Final Project - Text Analytics

Indonesian Tweets Sentiment Analysis

disusun oleh :

Nurul Aisyah - 1606889660
Ricky Chandra Johanes - 1606878013
Winda Wijaya - 1606836704

FAKULTAS ILMU KOMPUTER
PROGRAM STUDI SISTEM INFORMASI

APRIL 2019

DAFTAR ISI

KATA PENGANTAR

Pertama-tama puji syukur kami panjatkan kepada Tuhan Yang Maha Esa, karena atas berkat dan rahmatnya, kami dapat menyelesaikan proyek ini dengan baik dan tepat waktu. Laporan penelitian ini akan membahas tahapan serta eksperimen yang telah dilakukan dalam sentimen analisis terhadap *tweets* yang telah disediakan.

Laporan penelitian ini kami susun untuk memenuhi tugas mata kuliah Analitika Media Sosial. Tentunya, proposal ini tidak dapat diselesaikan dengan baik tanpa adanya bantuan dari berbagai pihak. Untuk itu, pada kesempatan ini kami mengucapkan terima kasih kepada:

1. Bapak Rahmad Mahendra, S.Kom., M.Sc., selaku dosen pengampu mata kuliah Analitika Media Sosial, sekaligus sebagai pembimbing dalam penulisan laporan penelitian ini.
2. Hadi Syah Putra, S.Kom., selaku asisten dosen yang membimbing kami dalam pengerjaan proyek akhir serta penyelesaian laporan penelitian ini.

Laporan penelitian kami masih jauh dari sempurna. Maka dari itu, kami mengharapkan pembaca dapat menyampaikan kritik dan saran yang membangun untuk perbaikan proposal penelitian ini di waktu yang akan datang. Demikian, kami mohon maaf apabila terdapat kesalahan ataupun kata-kata yang kurang sesuai dalam proposal penelitian ini.

Depok, 23 April 2019

Tim Penyusun

ABSTRAK

Perkembangan internet bertumbuh dengan pesat, hal tersebut dapat dilihat dari meningkatnya penggunaan sosial media dalam kehidupan sehari-hari. Salah satu sosial media yang umum digunakan adalah Twitter. Twitter sendiri sudah menjadi tempat umum untuk curhat, menyebarkan informasi, berbagi informasi, dan banyak lagi. Dengan banyaknya *tweets* yang ada per harinya tim peneliti melakukan penelitian terhadap sentimen yang ada pada *tweet-tweet,* melalui eksperimen berbagai model klasifikasi dan fitur.

DESKRIPSI TUGAS

Sebagai proyek akhir mata kuliah Analitika Media Sosial, tim peneliti diminta untuk melakukan sentimen analisis terhadap 8000 tweets bahasa Indonesia yang telah disediakan. tim peneliti diharapkan untuk bereksperimen dalam pengembangan model predictive analytics dan pemilihan fitur yang akan digunakan dalam sentimen analisis. tim peneliti melakukan sentimen analisis dengan mengklasifikan tweets bahasa Indonesia berdasarkan tipe polaritas yang terdiri dari positif (1) dan negatif (0). tim peneliti dibebaskan dalam memilih pendekatan dalam pengerjaan tugas ini.

METODOLOGI PENGERJAAN

❖ Kategori Metodologi

Metodologi yang digunakan dalam pengerjaan tugas ini adalah *experimental research* dan *case study research*. Sehingga, dalam pengerjaan tugas ini selain mencoba berbagai model *predictive analytics* dan mencoba berbagai kombinasi fitur, tim peneliti juga membaca artikel dan penelitian ilmiah terkait untuk membantu pengerjaan tugas. Selain itu, tim peneliti juga menerapkan metodologi *experimental research* dengan membandingkan hasil dari metode algoritma klasifikasi *logistic regression,* SVM, dan *multinomial naive bayes* yang dikombinasikan dengan fitur-fitur seperti, *unigram, POS tagging, sentiment lexicon*, dan lain-lainnya.

