ANTUSIASME WARGANET INDONESIA TERHADAP MARVEL CINEMATIC UNIVERSE PADA TAHUN 2018

Budi Gunawan, 1608107010009

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Syiah Kuala
Email: budi99@mhs.unsyiah.ac.id

ABSTRAK

Media sosial adalah suatu media yang menghubungkan orang-orang dari segala penjuru dunia sekaligus sebagai wadah tempat mengutarakan opini dan informasi. Salah satunya adalah Twitter yang banyak digunakan saat ini di Indonesia. Cuitan masyarakat pengguna Twitter dapat digunakan sebagai acuan kepuasan terhadap suatu produk di mana dalam penelitian ini akan dianalisis penyebab euforia masyarakat Indonesia, apa yang disukai pasar dan apa yang tidak terhadap produksi film Marvel ini sehingga akan berguna untuk evaluasi oleh pihak Marvel ke depan. Untuk mengetahui hal tersebut, maka dicari polaritas setiap *tweet* agar dapat diklasifikasikan ke dalam positif atau negatif. *Tweet* akan melalui tahap *preprocessing* terlebih dahulu untuk dilakukan *cleaning*, *normalisasi* dan penghilangan *stopword*. Klasifikasi dilakukan setelah mendapatkan label positif dan negatif dengan tujuan untuk memodelkan data dengan menggunakan metode *Random Forest Classifier* dan ekstraksi fitur meggunakan metode *TF-IDF Vectorizer*. Hasil penelitian diperoleh informasi bahwa terdapat 15.067 *tweet* positif tentang marvel, 3.735 *tweet* negatif tentang marvel, dan sisanya 11.406 netral dari total 30.191 *tweet* dengan akurasi sebesar 69% serta penyebab dari beragamnya reaksi masyarakat terhadap Marvel yang akan diuraikan dalam penelitian.

Kata kunci: twitter, tweet, marvel, polaritas, preprocessing, cleaning, normalisasi, stopword, Random Forest Classifier, TF-IDF Vectorizer

ABSTRACT

Social media is a media that connects people from all corners of the world as well as a place for expressing opinions and information. One of them is Twitter, which is widely used today in Indonesia. Tweet from Twitter user can be used as a reference to the satisfaction of a product where in this study will be analyzed the causes of euphoria of the Indonesian people, what the market likes and does not like about the production of this Marvel movie so that it will be useful for evaluation by Marvel in the future. To find this out, the polarity of each tweet is defined first so that it can be classified into positive or negative. Tweets will go through the preprocessing stage first to do cleaning, normalization and removal of stopwords. Classification is done after getting positive and negative labels with the aim to model the data using the Random Forest Clasifier method and feature extraction using the TF-IDF Vectorizer method. From the results, we obtained information that there are 15,067 positive tweets about the marvel, 3,735 negative tweets about the marvel, and the remaining 11,406 neutral out of a total of 30,191 tweets with an accuracy of 69% and the causes of the diverse community reactions to Marvel which will be explained in this study.

Keywords: twitter, tweet, marvel, polarity, preprocessing, cleaning, normalization, stopword, Random Forest Classifier, TF-IDF Vectorizer

1. Pendahuluan

Penggunaan media sosial seperti Facebook, Twitter, Pinterest, dan Youtube pada era abad informasi ini mengalami kenaikan yang sangat drastis dibandingkan dengan tahun - tahun sebelumnya. Data dari tahun 2015 saja memaparkan pengguna Twitter ada lebih dari 285 juta (Alim, 2015). Berdasarkan penelitian Semiocast, lembaga riset media sosial yang berpusat di Paris, Prancis, mengatakan bahwa jumlah pemilik akun Twitter di Indonesia merupakan yang terbesar kelima di dunia, dan berada pada posisi ketiga negara yang paling aktif mengirim *Tweet* perhari (Asih, 2012).

Tingginya pengguna Twitter menjadi peluang untuk masyarakat dalam melakukan jual beli, menyampaikan informasi, promosi, atau bahkan untuk mengutarakan perasaan dan opini,

termasuk juga dalam mengutarakan opini film. Opini tentang film dapat digunakan sebagai sarana untuk memberikan penilaian terhadap film yang diproduksi untuk masyarakat. Hal tersebut dapat dimanfaatkan oleh produser film untuk mengetahui bagaimana tanggapan masyarakat mengenai film yang telah diproduksi, begitu juga dengan masyarakat dapat meninjau film yang akan ditonton. Di sinilah analisis sentimen memiliki peran yang sangat penting. Analisis sentimen merupakan daerah penelitian perhitungan untuk mengekstraksi polaritas pendapat antar kelas (positif dan negatif) dari dokumen teks (Perdana & Pinandito, 2017). Meskipun begitu sebuah analisis teks tidak harus dengan media sosial tertentu, pemilihan pada media sosial Twitter, karena kelebihan Twitter memakai karakter dibawah 140, format pesan yang informal, teks pesan yang sangat banyak setiap harinya, berasal dari individu dengan latar belakang yang sangat bervariasi, dan mudah digunakan (Nahar & Lee, 2008).

Penggunaan analisis sentimen dapat diterapkan pada opini film pada dokumen Twitter berbahasa Indonesia. Pada penulisan opini film terkadang terdapat penulisan *Tweet* yang sulit dibaca. Hal ini disebabkan oleh beberapa faktor seperti penulisan kata yang disingkat, penggunaan bahasa modern atau *slang,* salah dalam mengetik huruf dan tidak baku dalam penulisan opini (Agarwal, et al., 2014). Maka itulah diperlukan beberapa tahapan untuk mengkonversi *tweet* agar mudah dikenali sebelum menentukan sentimen pada *tweet* tersebut. Teks twitter yang akan dianalisis haruslah benar-benar bersih agar dapat dideteksi dan tidak menjadi bias ketika dihitung polaritasnya. Setelah mendapatkan polaritasnya, barulah dapat dilakukan pelabelan sentimen apakah positif, netral ataupun negatif.

