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

DATA MINING - Kelas B
Alfian Ardiansyah - 15650063

Pendahuluan

Data mining merupakan proses eksplorasi agar memahami dan memprediksi data menggunakan algoritma *machine learning* agar dapat memprediksi potensi kejadian yang akan terjadi melalui data yang telah diolah. Hal ini membantu dalam pengambilan langkah keputusan untuk menprediksi masa depan. 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 menjadi bermanfaat. (Dey et al., 2017)

Dengan perkembangan teknologi, proses pengolahan data lebih mudah saat ini dan lebih dapat dipahami 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.

Framework yang digunakan untuk mendefinisikan alur kerja dalam penelitian ini adalah CRISP-DM. Model CRISP-DM (Cross-Industry Standard Process for Data Mining) yang menyediakan alur kerja proyek data mining. Yang memiliki enam fase yaitu *business understanding, data understanding, data preparation, modelling, evaluation,* dan *deployment*. (Pete et al., 2000)

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. Sebuah *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 : (Noor & Turan, 2020)

- 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 mengikuti.

- 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. Dan pembatasan karakter ini sangat membantu kejelasan sentimen yang ada.
- Seseorang dapat mengirim *tweets* dikarenakan pamer sesuatu, 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.

Pemahaman Bisnis

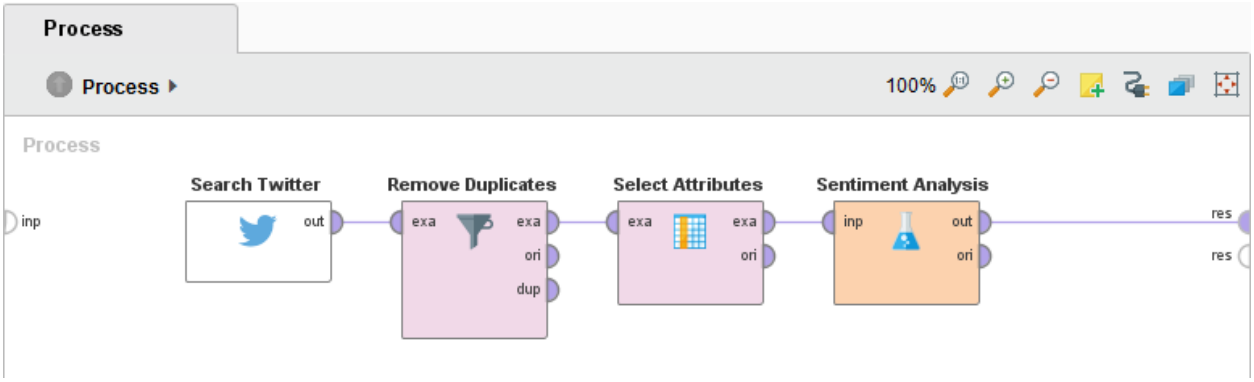
Analisis sentimen menggunakan twitter merupakan model yang dibuat agar mengetahui sentimen pengguna dari teks opini atau *tweets* yang ada. Didefinisikan sebagai proses teknik text mining yang diekstrak lalu selannjutnya dipelajari untuk menentukan polaritas pendapat dari dokumen teks yang diberikan. Dalam prinsip implementasi ini, peneliti 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. (Mundalik, 2018)

Meskipun kebanyakan hasil akhir sentimen analisis ini digunakan untuk urusan bisnis, hasil akhirnya bisa juga dapat digunakan untuk menggambarkan kondisi politik, dalam penelitian ini peneliti 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 pelayanan konsumen ke pemasaran, analisis sentimen dapat membantu semuanya. Tetapi bukan berarti *data mining* menggunakan algoritma *machine learning* ini dapat 100% digunakan untuk membuat keputusan namun lebih menjadi sebagai data pertimbangan. Meskipun begitu ini masih sangat bermanfaat untuk berkembangnya segala urusan maupun itu bisnis sampai politik.