❖ Tahapan Pengerjaan

Dalam pengerjaan tugas, tim peneliti memulai dengan berdiskusi bersama antaranggota kelompok terhadap pendekatan yang akan digunakan. Setelah tim peneliti sepakat terhadap pendekatan yang akan digunakan dalam pengerjaan tugas, tim peneliti melakukan konsultasi dengan salah satu asisten dosen dari mata kuliah Analitika Media Sosial. Selanjutnya, tim peneliti melanjutkan pengerjaan dengan mengikuti langkah-langkah pada tutorial *text classification.* Kemudian, pengerjaan dilanjutkan dengan masing-masing anggota mengerjakan dan bereksperimen masing-masing. Setiap hasil yang diperoleh dan metode yang digunakan untuk mendapatkan hasil tersebut pada setiap eksperimen yang dilakukan dibagikan dengan anggota lainnya.

Pada awalnya tim peneliti melakukan sentimen analisis terhadap *tweets* pada *tester set*. Setelah hasil dari prediksi yang didapatkan terhadap *tweets* pada *tester set* sudah sesuai, tim peneliti melanjutkan sentimen analisis terhadap *tweets* pada *test set*. Sehingga dalam melakukan sentimen analisis tersebut, langkah-langkah yang dilakukan berupa,

1. ***Pre-processing* terhadap *tweet***

   Adapun kegiatan *pre-processing* yang dilakukan sebagai berikut

   a. Normalisasi

      Normalisasi dilakukan agar kata-kata yang terdapat pada text menjadi konsisten, misalnya yg, ygg, yyygggg menjadi yang. Normalisasi *tweet* yang dilakukan terdiri dari pengubahan ke huruf kecil, pembuangan spasi yang berlebihan, *trimming*, pembuangan tanda baca, penghilangan huruf berulang. Fungsi *normalisasi (tweet)* menerima masukan berupa *tweet* awal atau mentah yang bertipe String.

   b. Penghapusan *stopwords* dan istilah spesial, seperti (*username, hyperlink,* dan lainnya)

      Stopwords adalah kata-kata yang umum digunakan sehingga memiliki frekuensi yang tinggi dan pada umumnya dibuang sebelum atau setelah *text processing*. Penghapusan *stopwords* dilakukan agar kata - kata yang tidak mengandung informasi lebih tidak mengambil ruang

(*space*) di database, atau membuat waktu *processing* menjadi bertambah. Oleh karena itu, tim peneliti membuang kata - kata tersebut dengan mendefinisikan daftar kata - kata yang dianggap sebagai *stopwords* dan dicatat pada sebuah dokumen yang bernama *stopwords.csv*. Lalu, kita menggunakan *library* pada python (Natural Language Toolkit) untuk membersihkan *tweet* dari *stopwords* yang telah didefinisikan. Adapun fungsi *remove_stopwords (tweet)* menerima masukan berupa *tweet* yang telah dilakukan normalisasi pada poin ***[a]***.

c. *Stemming*
Stemming merupakan proses pengambilan kata dasar dari kata pada *text* yang akan diproses. Pada pengerjaan tugas ini, tim peneliti melakukan dua percobaan, yaitu dengan melakukan *stemming* dan tanpa melakukan *stemming. Stemmer* yang digunakan untuk *stemming* dihasilkan dengan menggunakan Stemmer Factory. Fungsi *stemming (tweet)* menerima masukan berupa *tweet* bertipe String yang telah dilakukan penghapusan *stopwords* pada poin ***[b]***.

2. Ekstraksi Fitur

Terdapat beberapa fitur yang digunakan oleh tim peneliti dalam melakukan sentimen analisis, yaitu

1. *Bag-of-Words*
Fitur *bag-of-words* atau Unigram merupakan ekstraksi fitur yang merepresentasikan apakah sebuah kata muncul pada suatu text sebagai matriks. Fitur Unigram dibentuk menggunakan *library* Count Vectorizer dari Scikit-learn.