Untuk menguji data yang telah diolah tersebut maka dilakukan modeling, di mana data *tweet* yang telah dibersihkan diekstrak fiturnya untuk mendapatkan *score* dan dilatih dengan label yang telah didapat sebelumnya menggunakan salah satu metode klasifikasi.

Beberapa penelitian berkaitan dengan topik analisis sentimen dan klasifikasi sudah banyak dilakukan, di antaranya : penelitian *An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis* (Wan, 2015) menggunakan enam metode untuk klasifikasi yaitu Lexicon-based classifier, NB, Bayesian Network, SVM (Support Vector Machine), C4.5 (Decision Tree), Random Forest serta satu metode yang disebut dengan Ensemble Classifier yang menggabungkan lima metode (NB, Bayesian Network, SVM, C4.5 dan Random Forest) untuk mendapatkan akurasi yang lebih tinggi. Penelitian ini menggunakan empat kelas yaitu kelas positif (4288 *tweet*), negatif (35876 *tweet*), netral (40987 *tweet*) dan *irrelevant* (26715 *tweet*). Perolehan akurasi masing-masing saat tidak dikombinasikan dengan dataset dua kelas (menghilangkan kelas netral dan irrelevant) adalah Lexicon Based 67.9%, Naïve Bayesian 90%, Bayesian Network 91.4%, SVM 84.6%, Random Forest 89.8%. Metode Lexicon Based tidak ikut dalam kombinasi karena perolehan akurasinya paling sedikit yaitu 67,9%, perolehan akurasi ensemble dengan dataset dua kelas yaitu 91.7% sedangkan perolehan akurasi ensemble untuk dataset tiga kelas yaitu 84.2%. Berdasarkan penelitian tersebut, peneliti memutuskan untuk menggunakan salah satu metode klasifikasi yaitu metode Random Forest. Metode Random Forest adalah salah satu metode klasifikasi dalam *Decision Tree* dengan tingkat akurasi yang baik terutama untuk data dengan jumlah yang besar.

Studi ini mengambil topik utama yaitu marvel. Marvel sendiri merujuk kepada perusahaan yang bergerak di bidang *entertainment* yang berbasis superhero. Marvel telah menerbitkan komik dengan berbagai karakter yang sangat populer seperti *Spider-Man*, *X-Men*, *Hulk*, *The Fantastic Four*, *Iron Man*, dan masih banyak lagi. Sebagian besar karakter ciptaan *Marvel* beroperasi dalam dunia yang dikenal sebagai *Dunia Marvel*. Perkembangannya kemudian, banyak dari karakter Marvel tersebut yang muncul dalam media hiburan lain seperti serial kartun, film televisi, layar lebar, dan permainan video.

Marvel juga memiliki situs wikinya sendiri. Situs tersebut diluncurkan pada tahun 2006 dan memuat berbagai informasi dalam jagad Marvel. Pada tahun 2009, The Walt Disney Company menyatakan sepakat untuk membeli Marvel Entertainment sebesar USD 4 miliar dalam transaksi saham dan uang tunai. Dengan demikian, Walt Disney berhak atas karakter komik *superhero* atau karakter pahlawan berkekuatan super seperti *Spider-Man*, *Iron Man* dan *X-Men*. Kesepakatan tersebut akan memberi Disney kepemilikan lebih dari 5.000 karakter tokoh *Marvel Entertainment* (Setyanto & Adiwibawa, 2018).

Dalam penelitian ini, akan dilakukan analisis polaritas *tweet* yang membahas tentang Marvel yang dari hasil ini selain mendapatkan

akurasi dan model data diharapkan dapat memberikan informasi mengenai persebaran sentimen untuk cuitan tentang Marvel, tanggal di mana frekuensi cuitan tertinggi dan alasannya, penyebab munculnya cuitan positif dan negatif, *hashtag* apa yang paling populer dan cuitan positif atau negatif apa yang memiliki frekuensi tertinggi.

2. Literature Survey

2.1. *Pre-processing*

Tahap *text preprocessing* adalah tahap awal dari *text mining*. Tahap ini mencakup semua rutinitas, dan proses untuk mempersiapkan data yang akan digunakan pada operasi *knowledge discovery* sistem *text mining* (Feldman & Sanger, 2007). Tindakan yang dilakukan pada tahap ini adalah *toLowerCase*, yaitu mengubah semua karakter huruf menjadi huruf kecil dan *Tokenizing* yaitu proses penguraian deskripsi yang semula berupa kalimat-kalimat menjadi kata-kata dan menghilangkan delimiter-delimiter seperti tanda titik (.), koma (,), spasi dan karakter angka yang ada pada kata tersebut (Weiss et al, 2005).

Tahapan *pre-processing* pada penelitian ini meliputi proses *tokenizing, case folding, cleaning,* perbaikan kata tidak baku, dan *filtering stopwords.*

2.2. *Sentiment Analysis*

Sentiment Analysis (SA) atau *Opinion Mining* (OM) adalah studi komputasi atas pendapat, sikap, dan emosi orang terhadap suatu entitas. Entitas dapat mewakili individu, acara, atau topik (Medhat et al, 2014). *Opinion mining* bertugas untuk mengekstraksi dan menganalisis pendapat orang tentang suatu entitas, sementara *Sentiment Analysis* adalah untuk mengidentifikasi sentimen yang diungkapkan dalam suatu teks lalu menganalisisnya. Oleh karena itu, tujuan utama dari *Sentiment Analysis* adalah untuk menemukan pendapat, mengidentifikasi sentimen yang diungkapkan, dan kemudian mengklasifikasikan polaritasnya (positif, negatif, ataupun netral).