Pemahaman Data

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



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

Text	Id	polarity
Para Mantan Relawan Kapok Pilih Jokowi, Pindah ke Prabowo #DebatPilpres2019 #D	1116873535723429889	neutral
Gatot Nurmantyo Beri Kejutan Di Akhir Kampanye, Dukung Prabowo-Sandi #KlikRM	1116629536064454657	neutral
Dahlan Iskan: Hari Ini Saya Menjatuhkan Pilihan ke Pak Prabowo https://t.co/EhKiul	1116660541865377793	neutral
RT @ArieKuntung: “ANCAMAN” ustadz Adi Hidayat kepada pak Prabowo https://t.c	1117103072730181632	neutral
RT @wonhoseleraku: #DebatPilpres2019Prabowo: "You have been a president for 4	1117103072449187840	negative
@MichelAdamNew @prabowo @sandiuno Wahh... Ini serangan terhadap PDIP dan	1117103072142979072	neutral
RT @nusanewsid: Gara-gara Ustadz Somad, Sempat Dukung Jokowi Lalu Golput, Akt	1117103072084238338	neutral
RT @republikaonline: Aa Gym: Bismillah, Prabowo-Sandi Pilihan Hati https://t.co/s	1117103071971041282	neutral
RT @PakarLogika: Jokowi: Apa yang akan bapak lakukan untuk pengembangan e-sp	1117103071920709633	neutral
RT @PartaiSocmed: Betapapun pencapaian Prabowo tidak ada seujung kukunya SB	1117103071870341121	neutral
RT @Dae_1r0ne: Sikap ? Ini SIKAP Saya : Tetap Konsisten Sampai 17 April 2019 Saya	1117103071534784512	neutral
RT @MichelAdamNew: JAKA SEMBUNG AGAIN!01 membanggakan perbankan syaria	1117103071375400960	neutral
RT @detikcom: "...karena jawaban bapak tadi pertanian, saya jadi kok nggak sambu	1117103071006322688	neutral
RT @putrabanten80: Kode Keras Dari Aa Gym "Cuma Ngasih Tau Ajah"Pesanteren D	1117103070645604352	neutral
RT @RachlanNashidik: Pak Prabowo sebenarnya sedang berdebat dengan siapa? Ke	1117103069509103616	neutral
RT @PakarLogika: Jokowi: Apa yang akan bapak lakukan untuk pengembangan e-sp	1117103069303431168	neutral
RT @EmasPadi: @Beritasatu Cebong g ada yg komen yk??cb polling ulangRT ?? utk F	1117103069248966661	neutral
RT @Suara_Bawah: 05. Gue udah tahu bahwa Demokrat udah tahu setengah hati ma	1117103069173403648	neutral
@HeraLoebs @prabowo @sandiuno BKL #HoaxJkwMenangTotalDebat	1117103068837859328	neutral
@haierikaa @PartaiSocmed Skrng dah jelas kann yg dijawab prabowo ga melencen	1117103068825264128	neutral
RT @liputan6dotcom: Ditanya Jokowi soal E-Sport, Prabowo Jawab tentang Pangan	1117103068653350912	neutral
RT @Bambanghariyan: MANTUL.....Pernyataan Penutup Dari Prabowo Sandi, Bahwa	1117103068598824960	neutral

Data kedua sama berasal dari data orisinil namun telah dilakukan labeling secara manual oleh peneliti. Memiliki 101 tabel dan 3 kolom yang menunjukkan kolom No., TEXT, dan Polarity_manual. Polarity manual sebagai label di dalam rapidminer.