2. Leksikon (Kamus Sentimen & Kamus Koto)
Pada fitur ini tim peneliti menggunakan dua kamus, yaitu *leksikon* dari penelitian Vania dan *leksikon* dari penelitian Koto. *Leksikon* tersebut berupa kamus berisi kata-kata positif atau negatif. Pada setiap kamus terdiri dari dua *file*, satu *file* untuk kata-kata bersentimen positif dan satunya lagi untuk yang bersentimen negatif.

3. Part-of-Speech Tagging (POS Tagging)
Part-of-speech (POS) merupakan fitur dengan cara menetapkan setiap token kepada *part of speech*, yaitu kata benda, kata sifat, dan lainnya. kelas kata *part-of-speech* dapat digunakan untuk membantu mengetahui sentimen dari sebuah *tweets*. Pada percobaan ini, akan dihitung kemunculan kata sifat dan kata negasi berdasarkan pre-trained POS Tag dari penelitian Dinakarami et. al. yang sudah dikonversi ke bentuk CRF.Tagger agar bisa dibaca dar NLTK.

4. Ortografi

Fitur ortografi menggunakan huruf kapital serta tanda baca untuk mengetahui intonasi dari sebuah *tweets*. Sehingga, set data yang digunakan untuk ekstraksi fitur ini merupakan *data set* yang belum melalui tahap *pre-processing*. Fitur ortografi dapat membantu tim peneliti dalam mengetahui emosi dari sebuah *tweet*.

5. TF-IDF
   Fitur tf-idf atau *term frequency–inverse document frequency* merupakan matriks yang merepresentasikan seberapa pentingnya suatu kata pada set data teks. Fitur tersebut dibentuk menggunakan *library* Tfidf Vectorizer dari Scikit-learn.

3. Pengukuran Nilai F1

Setelah melakukan ekstraksi terhadap fitur yang mungkin akan digunakan, tim peneliti mengukur nilai F1 terlebih dahulu dari model-model dengan fitur-fitur yang akan digunakan. Dari pengukuran F1 pada beberapa fitur didapatkan nilai sebagai berikut,

```
LOGISTIC REGRESSION
('Jenis Fitur : ', 'Unigram')
('F1-Score :', 0.8476446145265975)
---------------
DECISION TREE
('Jenis Fitur : ', 'Unigram')
('F1-Score :', 0.7685186897529117)
---------------
LOGISTIC REGRESSION
('Jenis Fitur : ', 'Sentimen')
('F1-Score :', 0.7916838705373135)
---------------
DECISION TREE
('Jenis Fitur : ', 'Sentimen')
('F1-Score :', 0.7890882123922349)
---------------
LOGISTIC REGRESSION
('Jenis Fitur : ', 'POS')
('F1-Score :', 0.5170408403335769)
---------------
DECISION TREE
('Jenis Fitur : ', 'POS')
('F1-Score :', 0.5065480145412494)
---------------
LOGISTIC REGRESSION
('Jenis Fitur : ', 'Ortografi')
('F1-Score :', 0.5487824096670157)
---------------
DECISION TREE
('Jenis Fitur : ', 'Ortografi')
('F1-Score :', 0.5311882729703921)
---------------
LOGISTIC REGRESSION
('Jenis Fitur : ', 'Sentimen Koto')
('F1-Score :', 0.6885957118208752)
---------------
DECISION TREE
('Jenis Fitur : ', 'Sentimen Koto')
```

```
LOGISTIC REGRESSION
('Jenis Fitur : ', 'Unigram')
('F1-Score :', 0.8476446145265975)
---------------
DECISION TREE
('Jenis Fitur : ', 'Unigram')
('F1-Score :', 0.7644667236296823)
---------------
MULTINOMIAL NAIVE BAYES
('Jenis Fitur : ', 'Unigram')
('F1-Score :', 0.8465681807120788)
---------------
SUPER VECTOR MACHINE
('Jenis Fitur : ', 'Unigram')
('F1-Score :', 0.8347139289951933)
---------------
LOGISTIC REGRESSION
('Jenis Fitur : ', 'TFIDF')
('F1-Score :', 0.8485466308195153)
---------------
DECISION TREE
('Jenis Fitur : ', 'TFIDF')
('F1-Score :', 0.768917334899798)
---------------
MULTINOMIAL NAIVE BAYES
('Jenis Fitur : ', 'TFIDF')
('F1-Score :', 0.8658886074399833)
---------------
SUPER VECTOR MACHINE
```

