ANALISIS OPINI TERHADAP DC UNIVERSE PADA MEDIA SOSIAL TWITTER MENGGUNAKAN METODE NAÏVE BAYES

Andrew Patrick de Fretes[1)], **Anggit Dwi Hartanto**[2)],

[1)] *Informatika Universitas AMIKOM Yogyakarta*
[2)] *Informatika Universitas AMIKOM Yogyakarta*
Jl Ringroad Utara, Condongcatur, Depok, Sleman, Yogyakarta Indonesia 55283
Email : andrew.0278@students.amikom.ac.id[1)], *anggit@amikom.ac.id*[2)]

Abstract – DC Universe is a fictional universe in which a collection of superheroes and super villains based on characters that appear in comic books by DC Comics is in it. DC Comics itself is the largest and oldest comic book publisher that produces and displays superheroes and super villains.

To start a super hero-themed business, there are a number of business examples that can be used as references and can also be used to reap profits from the business, namely, rental of comics, selling merchandise, making clothing lines, making cosplay costumes, making superhero-themed foods, selling action figure.

In an effort to start a super hero-themed business, especially the DC Universe theme, it is necessary to pay attention / listen to DC Universe consumers in Indonesia. Classification of opinions or sentiment analysis is one way to find out about a person or group of people towards certain products, services, issues or groups from various social media platforms and the internet. Twitter is one of the social media that is loved by the people of Indonesia. This research tries to utilize what was written by Twitter social media users or better known as a tweet. Tweets will be processed by text mining and processed again using the Naïve Bayes Classifier algorithm.

Keywords - Social Media, Sentiment Analysis, Naïve Bayes Classifier.

1. Pendahuluan

1.1 Latar Belakang

DC Universe merupakan sebuah alam semesta fiksi dimana kumpulan pahlawan super dan penjahat super berdasarkan pada karakter yang muncul dalam buku-buku komik oleh DC *Comics* ada didalamnya [1]. DC *Comics* sendiri adalah penerbit buku komik terbesar dan tertua di amerika milik DC *Entertainment* yang memproduksi dan menampilkan pahlawan super dan penjahat super [2].

Untuk memulai sebuah bisnis bertema pahlawan super, ada beberapa contoh bisnis yang bisa jadi referensi dan juga dapat dimanfaatkan untuk meraup keuntungan dari bisnis tersebut yaitu, persewaan komik, menjual *merchandise*, membuat *clothing line*, membuat kostum cosplay, membuat makanan bertema pahlawan super, menjual *action figure*. [3]

Dalam upaya untuk memulai bisnis bertema pahlawan super khususnya bertema DC Universe, perlu memperhatikan / mendengarkan konsumen DC Universe di Indonesia. Dengan mengetahui sentimen masyarakat Indonesia dan hal-hal apa saja yang seringkali muncul / dibicarakan di media sosial terhadap DC Universe. Dengan mengetahui sentimen, pendapat, keinginan masyarakat Indonesia terhadap DC Universe, maka hal itu dapat digunakan sebagai acuan untuk memulai bisnis bertema pahlawan super di Indonesia. Oleh karena itu penulis merasa penting untuk mengajukan penelitian dengan judul Analisis Opini Terhadap DC Universe Pada Media Sosial Twitter Menggunakan Metode Naïve Bayes.

2. Dasar Teori

2.1 *Text Mining*

Text Mining merupakan bentuk lain dari *data mining* yang berusaha menemukan pola yang menarik dari sekumpulan *data* tekstual yang berjumlah besar. *Text mining* bertujuan untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber *data* yang digunakan dalam proses *text mining* adalah sekumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari *text mining* antara lain yaitu pengkategorisasian teks dan pengelompokkan teks. [4]

2.2 Analisis Sentimen

Analisis sentimen adalah proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu dokumen, kalimat atau fitur entitas/aspek opini [5]. Analisis sentimen dilakukan untuk melihat apakah opini yang dikemukakan seseorang dalam dokumen, kalimat atau fitur entitas/aspek bersifat positif, negatif atau netral [6].