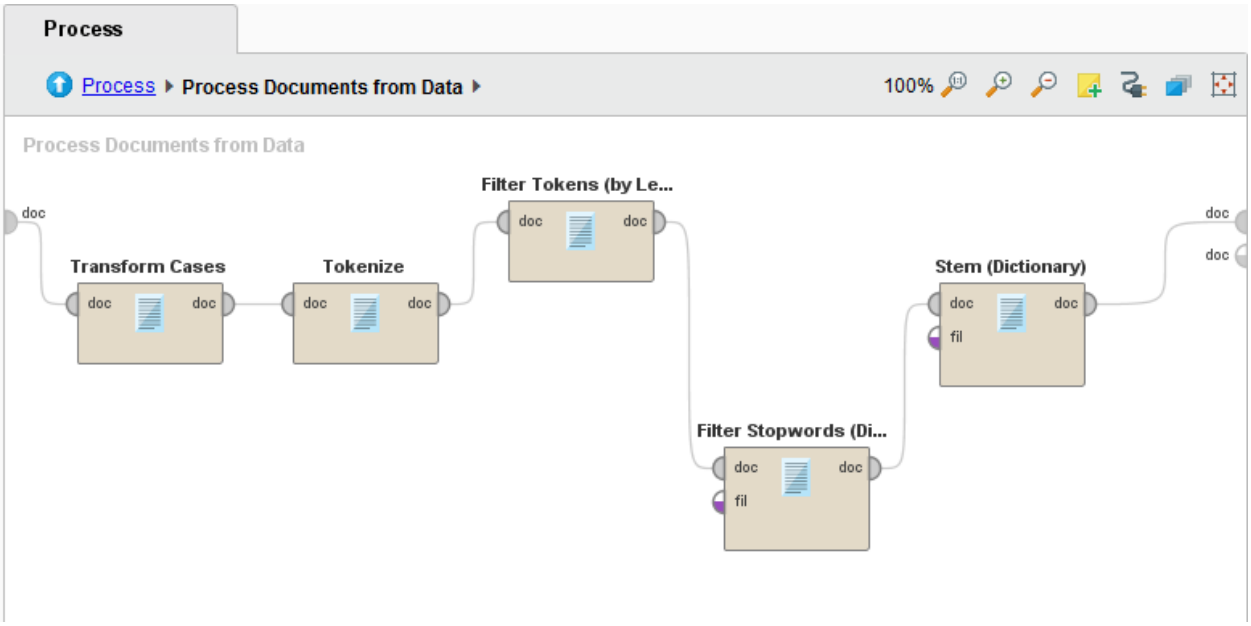
No.	TEXT	POLARITY MANUALLY
1	RT @anonLokal: Dengan berbagai cara kotor REZIM ini dan ANTEK2nya menghalangi Bpk @prabowo menjadi Presiden. BIADAB MEREKA!!	negative
2	GERAKKAN !Telah terjadi peembersihan sampah oleh kader partai pada saat kampanye Prabowo di Bandung	positive
3	RT @geloraco: RR: Ayah Prabowo Pernah Bantu Malaysia Rumuskan Kebijakan Ekonomi Pro Bumiputra	neutral
4	RT @ira_dyahloka: Emak emak dilawan ?? Biar dilarang_kampanye di mall yang penting Prabowo menang ?????❁❤ #PrabowoDanSandiPilihanRakyat #1…	positive
5	RT @AjiMustir: @prabowo Terimakasih pak telah datang ke manado saya bangga bisa salaman sama bapak prabowo, insya allah bapak dan bang sand…	positive
6	RT @aiek_esthreem: Rakyat sabar menunggu !!Rakyat sabar menanti !!Tapi Rakyat melawan jika ditantang !!18 Hari lagi Kita punya Presiden…	positive
7	RT @tempodotco: Prabowo: "Masih mau dicurangi atau tidak? Kalau tidak, 17 April jaga TPS. Bawa lontong, bawa ketupat, bawa sarung, bawa tik…	neutral
8	RT @SuaraWarganet: Ingat, 27 April 2019Tusuk Prabowo SandiTUMBANGKAN Presiden Selfie https://t.co/WmJDJkhTB0	neutral
9	RT @KasanMulyono: Prabowo Soal Larang Tahlil dan Dukung Radikalisme: Itu Fitnah#PKIvsPANCASILA □@prabowo□ □@sandiuno□ https://t.co/Vq9N…	negative
10	RT @Serulingbambu4: @Riast10 @_Ainz__ @MichelAdamNew @dherma98 @PadiOyan @Bang_tea2 @hadid501 @Sulistyo_1987 @Winarko_yuli @Tizels16	neutral
11	RT @fadlizon: Jumlah massa peserta Kampanye Akbar di Pakansari hingga ke jalan-jalan mencapai satu juta orang. Luar biasa Bogor. @prabowo @…	positive