4. Eksperimen dengan Kombinasi Model dan Fitur

Setelah mengetahui nilai F1 dari kombinasi fitur dan beberapa model klasifikasi. Selanjutnya tim peneliti melakukan percobaan dengan berbagai

kombinasi model dengan fitur. Pemilihan model dan fitur yang akan dicoba berdasarkan hasil pengukuran f1-nya. Selain itu, jika akurasi yang didapatkan sudah cukup tinggi, tim peneliti melakukan *hyper parameter tuning* dengan mengubah parameter pada model, seperti *max feature* pada Vectorizer.

PENDEKATAN YANG DIGUNAKAN

Pendekatan yang kami gunakan dalam analisis sentimen *tweet* disini adalah pendekatan *supervised.* Dimana kami memiliki data *training* sebanyak 3463 row yang berisi text dengan label sentimen text. Data training tersebut akan di-*training* menggunakan beberapa model untuk memprediksi data *test_set* yang terdiri dari 8000 *tweets*. Selain itu, tim peneliti juga memilih untuk menggunakan pendekatan *machine learning* dibandingkan dengan *rule based*. Sehingga, dalam memprediksi, tim peneliti menggunakan model-model klasifikasi untuk mengklasifikasi *tweets* pada *test_set* bersentimen negatif atau positif berdasarkan *train* dari *train_set*.

PEMBAHASAN

❖ Eksperimen dan Hasil

Adapun eksperimen yang dilakukan dalam pengerjaan proyek ini adalah dengan mencoba mengimplementasikan algoritma *Machine Learning* guna memprediksi sentimen dari setiap *tweet* yang ada pada *tester set* (10 *tweets*) dan *test set* (8000 *tweets*) yang diberikan. Adapun algoritma - algoritma yang digunakan antara lain *Multinomial Naive Bayes, Logistic Regression,* dan *Super Vector Machine.*

Eksperimen pada Tester Set :

Adapun tahapan - tahapan yang dilakukan pada eksperimen yang dilakukan menggunakan *tester set* adalah sebagai berikut :

1. ***Logistic Regression* dengan fitur *Sentiment Lexicon***
   Pada eksperimen pertama, kami menggunakan fitur *sentiment lexicon* dan mengklasifikasi dengan menerapkan model *logistic regression.* Kami melakukan percobaan ini paling awal, karena sesuai dengan tutorial pada *text analysis*, serta hasil pengukuran f1 yang didapatkan merupakan nilai f1 kedua tertinggi tepat dibawah fitur *unigram* dengan nilai 0,79 dari 1. Dari eksperimen tersebut kami mendapatkan akurasi dari sentimen analisis terhadap *tester set* sebesar 80%.

2. ***Logistic Regression* dengan fitur *Bag-of-Words* atau *Unigram***
   Pada eksperimen kedua, kami menggunakan fitur unigram dengan menerapkan model *logistic regression.* Kami memilih fitur unigram untuk eksperimen selanjutnya karena hasil pengukuran F1 yang tertinggi. Dengan menggunakan fitur dan model tersebut hasil pengukuran F1 yang didapatkan sebesar 0.846 dari 1. Hasil akurasi yang dihasilkan pada data tester didapatkan sebesar 100% namun kami tidak berhasil melakukan submit karena sudah melewati deadline.

Eksperimen pada Test Set :

Adapun pendekatan - pendekatan yang digunakan pada eksperimen yang dilakukan dengan menggunakan *test set* adalah sebagai berikut.