2.3 Evaluasi

Melakukan evaluasi performa sistem untuk menguji hasil dari klasifikasi dengan mengukur nilai

performa dari sistem yang telah dibuat. Parameter pengujian yang digunakan untuk evaluasi hasil dari klasifikasi yaitu akurasi yang perhitungannya diperoleh dari tabel *confusion matrix* (matrik klasifikasi atau tabel kontigensi) [7]. Tabel *confusion matrix* Positif, Negatif, Netral dapat dilihat seperti yang ditunjukan pada tabel 1.

Tabel 1. *Confusion Matrix* Positif, Netral & Negatif

CONFUSION MATRIX		**Prediksi**		
		Positif	Negatif	Netral
Aktual	Positif	True Positive (TP)	False Positive (FP)	False Positive (FP)
	Negatif	False Negative (FNeg)	True Negative (TNeg)	False Negative (FNeg)
	Netral	False Neutral (FNeut)	False Neutral (FNeut)	True Neutral (TNeut)

Matrix tersebut memiliki enam nilai yang digunakan sebagai acuan dalam perhitungan *confusion matrix* :

A. *True Positive* (TP) adalah kelas yang diprediksi sistem positif dan fakta kelasnya positif.

B. *True Negative* (TNeg) adalah kelas yang diprediksi sistem negatif dan fakta kelasnya negatif.

C. *True Neutral* (TNeut) adalah kelas yang diprediksi sistem netral dan fakta kelasnya netral.

D. *False Positive* (FP) adalah kelas yang diprediksi sistem positif dan fakta kelasnya tidak positif.

E. *False Negative* (FNeg) adalah kelas yang diprediksi sistem negatif dan fakta kelasnya tidak negatif.

F. *False Neutral* (FNeut) adalah kelas yang diprediksi sistem netral dan fakta kelasnya tidak netral.

Untuk menghitung akurasi prediksi kelas positif, negatif dan netral menggunakan rumus pada persamaan 1 sebagai berikut,

$$Akurasi = \frac{TP + TNeg + TNeut}{(TP + TNeg + TNeut + FP + FNeg + FNeut)} \times 100\% \quad (1)$$

3. Metode Penelitian

3.1 Metode Pengumpulan Data

Dataset berupa teks berbahasa Indonesia yang diambil dari *website* http://www.twitter.com/ memanfaatkan *search* API *Twitter* yang telah disediakan oleh *Twitter* dengan kata kunci DC Universe. Data yang diambil untuk penelitian ini akan dibedakan menjadi dua data yaitu : data latih dan data uji.

3.1.1 Data Latih

Data latih diambil dalam kurun waktu selama 8 bulan terhitung mulai bulan Juli 2019 – Februari 2020. *Tweet* yang sudah dikumpulkan kemudian diberi label sentimen / diklasifikasikan secara manual menjadi tiga kategori yaitu positif, netral dan negatif dengan *tweet* yang berjumlah 364. Dengan jumlah tweet masing-masing kategori sentimen yang dapat dilihat pada Tabel 2.

Tabel 2. Data Uji

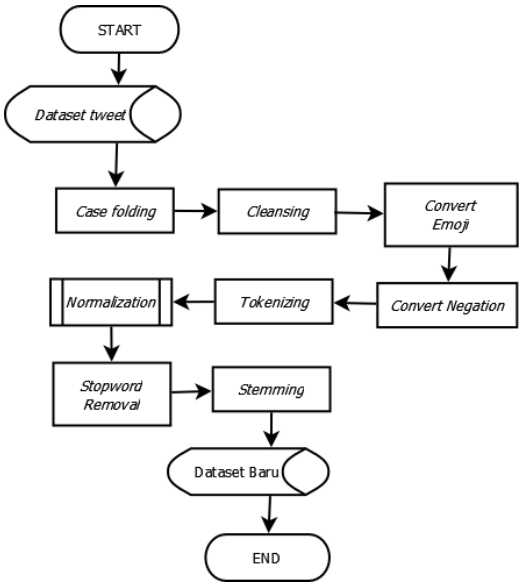
Positif	Negatif	Netral
139	101	124

3.1.2 Data Uji

Data uji pada penelitian ini merupakan data yang diambil secara *realtime* sejak 7 hari kebelakang dengan fungsi *crawling* pada sistem memanfaatkan API *twitter*.