12	@marierteman @JihanfahiraREAL @prabowo @sandiuno Ihh bapak kok punya senior cemen ngomong nya khilafa terus. Yg jelas #PKIvsPANCASILA mangka	negative
13	RT @putrabanten80: Masya Allah...Spontanitas Dari Relawan "PADI Probumi" Yang Menyumbang Dana Sebesar 200juta Buat Perjuangan Indonesia Adi…	neutral
14	Pak Woo berjanji akan pulangkan Habib Rizieq kalau terpilih!KALAU terpilih!!!!??Siapa si Habib Rizieq itu???@YRadianto@heniunique#PANCASIALAvsKHILAF	positive
15	RT @Parodi_Negeri: ?? : Pret, ternyata wajah JkW itu mirip Prabowo.?? : Ah, masa ??? : Iya pret ! Kemarin waktu mau kampanye, beliau lewat…	neutral
16	RT @CNNIndonesia: Prabowo Sindir Politikus Lupa Balas Budi Usai Jadi Wali Kota https://t.co/knoJNFQuPN	neutral
17	RT @RelawanProSandi: Luarbiasa! Kampanye gembira, sampahpun segera bersih berkah relawan2 #BahagialahPemilihPrabowoSandi Lokasi Kampanye…	positive
18	RT @fadlizon: Kampanye di Bogor, Prabowo Diarak Massa hingga ke Atas Panggung https://t.co/rEigXNsQxp	neutral
19	RT @merdekadotcom: BPN Prabowo: Saya Kasih Bocoran, Novel Baswedan akan Jadi Jaksa Agung https://t.co/fqkb63EkMm	neutral
20	@isbejoIC @yuniepertiwi @eae18 @prabowo Tukang sobek makalah anak pmii	negative
21	RT @_indrajaya78_: KH.Mahfudz Asirun, Rois Suriah PWNU "Tanggal 17 April semua wajib datang ke TPS (Tusuk @prabowo @sandiuno)" ?? #Mena…	neutral
22	RT @CakKhum: Saya Mendukung Novel Baswedan Menjadi Jaksa Agung Saat Prabowo Dilantik Presiden.Sikat habis maling² duit rakyat, bakal penu…	positive