1. **Logistic Regression**
   Eksperimen pertama kali yang kami lakukan adalah *Logistic Regression* dengan menggunakan fitur *Bag-of-Words* atau *Unigram*. Hal ini kami lakukan karena pendekatan ini berhasil mendapat akurasi 100% saat diterapkan pada *tester set*. Akurasi yang didapatkan dari eksperimen ini adalah 85,60% (*max_features* = 2000)..

Setelah dianalisis lebih lanjut, tim peneliti memutuskan bahwa sepertinya akan mengalami peningkatan akurasi jika jumlah kata yang unik atau tingkat intensitasnya kecil diambil lebih banyak sehingga tim peneliti berupaya dengan menambah parameter *max_features* nya menjadi 100000. Akan tetapi, hal tersebut menyebabkan terjadinya penurunan akurasi. Tim peneliti mengasumsikan bahwa penurunan tersebut disebabkan oleh kata - kata unik yang ada terlalu unik atau intensitasnya terlalu kecil sehingga menyebabkan prediksi semakin sukar dilakukan karena kemungkinannya menjadi banyak. Adapun hal tersebut direpresentasikan pada tabel dibawah ini.

No.	Jumlah Kata Unik / *Max Features*	Akurasi
1.	2000	85.60%
2.	20000	86.05%
3.	100000	85.68%

Keterangan : Angka - angka yang ada pada tabel tersebut merupakan hasil percobaan yang telah dilakukan.

Adapun poin yang ingin disampaikan adalah kisaran jumlah kata unik atau *max_features* memiliki titik puncak sehingga akurasi akan kembali menurun jika *max_features* yang diterapkan berjumlah diatas titik puncak tersebut. Berdasarkan hasil penelitian yang kami lakukan, *max_features* terbaik berada di antara 6600 sampai dengan 20000. Akurasi yang didapatkan dari eksperimen ini (dengan jumlah *max_features* teroptimal, yaitu 20000) adalah 86.05% (dengan unigram dan *max_features = 20000*).

Lalu, peneliti kembali menganalisis dalam mengambil keputusan berikutnya untuk meningkatkan akurasi berdasarkan komparasi kedua teknik ekstraksi fitur yang telah dilakukan. Pada akhirnya, kita berpikir untuk menerapkan pendekatan *Logistic Regression* hanya pada kata - kata yang terpenting saja (*Term Frequency, Inverse Document Frequency*) agar proses perhitungan akan menjadi lebih cepat dan asumsinya perhitungan akurasi akan lebih meningkat jika kita memprediksi sentimen setiap *tweet* berdasarkan kata - kata terpenting yang ada pada *train set*. Ternyata, asumsi tersebut salah akurasi yang didapatkan dari eksperimen tersebut adalah sama dengan eksperimen sebelumnya, yaitu 86.05% (dengan *tfidf* dan *max_features = 20000*).

2. **Multinomial Naive Bayes**
   Sebagai eksperimen kedua, tim peneliti memilih untuk menggunakan model Multinomial Naive Bayes. Berdasarkan eksperimen sebelumnya, kami melanjutkan penggunaan fitur *tf-idf.* Eksperimen ini dilakukan dengan menggunakan fitur leksikon Vania dan leksikon Koto dan *tf-idf* terhadap *tweets* pada *test set* tanpa melakukan *pre-processing.* Akurasi yang didapatkan dari eksperimen ini adalah 86.16.

3. **Super Vector Machine**
   Sebagai eksperimen ketiga, tim peneliti memilih untuk menggunakan model Super Vector Machine. Berdasarkan eksperimen sebelumnya, kami melanjutkan penggunaan fitur *tf-idf.* Eksperimen ini dilakukan dengan menggunakan fitur *tf-idf* terhadap *tweets* pada *test set* tanpa melakukan *pre-processing.* Akurasi yang didapatkan dari eksperimen ini adalah 87.16. Kemudian kami melakukan eksperimen ini pada tweets dengan melakukan *pre-processing* dan didapatkan hasil akurasi sebesar 87.40.