2.3. *Feature Extraction*

Ekstraksi fitur bertujuan untuk mendapatkan nilai atau skor dari sebuah teks. Para peneliti mengekstraksi fitur dengan melakukan pembobotan kata-kata. Kata-kata dalam dokumen diubah menjadi bobotnya. Pembobotan kata-kata dilakukan dengan menggunakan istilah Frequency-Inverse Document Frequency (TF-IDF). TFIDF sering digunakan karena relatif sederhana tetapi dapat menghasilkan akurasi dan daya ingat yang tinggi (Guo & Yang, 2016). Nilai TF-IDF dihitung menggunakan (1).

$$\text{tf-idft, d} = \text{tft, d} \times \text{idft} \quad (1)$$

Dimana tf-idft, d adalah nilai TF-IDF dalam istilah t dalam dokumen d. Nilai tft, d dihasilkan dari nilai Frekuensi istilah dalam istilah t dari dokumen d. Sedangkan nilai idft adalah frekuensi dokumen terbalik dari istilah t. Nilai idft dihasilkan dari (2).

$$\text{idft} = \log (N / \text{dft}) \quad (2)$$

N adalah jumlah semua dokumen. Sementara itu, dft adalah jumlah dokumen yang mengandung istilah t. Pembobotan kata-kata menggunakan perpustakaan Phyton, Scikit-learn (Pedregosa et al, 2011).

2.4. *Metode Klasifikasi Random Forest*

Random Forest pertama kali dikenalkan oleh Breiman pada Tahun 2001. Dalam penelitiannya menunjukkan kelebihan Random Forest antara lain dapat menghasilkan *error* yang lebih rendah, memberikan hasil yang bagus dalam klasifikasi, dapat mengatasi *data training* dalam jumlah sangat besar secara efisien, dan metode yang efektif untuk mengestimasi *missing data* (Breiman,2001).

Metode Random Forest (RF) merupakan metode yang dapat meningkatkan hasil akurasi, karena dalam membangkitkan simpul anak untuk setiap node dilakukan secara acak. Metode ini digunakan untuk membangun pohon keputusan yang terdiri dari root node, internal node, dan leaf node dengan mengambil atribut dan data secara acak sesuai ketentuan yang diberlakukan. Root node merupakan simpul yang terletak paling atas, atau biasa disebut sebagai akar dari pohon keputusan. Internal node adalah simpul percabangan, di mana node ini mempunyai output minimal dua dan hanya ada satu input. Sedangkan leaf node atau terminal node merupakan simpul terakhir yang hanya memiliki satu input dan tidak mempunyai output.

Pohon keputusan dimulai dengan cara menghitung nilai entropy sebagai penentu tingkat ketidakmurnian atribut dan nilai *information gain.* Untuk menghitung nilai entropy digunakan rumus seperti pada persamaan 1, sedangkan nilai *information gain* menggunakan persamaan 2. (Schouten, 2016)

$$Entropy\ (Y) = -\sum_i p(c|Y) \log_2 p(c|Y) \quad (1)$$

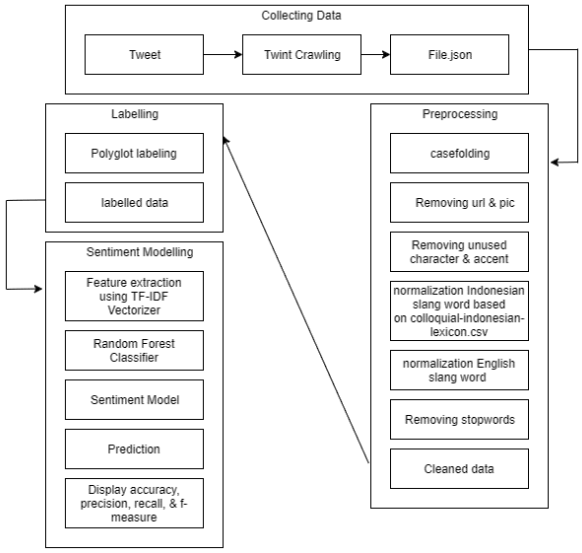
di mana Y adalah himpunan kasus dan p(c|Y) merupakan proporsi nilai Y terhadap kelas c.

$$Information\ Gain\ (Y,a) = Entropy\ (Y) - \sum_{v \epsilon values(a)} \frac{|Y_v|}{|Y_a|} Entropy\ (Y_v) \quad (2)$$

di mana Values(a) merupakan semua nilai yang mungkin dalam himpunan kasus a. Yv adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a. Ya adalah semua nilai yang sesuai dengan a.

3. Metode Penelitian

Dalam penelitian ini, data yang digunakan adalah data twitter yang difilter dengan Bahasa Indonesia dan kata kunci Marvel berjumlah 30.191 *tweet*. Berikut beberapa langkah dan tahapan penelitian yang dilakukan :



Gambar 1 : Langkah-langkah penelitian

3.1. *Collecting Data*

Peneliti mengumpulkan data yang berasal dari *tweet* maupun dari *retweet* masyarakat di twitter. Data diambil dari kata kunci "marvel" dan difilter dengan Bahasa Indonesia saja meskipun juga terdapat Bahasa Inggris dan Bahasa Melayu dalam output data tersebut. *Crawling* dilakukan menggunakan *tools* bernama Twint yang memungkinkan untuk melakukan *crawl* data twitter tanpa perlu menggunakan twitter API. Data yang didapatkan dapat disimpan ke dalam berbagai file di mana dalam penelitian ini, data disimpan dalam bentuk json. Data *tweet* dalam penelitian ini juga difilter untuk tahun 2018 saja dengan total *tweet* yang didapatkan sejumlah 30.191 *tweet*.