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.

sebelum	sesudah
--	sedih
--"	sedih
%-(	sedih
)":	sedih
):	sedih
)-:	sedih
_	sedih
___	sedih
:"(	sedih
:#	sedih
:-&	sedih
:(	sedih
:'(	sedih
:-(	sedih
:((	sedih
:(((	sedih
:((((	sedih
:(((((	sedih
:((((((	sedih
:(((((((	sedih
:/	sedih
:-/	sedih

Persiapan Data

Data yang akan peneliti 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 masih dibutuhkan mengubah data nominal ke bentuk 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' peneliti menggunakan *dictionary* yang artinya kata hentinya ditentukan secara manual, begitu pula dengan operator 'Stem'. Perlu diingat sebelum melakukan operasi-operasi ini perlu ditentukan kolom polarity sebagai polynomial label, Berikut adalah operator yang ada di dalam proses dokumen dari sebuah data.



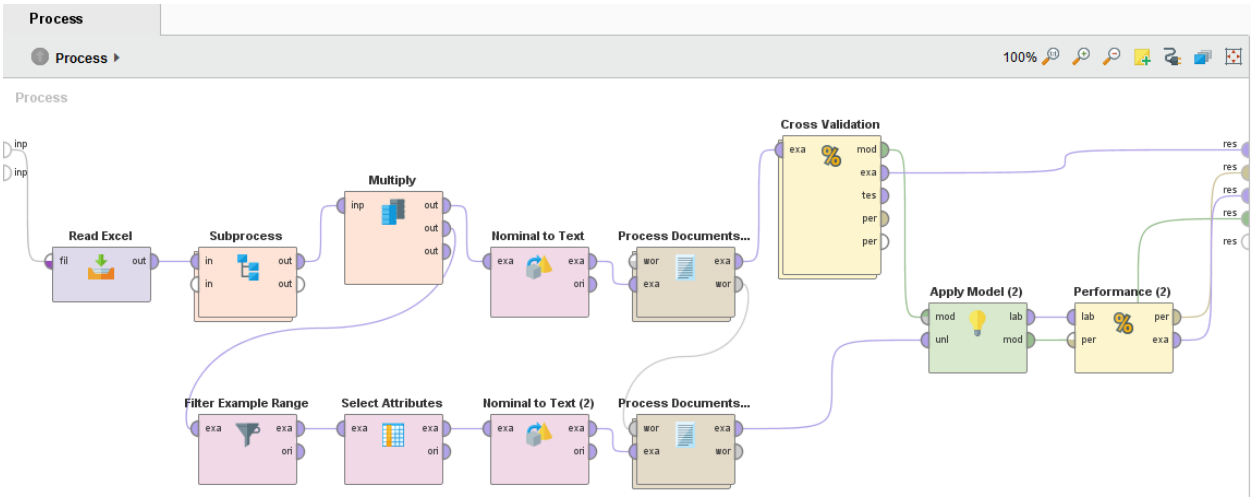
1. **Transform cases**
   Operator ini digunakan untuk mengubah semua karakter menjadi huruf kecil.
2. **Tokenize**
   Operator ini akan memisahkan kalimat menjadi tiap kata atau yang disebut tokenisasi.
3. **Filter Tokens (by length)**
   Operator ini memfilter token tergantung dengan panjangnya. Penulis akan memberikan range atau jarak dari 4 sampai 25. Jadi, operator akan filter setiap kata kecuali kata yang ada di range tersebut.

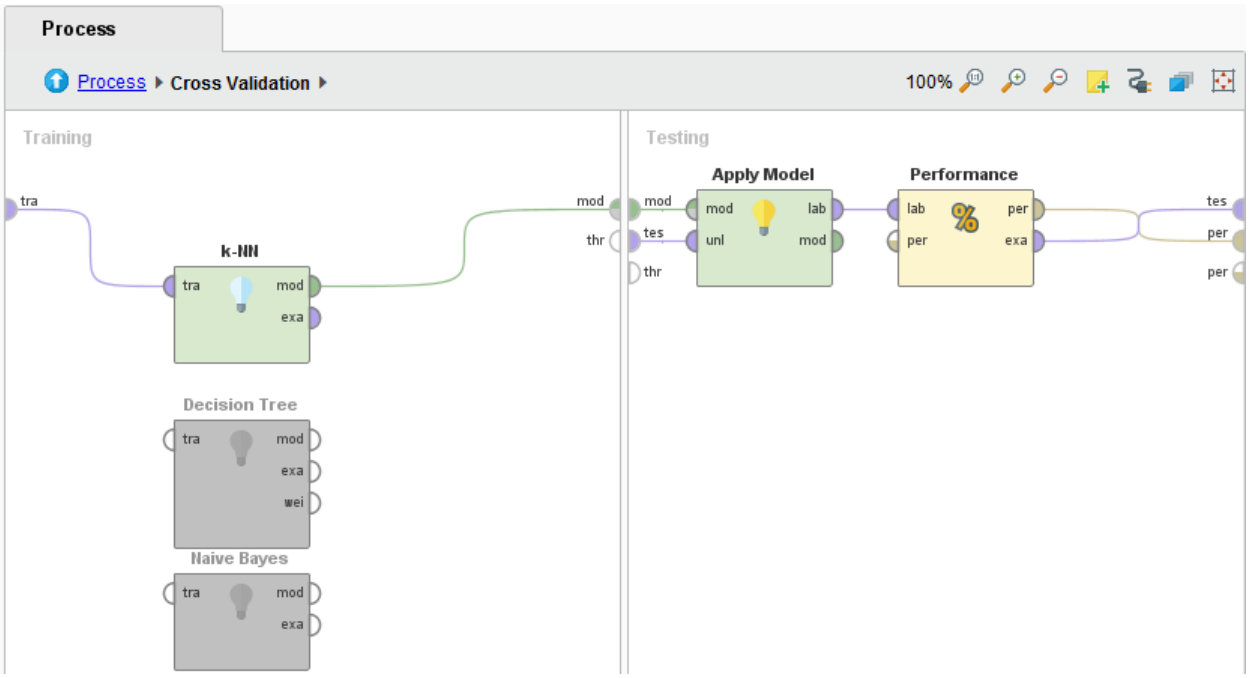
4. **Filter Stopwords (Dictionary)**
   Operator ini akan menghilangkan *stopwords*. Namun untuk mengenali dan bekerja pada bahasa indonesia butuh pengenalan secara manual. Oleh karena itu peneliti memakai 'Filter Stopwords (Dictionary)' dengan menggunakan file "stopword_indo_new.txt" yang katanya sudah ditentukan sebelumnya.
5. **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.