ERROR ANALYSIS

1. Logistic Regression

Pendekatan *Logistic Regression* memiliki beberapa kelemahan, yaitu tidak bisa melacak *tweet* yang memiliki sentimen tertentu baik positif ataupun negatif, tetapi sentimen tersebut akan terbentuk jika dan hanya jika kalimat dibaca secara utuh. Misalkan, pada *gold standard label,* kita dapat melihat beberapa contoh, yaitu sebagai berikut.

1. Ku inget Ayahku yang berpulang hampir 3 bulan lalu. kalo ada kesempatan sekali lagi rasanya aku gamau ada yg waktu yg kebuang brsama ayah.

   Pada *tweet* tersebut, pendekatan Logistic Regression kesulitan untuk menilai sentimennya negatif atau positif jika hanya melihat salah satu kata atau beberapa kata saja. Hampir tidak ada kata yang menjurus ke arah sifat negatif atau perasaan negatif pada *tweet* sejenis ini.

2. Definitely Caca, Kak Ika! Auranya udahlah paling pas. Kualitas aktingnya ga perlu dipertanyakan. Ditambah gatau Caca ini ada apanya, tapi akan gemes aja kalo jadi Alex yang galak tapi penyayang tapi cantik tapi sibuk tapi masih sanggup menghadapinya

   Pada *tweet* tersebut, kita melihat banyak sekali kombinasi kata yang memiliki sentimen yang berbeda. Misalkan, terdapat kata cantik, namun juga terdapat kata gemes ataupun sibuk yang dapat berdampak negatif.

3. kontradiksi dengan freport yang jelas2 merusak alam :d

   Pada *tweet* tersebut, kita melihat bahwa terdapat *emoticon* yang jika dilihat secara tunggal, kita bisa saja menganggap bahwa sentimennya bersifat positif atau negatif, ternyata malah sebaliknya karena *emoticon* tersebut bersifat sarkasme.

4. Semilir angin sore, tidak ada deru berisik kota, hanya ada ruang di antara senduro, di sudut Surya Kencana. . #CatatanPerjalanan #SuryaKencana #GunungGede [URL]

   Pada *tweet* tersebut, mesin kesulitan untuk mengetahui bahwa kondisi yang digambarkan bersifat positif. Hal - hal ini membutuhkan kemampuan akal pikiran manusia yang tahu bahwa adanya angin sore dan tidak ada deru berisik merupakan suatu hal yang positif. Dengan kata lain, pada *tweet* seperti ini prediksi menggunakan pendekatan *Logistic Regression* bisa saja benar ataupun salah bergantung pada jumlah katanya dan sifat - sifat atau sentimen - sentimen yang bisa saja terbentuk dari sejumlah kata atau suatu kata tunggal. Misalnya, jika ada kata “Awalnya, kita sedih karena satu dan lain hal. Lalu, ada angin begitu sejuk dan tidak ada berisik kota”

mungkin akan bersentimen negatif karena terdapat kata sedih, padahal hal tersebut belum tentu berakhir sedih karena terdapat kata “Awalnya”.

2. Multinomial Naive Bayes

Hasil prediksi dari dengan model *multinomial naive bayes* menghasilkan akurasi 86.16, yaitu lebih tinggi dari penggunaan *logistic regression*. Selain itu, jika diperhatikan pada *gold standard* terdapat beberapa jenis *tweet* yang gagal diprediksi, seperti

1. *tweet* yang pada umumnya gagal terprediksi dengan tepat merupakan *tweet* yang bersifat sarkasme, sebagai contoh

   “*Jadi wanita jangan suka menghancurkan hubungan orang. Jgn bangga berhasil merusak kebahagian orang. Silahkan saja, tapi ga berkah bahagianya nanti hehe*.”.