3.2. *Preprocessing*

Data *tweet* yang telah didapatkan sebelumnya masih berupa data mentah sehingga perlu diolah agar dapat dengan mudah dicari arah sentimennya. Oleh karena itu, dilakukan tahap *preprocessing* untuk mendapatkan data bersih agar dapat diproses ke tahap berikutnya. Tahapan yang dilakukan adalah :

1. *Case Folding*
   Tahapan ini merupakan proses mengubah bentuk huruf menjadi huruf kecil (*lower case*) agar bentuk huruf menjadi seragam.
2. *Cleaning data*
   Dalam tahapan ini, meliputi pembersihan url, gambar, karakter unik dan juga aksen bahasa.
3. *Normalisasi*
   Pada normalisasi, dilakukan konversi bahasa slang atau akronim menjadi suatu kata utuh yang memiliki makna. Dalam penelitian ini dilakukan normalisasi dalam Bahasa Indonesia maupun Bahasa Inggris dikarenakan adanya *mixed language*. Acuan normalisasi dalam Bahasa Indonesia adalah list kata yang didapatkan dari *link* berikut : https://github.com/nasalsabila/kamus-alay/blob/master/colloquial-indonesian-lexicon.csv
   sedangkan acuan normalisasi dalam Bahasa Inggris didapatkan dari : https://github.com/rishabhverma17/sms_slang_translator/blob/master/slang.txt
4. *Stopword removal*
   Merupakan proses menghilangkan daftar kata-kata yang tidak mendeskripsikan sesuatu seperti kata depan. Dalam penelitian ini, *stopword* yang dihilangkan adalah *stopword* Bahasa Indonesia dan Bahasa Inggris.

3.3. *Labelling*

Hasil *preprocessing* adalah sekumpulan data *tweet* bersih yang siap untuk diekstrak skor sentimennya. Peneliti menggunakan *library* yang telah tersedia bernama *Polyglot* yang mampu mengidentifikasi banyak bahasa termasuk di antaranya Bahasa Indonesia untuk diimplementasikan ke dalam kode Python. Dalam menentukan nilai polaritasnya apabila nilai polaritasnya melebihi 0, maka *tweet* tersebut dapat dikatakan positif, sedangkan apabila nilai polaritasnya di bawah 0, maka *tweet* dapat dikatakan negatif, dan jika sama dengan 0, *tweet*

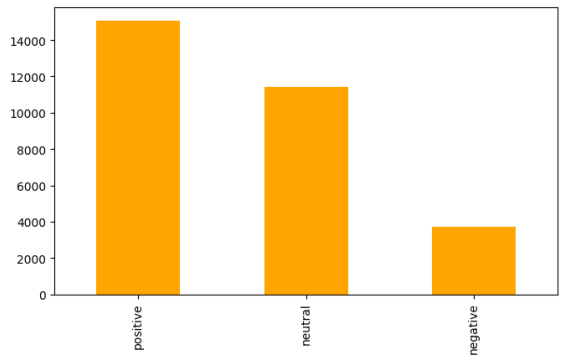
dianggap netral. Selanjutnya, setelah mendapatkan label sentimen untuk tiap *tweet*, simpan keseluruhan data termasuk kolom baru untuk label sentimen ke dalam sebuah file CSV agar mudah untuk diakses. Sebelum menyimpan, pastikan hanya menyeleksi kolom yang benar-benar ingin digunakan agar tidak memberatkan.

3.4. *Sentiment Modelling*

Dalam memodelkan sentimen yang telah didapatkan tersebut, peneliti terlebih dahulu mengekstraksi fitur dari tiap *tweet* menggunakan *TF-IDF Vectorizer*. Setelah mendapatkan fitur tersebut dan kelas berupa label sentimen, barulah dilakukan *modelling* dengan metode klasifikasi *Random Forest*. Data dibagi menjadi data *train* 80% dan data *testing* 20%. Pengujian model yang digunakan adalah dengan *supplied test set* sehingga dari hasil tersebut, didapatkan akurasi, presisi, recall, dan f-measure dari klasifikasi data twitter tersebut.

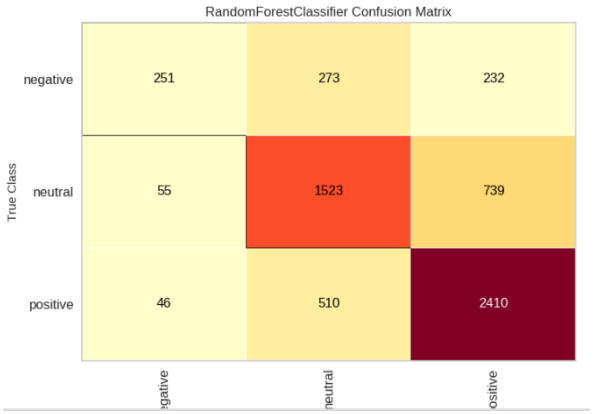
4. Hasil dan Pembahasan

Penelitian ini bertujuan untuk mengetahui sentimen masyarakat terhadap produk marvel dan alasannya yang hasilnya diharapkan mampu membantu pihak marvel dalam meningkatkan pangsa pasar. Pada penelitian ini digunakan data dengan kata kunci "marvel" berjumlah 30.191 data *tweet* dengan rentang waktu 01 Januari 2018 hingga 31 Desember 2018. Hasil dari analisis sentimen ini melahirkan tiga kelas, yaitu positif, negatif, dan netral yang ditentukan dari polaritas masing-masing *tweet*. Berikut persebaran sentimen yang didapatkan :