Pemodelan

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 orisinil yang belum dilabeli secara manual untuk membandingkan performa model mana yang terbaik dengan menggunakan data yang berbeda. Setelah mencoba berbagai macam model, model yang paling akurat akan digunakan untuk evaluasi dan penerapan.

Disini peneliti 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 emosi, proses dokumen dari data, dan *dictionary* bahasa indonesia yang ditentukan secara 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. (Mierswa, 2017)





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

PerformanceVector (Performance (2))

ExampleSet (Cross Validation)

Criterion

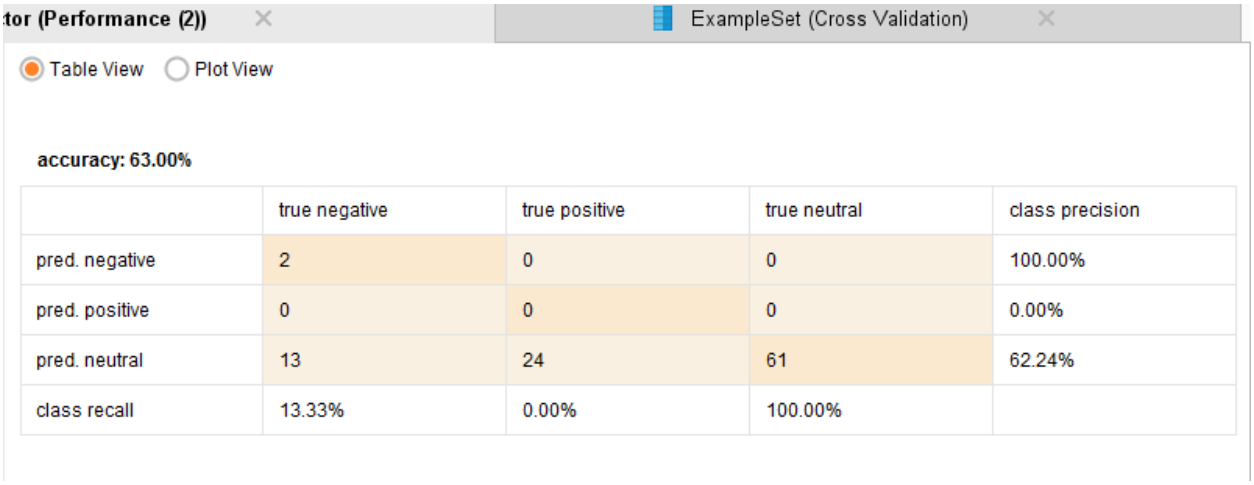
accuracy

Table View Plot View

accuracy: 74.00%

	true negative	true positive	true neutral	class precision
pred. negative	5	1	0	83.33%
pred. positive	3	11	3	64.71%
pred. neutral	7	12	58	75.32%
class recall	33.33%	45.83%	95.08%	

2. Decision Tree



accuracy: 63.00%

	true negative	true positive	true neutral	class precision
pred. negative	2	0	0	100.00%
pred. positive	0	0	0	0.00%
pred. neutral	13	24	61	62.24%
class recall	13.33%	0.00%	100.00%	

3. Naive Bayes

accuracy: 100.00%

	true negative	true positive	true neutral	class precision
pred. negative	15	0	0	100.00%
pred. positive	0	24	0	100.00%
pred. neutral	0	0	61	100.00%
class recall	100.00%	100.00%	100.00%	