3.2 *Preprocessing*

Preprocessing merupakan tahapan untuk mempersiapkan proses penyeleksian *data* agar *data* lebih terstruktur dan berkualitas serta dapat diolah dengan cepat dan tepat. p*reprocessing* sangat penting dalam pembuatan analisis sentimen, terutama untuk sentimen pada media sosial yang sebagian besar berisi kata – kata atau kalimat yang tidak formal dan tidak terstruktur serta memiliki noise yang besar [8]. Diagram alur tahapan-tahapan proses preprocessing dapat dilihat pada Gambar 1 sebagai berikut.



Gambar 1. Diagaram Alur *Text Preprocessing*

3.3 Pembobotan TF-IDF

TF-IDF (*Term Frequency-Inverse Document Frequency*) digunakan merupakan teknik pembobotan kata yang mempunyai ide dasar untuk memberikan bobot pada setiap kata atau dokumen, dimana perhitungan bobot *term* dalam sebuah dokumen dilakukan dengan mengalikan nilai *Term Frequency* (TF) dengan *Inverse Document Frequency* (IDF), *term* dapat berupa kata, frasa atau tipe sintatik lainnya. [9]

$$a_{ij} = tf_{ij} \times idf_j \tag{2}$$

$$a_{ij} = tf_{ij} \times \log\left(\frac{N}{n_j}\right) \tag{3}$$

Ketika N sama dengan n_j, maka a_{ij} menjadi nol, ini sering muncul dalam dataset yang kecil, jadi perlu diterapkan beberapa teknik perataan untuk meningkatkan rumus pada persamaan 3, maka dari persamaan tersebut dapat ditulis sebagai berikut [10]:

$$a_{ij} = \log(tf_{ij} + 1) \times \log\left(\frac{N+1}{n_j}\right) \tag{4}$$

Keterangan :

t : adalah *term* (kata)

d : adalah dokumen atau kumpulan teks

a_{ij} : adalah bobot *term* t_j terhadap dokumen d_i

tf_{ij} : adalah jumlah kemunculan *term* t_j dalam dokumen d_i

N : adalah jumlah total dokumen dalam *dataset*

n_j : adalah jumlah dokumen yang mengandung *term* t_j

3.4 *Naïve Bayes Classifier*

Naïve Bayes Classifier, merupakan salah satu metode yang digunakan dalam *data mining* yang didasarkan pada teori keputusan *Bayes*. *Naive Bayes Classifier* memiliki kemampuan klasifikasi seperti metode *decision tree*, *neural network*, *k-nearest-neighbourhood classifier*, *classification (IF-THEN) rule*. Kelebihan dari *Naïve Bayes Classifier* adalah sederhana tetapi memiliki akurasi yang tinggi dan dapat menangani permasalahan banyak kelas. [11]

$$P(V_j) = \frac{docs_j}{training} \tag{5}$$

$$P(w_k|v_j) = \frac{n_k+1}{n+|Kosakata|} \tag{6}$$

Keterangan :