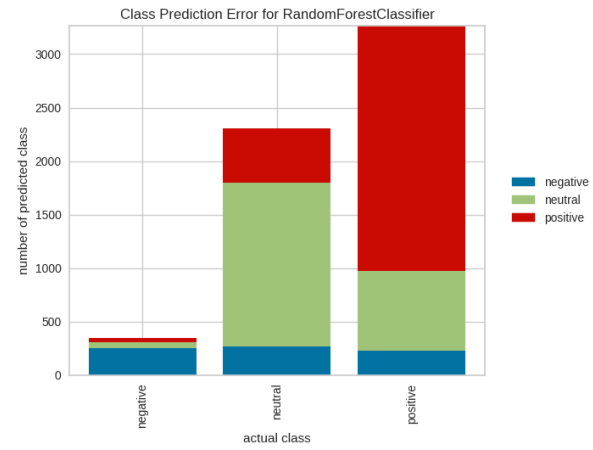
Gambar 2 : Persebaran sentimen

Dari gambar 2, dapat terlihat bahwa cuitan masyarakat tentang marvel lebih banyak positif daripada negatif di mana cuitan positif sebanyak 15.067 *tweet* dan cuitan negatif hanya sebanyak 3735 *tweet*. Untuk menguji akurasi dari klasifikasi ini, maka dibangunlah fitur menggunakan *TF-IDF Vectorizer* dengan data training sebesar 80% dan data testing sebesar 20%. Kemudian data tersebut dimodelkan dan diuji sehingga mendapatkan hasil seperti berikut :



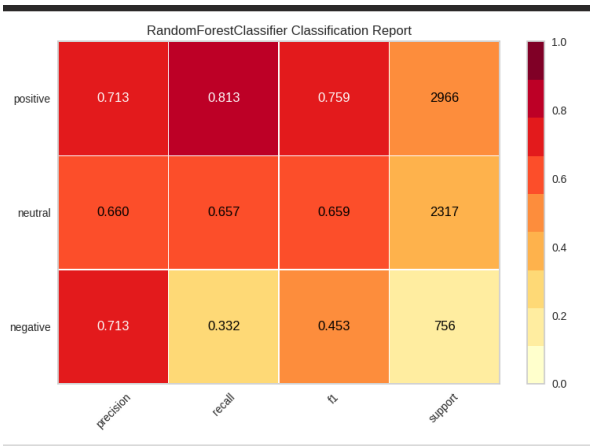
Gambar 3 : Confusion matrix

Pada *confusion matrix* tersebut, dapat terlihat bahwa data positif dan netral sebagian besar telah terklasifikasi dengan benar, namun untuk data negatif masih terdapat kesalahan klasifikasi sehingga akurasi dari klasifikasi ini sendiri hanya sebesar 69%.



Gambar 4 : *Class prediction error*

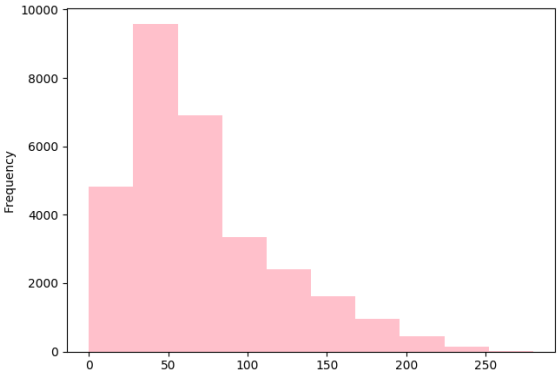
Pada diagram di atas juga terlihat bahwa kelas positif dan netral terklasifikasi dengan baik, berbeda dengan kelas negatif yang lebih bias sehingga klasifikasi untuk kelas negatif menjadi kurang tepat. Hal itu bisa disebabkan karena bermacam-macam hal, seperti teks yang kurang bersih karena penggunaan *slang word* dan akronim di luar acuan yang telah digunakan.



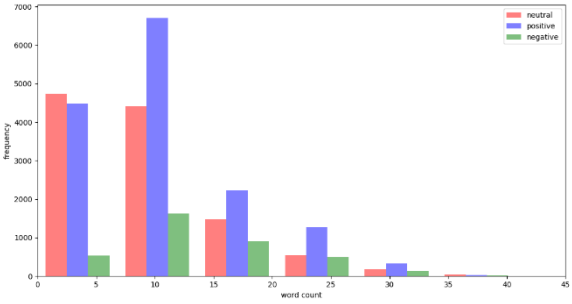
Gambar 5 : *Classification report*

Dari gambar 5, dapat diketahui bahwa nilai presisi sebesar 69%, nilai *recall* sebesar 60%, dan nilai *f-measure* sebesar 62% dengan tingkat akurasi sebesar 69%. Hasil tersebut menunjukkan bahwa tingkat akurasi ini masih tergolong rendah.

