penerapan analisis sentimental salah satu topik politik, yaitu tweets yang mengandung kata prabowo.

Media sosial tidak hanya penting bagi kehidupan sehari-hari tetapi juga menjadi sumber daya yang berharga untuk urusan bisnis. Dengan usaha dan biaya yang sedikit, urusan bisnis dapat memiliki wawasan yang dapat digunakan untuk keputusan selanjutnya. Dari customer service ke pemasaran, analisis sentimen dapat membantu semuanya. Tetapi bukan berarti data mining menggunakan algoritma machine learning ini dapat 100% digunakan untuk membuat keputusan. Meskipun begitu ini masih sangat bermanfaat untuk berkembangnya segala urusan maupun itu bisnis sampai politik.

**Data Understanding**

Dataset didapatkan dari crawling dan filter data dari twitter yang mengandung kata prabowo. Ada dua macam data yaitu data original yang didapat langsung dari crawling dan filtering dari twitter dengan dan data pengolahan lanjut tersebut yang dilabeli secara manual.

Data pertama yaitu data original yang tidak dilabeli secara manual, memiliki 296 tabel dan 2 kolom yang menunjukkan kolom Text, Id, polarity. Dengan kolom polarity sebagai label.

Data kedua sama berasal dari data original tersebut tetapi dilakukan manual labeling oleh peneliti. Memiliki 101 tabel dan 3 kolom yang menunjukkan kolom No., TEXT, dan Polarity\_manual. Polarity manual sebagai label di dalam rapidminer.

Setelah file data excel dimasukkan proses selanjutnya yaitu peneliti melakukan labelling manual untuk mendeteksi emoticon yang gunanya mengenali emosi sedih dan senang. Memiliki 97 tabel dan 2 kolom yang tidak memiliki keterangan. Selanjutnya yang akan di multiply untuk jalur proses yang berbeda.

**Data Preparation**

Data yang akan kita gunakan masih butuh untuk disaring dan disiapkan untuk proses selanjutnya. Tentunya data crawling dan filtering dari twitter sendiri masih belum rapi dan bersih sehingga penyaringan ulang masih sangat dibutuhkan. Namun sebelum itu juga masih dibutuhkan mengubah data nominal yang ada menjadi teks lalu peneliti akan menggunakan operator yang bernama ‘Process Document From Data’. Di dalam operator tersebut akan digunakan berbagai macam operasi untuk memfilter data. Proses ini dilakukan dua kali di aliran yang berbeda. Di dalam operator ‘filter stopwords’ kita menggunakan dictionary yang artinya stopwordsnya ditentukan secara manual, begitu pula dengan operator ‘Stem’. Perlu diingat sebelum melakukan operasi-operasi ini perlu ditentukan kolom polarity sebagai polynomial label, operator itu adalah.

* **Transform cases**

Operator ini digunakan untuk mengubah semua karakter menjadi huruf kecil.

* **Tokenize**

Operator ini akan memisahkan kalimat menjadi tiap kata atau yang disebut tokenisasi.

* **Filter Tokens (by length)**

Operator ini memfilter token tergantung dengan panjangnya. Penulis akan memberikan range atau jarak dari 4 sampai 25. Jadi, operator akan filter setiap kata kecuali kata yang ada di range tersebut.

* **Filter Stopwords (Dictionary)**

Operator ini akan menghilangkan stopwords. Namun untuk mengenali dan bekerja pada bahasa indonesia butuh pengenalan secara manual. Oleh karena itu peneliti memakai ‘Filter Stopwords (Dictionary)’ dengan menggunakan file “stopword\_indo\_new.txt” yang katanya sudah ditentukan sebelumnya.

* **Stem (Dictionary)**

Guna operator adalah mengubah semua kata mejadi bentuk dasarnya. Dengan menggunakan file “stem.txt”. Contohnya seperti kata lah,kah,ku,mu,isme,di, dan lain-lain.

**Modelling**

Akan digunakannya 3 model yang berbeda dalam proses ini untuk menentukan model mana yang memiliki tingkat keakuratan paling tinggi dengan dataset yang telah ditentukan polaritasnya secara manual. Juga akan digunakannya data original dan data yang sudah dilabeli manual untuk membandingkan performa model mana yang terbaik dengan data yang berbeda. Setelah mencoba berbagai macam model, model yang paling akurat akan digunakan untuk evaluasi dan deployment.