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

Evaluasi

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

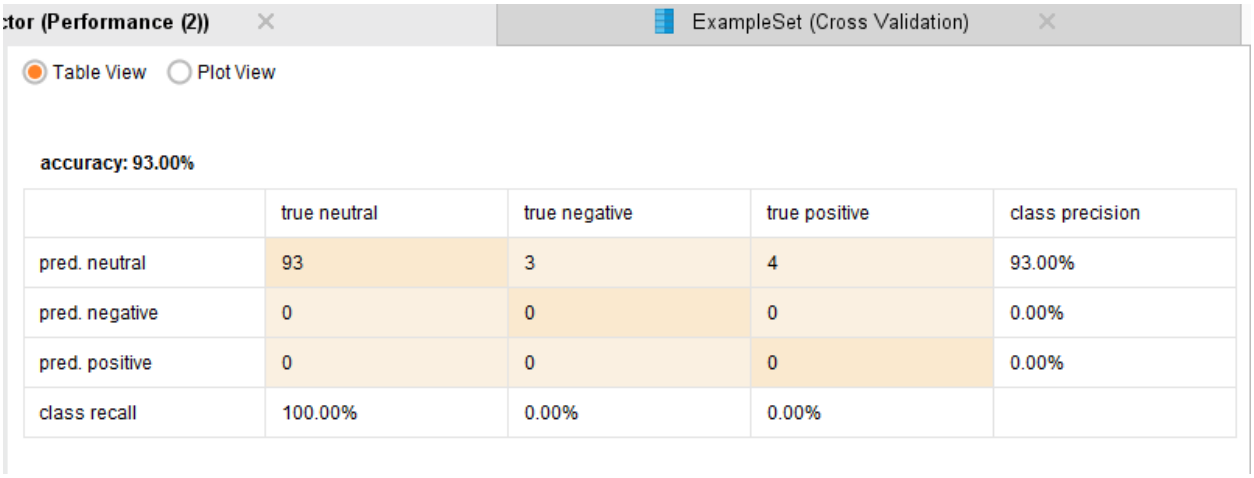
- Evaluasi hasil
- Review proses
- Menentukan langkah selanjutnya

Sekarang telah didapatkan 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* 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 orisinil yang belum dilabeli secara manual atau murni dari *crawling* dan *filtering* proses pencarian *tweets* oleh program rapidminer.

Penerapan

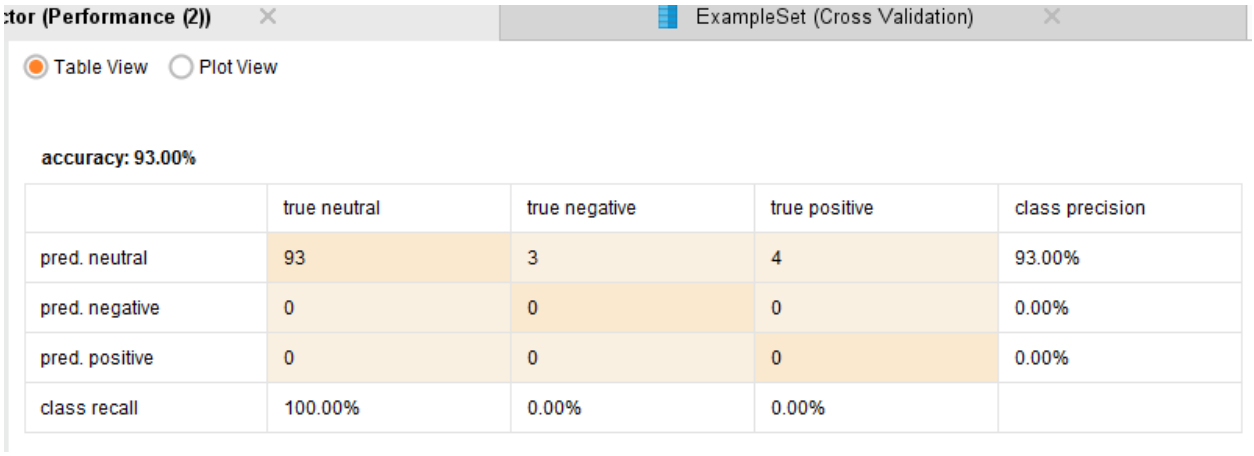
Bagian ini adalah proses akhir dalam *framework* CRISP-DM. Peneliti telah melakukan berbagai macam algoritma model terhadap salah satu dataset. Namun di bagian ini akan dilakukan perbandingan dengan melakukan percobaan dengan dataset yang sama namun belum labeli secara manual. Jadi peneliti akan menerapkan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