Setelah melakukan pengujian tersebut, selanjutnya peneliti akan melakukan analisis terhadap beberapa data yang telah didapatkan. Berikut data-data yang berhasil dihimpun :



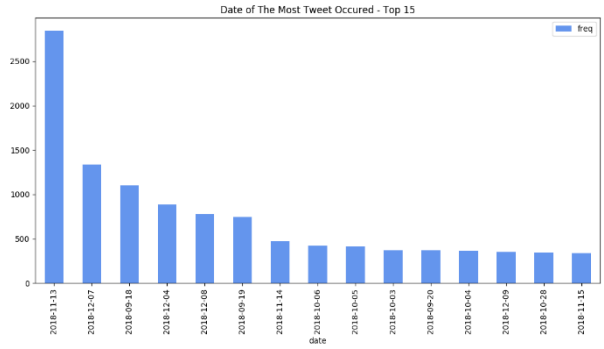
Gambar 6 : Frekuensi jumlah karakter teks twitter yang digunakan



Gambar 7 : Frekuensi jumlah kata yang digunakan per sentimen

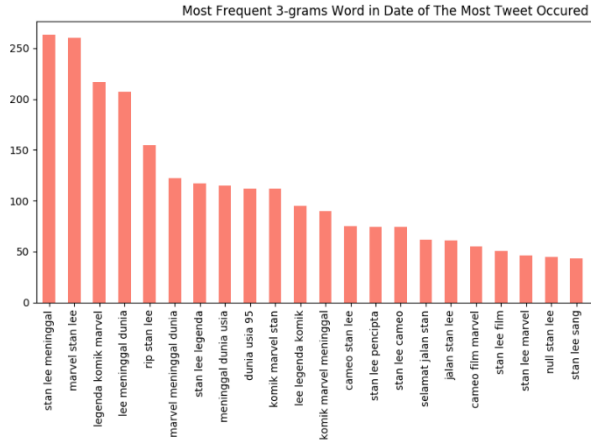
Pada gambar 6 menunjukkan rata-rata pengguna twitter yang membicarakan tentang marvel hanya memberikan komentar singkat dengan jumlah karakter di bawah 100 dari 280 karakter yang bisa diinput. Hal tersebut tentu berpengaruh terhadap rendahnya akurasi yang didapatkan dari klasifikasi kelas berdasarkan polaritas tersebut. Pada gambar 7 dapat terlihat *tweet* bernada positif dan negatif rata-rata terdiri atas 10 kata, berbeda dengan *tweet* netral yang banyak menggunakan 5 kata ke bawah karena *tweet* dengan 5 kata ke bawah sulit untuk diprediksi arah sentimennya, selain karena semakin sedikit kata, maka semakin minim informasi yang didapat di *tweet* tersebut.

Selanjutnya, peneliti akan melihat kapan euforia *tweet* mengenai marvel ini meningkat untuk melihat penyebabnya.

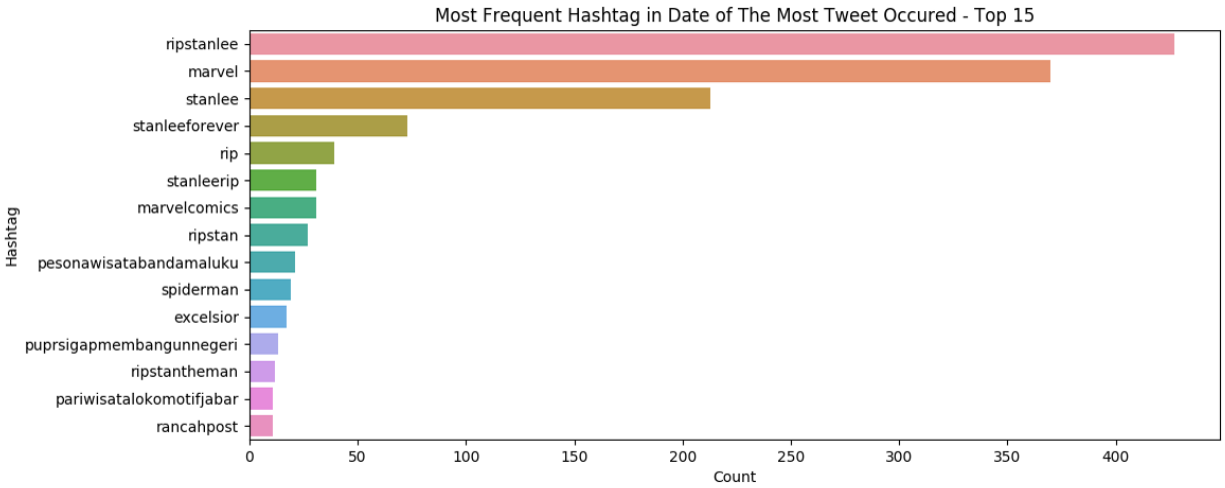


Gambar 8 : Tanggal dengan *tweet* marvel terbanyak

Dari gambar 8, dapat diketahui bahwa *tweet* mengenai marvel terbanyak ada pada tanggal 13 November 2018, oleh karena itu, peneliti akan mencari tahu dengan melihat 3-gram *word* terbanyak dan juga *hashtag* terbanyak pada tanggal 13 November 2018.