Disini kita akan menggunakan cross validation yang cocok untuk data yang sudah diolah dan tidak terlalu besar, juga dikarenakan sudah banyak dilakukan preprocessing seperti labeling manual, replace menggunakan emotion, Process Documents dari data, dan dictionary bahasa indonesia manual. Digunakan k-fold sebanyak 10 kali, jumlah ini direkomendasikan sebagai jumlah standar dalam menentukan model mana yang paling baik karena kecenderungan memberikan estimasi akurasi yang kurang bias akhirnya tidak berat sebelah terhadap salah satu model tertentu.

Algoritma yang akan digunakan adalah k-NN, decision tree, dan naive bayes dengan diterapkannya operator ‘apply model’ untuk menerapkan tiap model ke bagian testing dataset. Lalu akan digunakan operator ‘performance’ untuk mengetahui performa dari tiap model. Performa yang didapatkan tiap model adalah sebagai berikut :

1. **k-NN**
2. **Decision Tree**
3. **Naive Bayes**

Dari semua model ini, naive bayes memiliki tingkat akurasi tinggi hingga mencapai 100%

**Evaluation**

Sekarang kita telah mendapatkan hasil dari berbagai macam model. Tetapi sebelum ke proses selanjutnya kita perlu melihat semua proses yang telah dilakukan. Ada beberapa langkah untuk memeriksa apakah kita telah mencapai tujuan yang diinginkan dengan meninjau model. Juga, kita akan memeriksa apakah ada masalah potensi kesalahan dipertimbangkan diperbolehkan atau tidak. Langkah-langkah evaluasi tersebut sebagai berikut.

* Evaluasi hasil
* Review proses
* Menentukan langkah selanjutnya

Sekarang kita telah mendapatkan performa dari berbagai macam model dan akhirnya akan digunakan model naive bayes dikarenakan memiliki tingkat akurasi paling tinggi yaitu sebesar 100%. Naive bayes adalah merupakan keluarga “probalistic classifier” sederhana yang didasarkan penerapan teorema bayes dengan asumsi independensi yang kuat (naive) diantara fitur-fitur model lainnya. Namun pernyataan ini masih perlu di cek apakah naive bayes masih cocok dengan dataset lainnya. Maka dari itu dibutuhkannya perbandingan dengan menggunakan dataset telah disediakan sebelumnya yaitu dengan menggunakan dataset original yang belum dilabeli secara manual atau murni dari crawling dan filtering proses pencarian tweets oleh rapidminer.

**Deployment**

Bagian ini adalah proses akhir dalam metodologi CRISP-DM. Peneliti telah melakukan berbagai macam algoritma model terhadap salah satu dataset. Namun di bagian ini akan dilakukan perbandingan dengan melakukan percobaan dengan dataset yang sama namun belum labeli secara manual. Jadi peneliti akan menerapkan kembali proses modelling dengan dataset tersebut.

1. **k-NN**
2. **Decision Tree**
3. **Naive Bayes**

Ternyata walaupun menggunakan data asli yang belum diolah secara manual akurasi oleh naive bayes tetap memiliki akurasi 100%

**Conclusion**

Telah disimpulkan bahwa sentimen analisis ini dapat diselesaikan dengan metodologi CRISP-DM untuk melakukan penelitian ini.

Walaupun level dataset berbeda pertama telah dilabeli secara manual dan yang lainnya data yang tidak diolah lebih lanjut, persentase hasilnya saling mendekati walaupun dengan model yang berbeda-beda.

Pendefinisian dataset yang belum diolah banyak memiliki nilai polaritas netral, dikarenakan proses crawling dan filtering dari twitter tersebut hanya mengenali dalam bahasa inggris, maka dari itu dasaset tersebut perlu dilabeli lebih lanjut agar mengenali berbagai macam reaksi dalam bahasa indonesia.

Setelah dataset telah diperbaharui polaritas secara manual, maka data ini pantas dipakai untuk dilakukan sentimen analisis yang berkaitan dalam bahasa indonesia.

Proses selanjutnya adalah pendefinisian filter stopword secara custom dengan memakai kata henti dan stemming yaitu pengenalan kata imbuhan yang dikenal dalam bahasa indonesia di dalam ‘process documents from data’.

Secara keseluruhan, prediksi akurat sesuai dengan dataset yang dipakai dengan naive bayes memiliki tingkat keakuratan yang paling tinggi. Namun bukan berarti keakuratan tinggi sesuai dengan tujuan pengolahan data seperti kasus dalam penelitian ini.