   Tanpa diartikan kalimat tersebut bersentimen positif dengan adanya kalimat-kalimat yang positif, namun sebenarnya kalimat tersebut bermaksud untuk menyindir atau menyinggung seseorang.
2. *Tweet* lain yang gagal diprediksi adalah *tweet* dengan kalimat kontradiktif, dimana pada awal kalimat bersentimen negatif, namun pada akhirnya bersentimen positif, begitupun sebaliknya. Salah satu contohnya adalah

   “*Di SD-SMP gue dulu kalau ada yg ketahuan bawa artikel terkait Pornografi, dijejerin depan kelas. Ditendangin tulang keringnya satu satu oleh wali kelas. Gue ngga pernah (ketahuan). . . [URL]”.*

3. Super Vector Machine

Pendekatan dengan Super Vector Machine sudah dapat memprediksi sentimen dengan lebih baik, beberapa *tweet* yang gagal diprediksi oleh *multinomial naive bayes* berhasil diprediksi dengan tepat. Namun, tetap ada beberapa *tweet* yang masih gagal diprediksi sentimennya. Beberapa *tweet* yang tidak berhasil diprediksi adalah sebagai berikut,

1. Terdapat beberapa *tweet* yang akan susah diprediksi sentimennya apabila dilihat kata per kata, sebagai contoh

   “*Ku inget Ayahku yang berpulang hampir 3bulan lalu. kalo ada kesempatan sekali lagi rasanya aku gamau ada yg waktu yg kebuang brsama ayah.”*

   Tweet tersebut gagal diprediksi dengan *logistic regression* juga.
2. *Tweet* dengan kalimat ekspresi ataupun *emoticon* yang tidak sesuai dengan sentimen, sebagai contoh

   “*haha jangan marah terus tha nanti cepat tua sekarang yang kudu kamu lakuin ikhlasin saja kalau bojong tidak-ada ningtyas”.*

   Dapat dilihat *tweet* tersebut menggunakan kata ‘haha’ yang seharusnya menandakan sesuatu yang lucu atau positif, tetapi tanpa ‘haha’ dapat dilihat bahwa *tweet* ini bersentimen negatif.

KESIMPULAN DAN SARAN

Dari banyaknya percobaan yang dilakukan, pada akhirnya hasil akurasi tertinggi yang dapat diperoleh oleh tim peneliti adalah dengan menggunakan SVM dengan akurasi 87,40. Dari percobaan-percobaan yang dilakukan tersebut, tim peneliti juga mendapatkan bawa *tweets* yang dilakukan *pre-processing* belum tentu menghasilkan akurasi yang lebih tinggi dibandingkan yang tanpa *pre-processing,* seperti bagaimana eksperimen dengan *stemming* menghasilkan akurasi yang lebih rendah. Selain itu, parameter *max features* yang banyak belum tentu berdampak baik, karena prediksi akan *over fit*, dimana untuk mencapai sentimen 1 atau 0 akan sulit. *Tweet-tweet* yang mengandung sarkasme akan mempersulit pembelajaran mesin untuk memprediksi sentimen, karena pembelajaran mesin yang dilakukan dengan memperhatikan setiap kata pada *tweet* tersebut dan tidak arti *tweet* secara keseluruhan.

Pada penelitian kedepannya, saat akan dilakukan *pre-processing* dapat dipikirkan terlebih dahulu kegiatan *pre-processing* apa saja yang diperlukan. Selain itu, pada saat melakukan *hyper parameter tuning*, peneliti dapat melakukan beberapa percobaan, karena belum tentu semakin besar akan semakin lebih baik.

DAFTAR PUSTAKA

Clara Vania, Moh. Ibrahim, and Mirna Adriani. Sentiment Lexicon Generation for an Under-Resourced Language. CICLING 2014 (IJCLA)

Fajri Koto, and Gemala Y. Rahmaningtyas "InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs". IEEE in the 21st International Conference on Asian Language Processing (IALP), Singapore, 2017.

LAMPIRAN

❖ **Source Code**

Terlampir pada *file* yang terpisah dan dapat dilihat pada https://github.com/kaelky/TextAnalytics_RaceNWin

❖ **Model**