1. k-NN



	true neutral	true negative	true positive	class precision
pred. neutral	93	3	4	93.00%
pred. negative	0	0	0	0.00%
pred. positive	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

2. Decision Tree



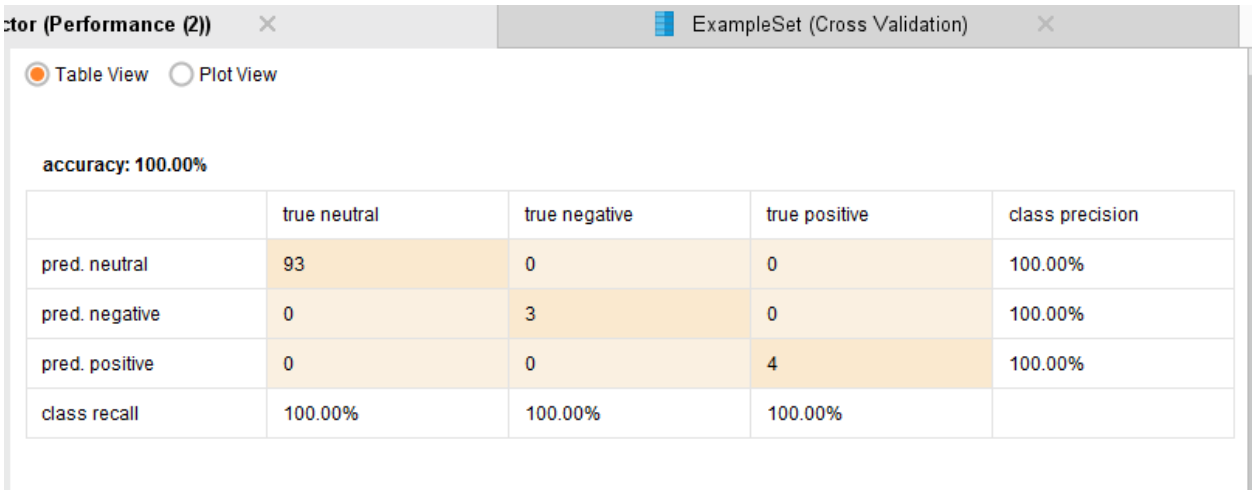
tor (Performance (2)) | ExampleSet (Cross Validation)

Table View | Plot View

accuracy: 93.00%

	true neutral	true negative	true positive	class precision
pred. neutral	93	3	4	93.00%
pred. negative	0	0	0	0.00%
pred. positive	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

3. Naive Bayes



tor (Performance (2)) | ExampleSet (Cross Validation)

Table View | Plot View

accuracy: 100.00%

	true neutral	true negative	true positive	class precision
pred. neutral	93	0	0	100.00%
pred. negative	0	3	0	100.00%
pred. positive	0	0	4	100.00%
class recall	100.00%	100.00%	100.00%	

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

Conclusion

Dapat disimpulkan bahwa sentimen analisis ini dapat diselesaikan dengan *framework* CRISP-DM dalam melakukan penelitian ini. Walaupun level dataset berbeda, yang 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 kustom 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. Tujuan *data mining* adalah mencari nilai data yang paling benar bukan keakuratannya saja.

Referensi

Dey, N., Ashour, A. S., & Nguyen, G. N. (2017). Deep learning for multimedia content analysis. *Mining Multimedia Documents*, *1*(4), 193–203. https://doi.org/10.1201/b21638

Mierswa, I. (2017). *How to Correctly Validate Machine Learning Models*. 26. https://rapidminer.com/resource/correct-model-validation/

Mundalik, A. (2018). *Aspect Based Sentiment Analysis Using Data Mining Techniques Within Irish Airline Industry Aishwarya Mundalik Supervisor :*

Noor, I. M., & Turan, M. (2020). Sentiment Analysis on New Currency in Kenya using Twitter Dataset. *IJID (International Journal on Informatics for Development)*, *8*(2), 81. https://doi.org/10.14421/ijid.2019.08206

Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., & Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76.