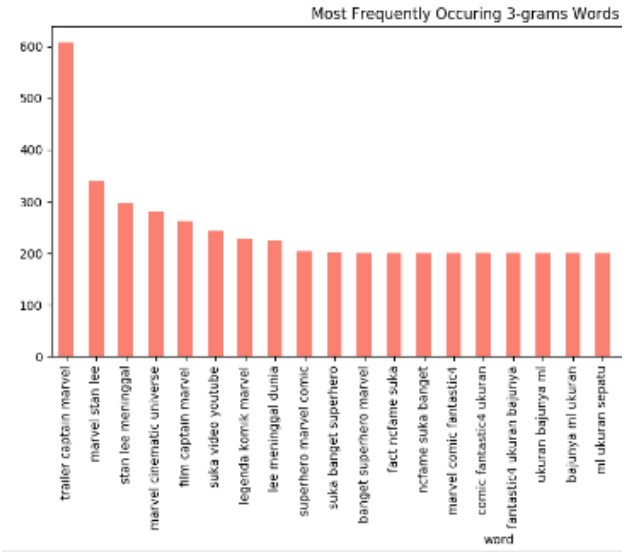
Gambar 9 : 3-gram *word* yang sering muncul di tanggal 13 November 2018



Gambar 10 : *Hashtag* yang sering muncul di tanggal 13 November 2018

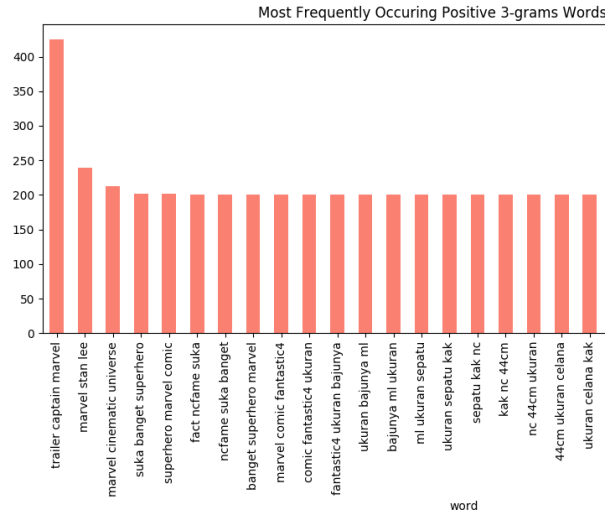
Dari gambar 9 dan gambar 10 dapat diketahui alasan mengapa terdapat banyak *tweet* tentang marvel di tanggal 13 November 2018. Pada gambar 9, kata 3-gram paling banyak di tanggal tersebut adalah "stan lee meninggal", sedangkan pada gambar 10, *hashtag* terbanyak adalah "ripstanlee". Hal tersebut telah memberi jawaban atas pertanyaan sebelumnya, yaitu pada tanggal 13 November 2018, penggemar marvel sedang berduka karena telah kehilangan sosok pendiri marvel sehingga banyak yang menyampaikan belasungkawa.

Setelah itu, peneliti juga akan menganalisis cuitan apa yang paling sering dilontarkan mengenai marvel di tahun 2018.



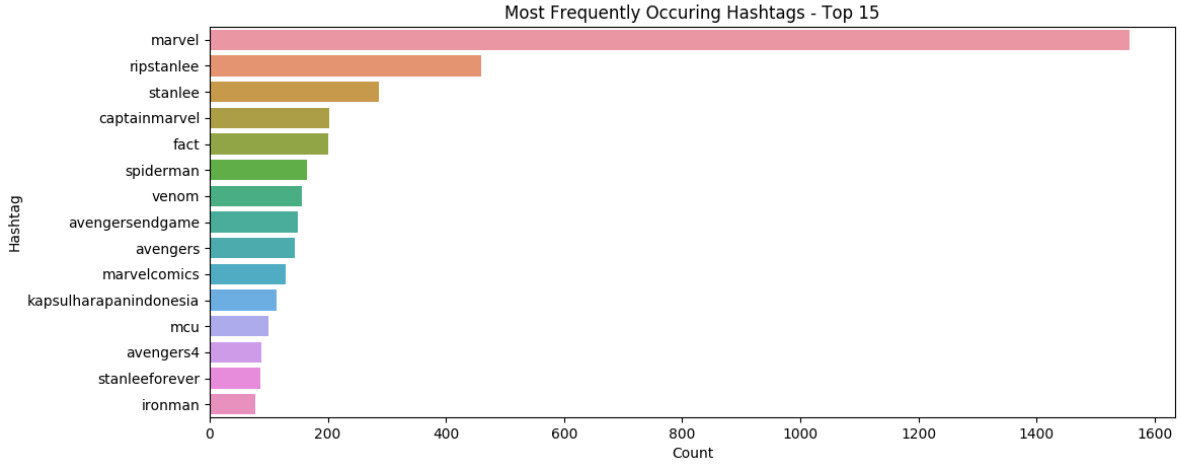
Gambar 11 : 3-gram *word* yang sering muncul di tahun 2018

Pada gambar 11, ditunjukkan bahwa 3-gram *word* yang paling sering disebutkan di tahun 2018 adalah "trailer captain marvel". Perlu diketahui, trailer pertama film captain marvel ini dirilis pada tanggal 19 September 2018 sehingga tidak mengherankan banyak yang *menge-tweet* tentang marvel. Setelah ini, peneliti akan mencari cuitan bernada positif terbanyak dan juga negatif.

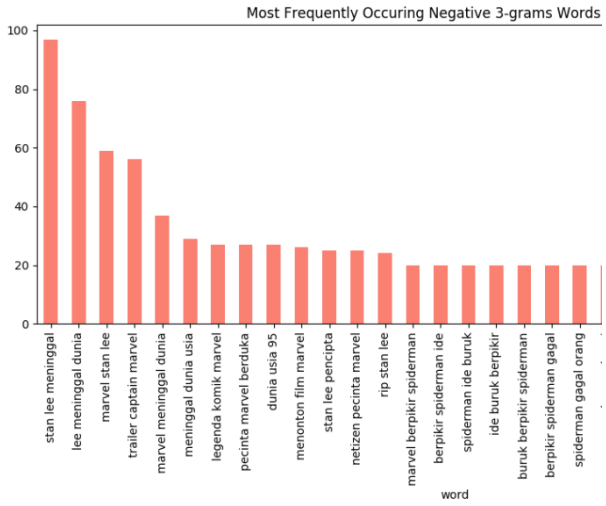


Gambar 12 : 3-gram *word* positif yang sering muncul di tahun 2018

Dari gambar 12, ditunjukkan bahwa kata 3-gram "trailer captain marvel" merupakan kata 3-gram positif terbanyak di tahun 2018. Hal itu menunjukkan bahwa masyarakat khususnya masyarakat Indonesia sangat antusias akan hadirnya film layar lebar yang berjudul "captain marvel" ini sehingga mendapatkan respon yang positif dari masyarakat.



Gambar 13 : *Hashtag* yang sering muncul di tahun 2018



Gambar 14 : 3-gram *word* negatif yang sering muncul di tahun 2018

Berbeda halnya dengan gambar 12, pada gambar 14, ditunjukkan bahwa kata 3-gram negatif terbanyak di tahun 2018 adalah “stan lee meninggal”. Hal tersebut disebabkan kata “meninggal” memiliki konotasi yang negatif sehingga arah sentimen juga mengarah ke negatif walaupun sebenarnya bukan negatif terhadap produk marvel itu sendiri.

Kemudian peneliti juga mencari *hashtag* tentang marvel yang populer di tahun 2018. Hasilnya ditunjukkan pada gambar 13, di mana karena topik yang sedang dibahas adalah mengenai marvel, maka *hashtag* yang paling populer adalah “marvel” juga dan disusul oleh “ripstanlee”, “stanlee”, “captainmarvel” di mana kata-kata yang muncul tersebut adalah kata-kata yang sedang tren di tahun 2018, meskipun ada 1 atau 2 yang tidak berhubungan dengan marvel.

5. Kesimpulan

Berdasarkan hasil penelitian yang sudah dilakukan, maka dapat ditarik kesimpulan sebagai berikut :

1. Sentimen masyarakat Indonesia terhadap marvel cenderung positif.
2. Tingkat akurasi dalam klasifikasi sentimen menggunakan *Random Forest* tergolong rendah yaitu sebesar 69% dengan *precision*, *recall*, dan *f-measure* masing-masing sebesar 69%, 60%, dan 62%.
3. Jumlah karakter rata-rata yang terdapat dalam *tweet* tentang marvel adalah sebesar 50 karakter.
4. *Tweet* bernada positif dan negatif rata-rata terdiri atas 10 kata, sedangkan netral di bawah 5 kata.
5. *Tweet* tentang marvel terbanyak pada tanggal 13 November 2018 karena sedang berbelasungkawa terhadap Stan Lee yang mendirikan Marvel.
6. Penggemar marvel dari Indonesia di tahun 2018 banyak membicarakan tentang *trailer Captain Marvel*.
7. Rata-rata *tweet* dengan sentimen positif membicarakan tentang *trailer Captain Marvel*.
8. Rata-rata *tweet* dengan sentimen negatif membicarakan tentang kematian Stan Lee, di mana meninggal memiliki konotasi negatif.

Daftar Pustaka

Agarwal, A. et al. 2014. *Sentiment Analysis of Twitter Data*. Department of Computer Science Columbia University. Available at : http://www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf.

Alim, S. 2015. *Analysis of Tweets Related to Cyberbullying: Exploring Information Diffusion and Advice Available for Cyberbullying Victims*. International Journal of Cyber Behavior, Psychology and Learning, 5(4), 31-52 (October-December 2015).

Asih, R. 2012. *Indonesia Pengguna Twitter Terbesar Kelima Dunia*. http://www.tempo.co/read/news/2012/02/02/072381323/ [Diakses 16 Desember 2019].

A. Guo and T. Yang. 2016. *Research and improvement of feature words weight based on TFIDF algorithm*. 2016 IEEE Information Technology, Networking, Elecronic, and Automation Control Conference, pp. 415-419.

Breiman, L. 2001. *Random Forest*. Machine Learning, 45(1), pp.5-32.

Feldman, R & Sanger, J. 2007. *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press : New York.

F. Pedregosa et al. 2011. *Scikit-learn: Machine Learning in Python*. J. Mach. Learn. Res., vol. 12, pp.2825-2830.

Pang, B. & Lee, L. 2008. *Opinion Mining and Sentiment Analysis*. Foundations and Trends In Information Retrieval Vol. 2, No. 1–2 (2008) 1–135 C 2008 B. Pang and L.Lee.

Perdana, R. S. & Pinandito, A. 2017. *Combining Likes-Retweet Analysis and Naive Bayes Classifier Within Twitter for Sentiment Analysis*. Malang : Fakultas Ilmu Komputer Universitas Brawijaya.

Schouten, K., F. Fasincar, and R. Dekker. 2016. *An Information gain-Driven Feature Study for Aspect-Based Sentiment Analysis*. Natural Language Processing and Information Systems, pp. 48-59.

Setyanto, D. & Adiwibawa, B. 2018. *Membaca Warna pada Karakter Superhero Marvel*. Jurnal Desain Komunikasi Visual, Manajemen Desain dan Periklanan Vol. 03 No. 02 (September 2018)

Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, F.J. 2005. *Text mining : Predictive Methods for Analyzing Unstructured Information*. Springer : New York.

W. Medhat, A. Hassan, dan H. Korashy. 2014. *Sentiment Analysis Algorithms and Applications: A Survey*. Ain Shams Eng.J., vol. 5, hal. 1093-1113.

Y. Wan. 2015. *An Ensemble Sentiment Classification System of Twitter Data for Airline Sevices Analysis*. IEEE 15th International Conference on Data Mining Workshops.