$docs_j$: Jumlah dokumen pada kategori j

training : Jumlah dokumen yang digunakan dalam proses data latih

w_k : Kata (*word*) k dalam semua dokumen yang diberi label sebagai j

v_j : Semua kata (Kosakata) dalam kelas j

n_k : Jumlah munculnya kata w_k dalam kelas v_j

n : Jumlah kosakata yang ada pada kategori v_j

Kosakata : Total kata unik dalam dokumen atau data latih

4. Pembahasan

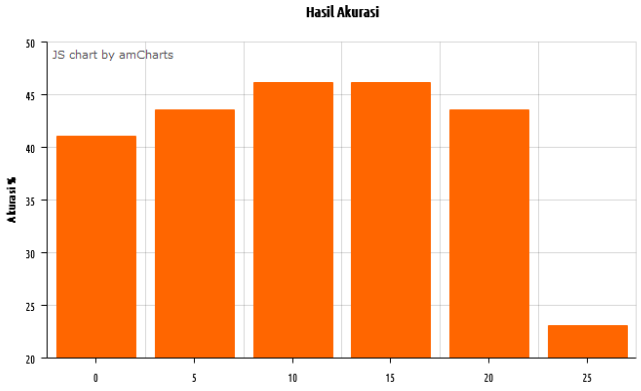
Pada tahap pengujian sistem peneliti melakukan ujicoba pengaruh dari menyeleksi *terms* berdasarkan nilai bobot TF-IDF. Peneliti menemukan bahwa penyeleksian *terms* akan mempengaruhi tingkat akurasi sistem. Pengujian ini dilakukan dengan cara menggunakan menggunakan data uji yang di *crawling* pada pertengahan bulan maret 2020. Evaluasi pengujian akurasi sistem dilakukan dengan memberi berbagai macam batas nilai minimum berdasarkan nilai bobot *terms* antara lain 0, 5, 10, 15, 20, 25. Hasil prediksi klasifikasi, dihitung akurasinya secara manual dengan mengimplementasikan persamaan 1. Sehingga hasil dari akurasi sistem ini didapat dari 6 kali percobaan dengan berbagai batasan

minimum bobot. Hasil dari pengujian sistem ini dapat dilihat pada tabel 3 sebagai berikut.

Tabel 3. Tabel Hasil Akurasi

0	5	10	15	20	25
41.02	43.58	**46.15**	**46.15**	43.58	23.07

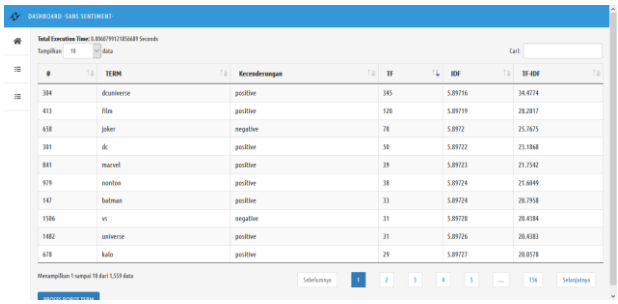
Atau disajikan dalam bentuk diagram batang seperti pada gambar 2 sebagai berikut.



Gambar 2. Diagram Batang Hasil Akurasi

Berdasarkan data tersebut, terlihat bahwa batas minimum bobot terms 10 & 15 meningkatkan akurasi sedangkan batas minimum 25 menurunkan nilai akurasi.

Pada halaman pembobotan TF-IDF juga dapat digunakan untuk mengetahui *terms* yang paling sering keluar pada koleksi data. 10 Terms yang paling sering muncul pada teks *tweet* mengenai DC Universe dengan *terms* 'dcuniverse' sebagai pengecualian dapat dilihat pada gambar 3 sebagai berikut.



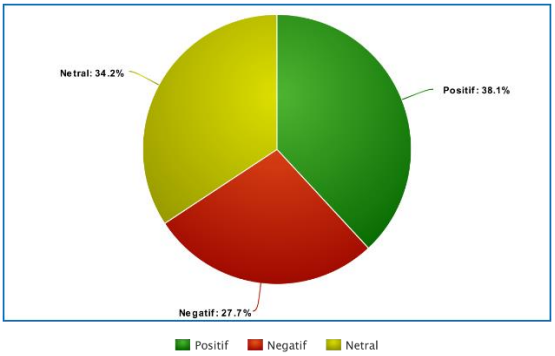
Gambar 3. *Terms* Yang Sering Muncul

5. Penutup

5.1 Kesimpulan

Berdasarkan penelitian yang sudah dilakukan, dapat disimpulkan bahwa :

1. Berdasarkan koleksi data latih sentimen yang diambil antara bulan juli 2019-februari 2020, dimana sentimen positif lebih mendominasi. Dengan komposisi sebagai berikut : *tweet* yang mempunyai sentimen positif sebanyak 139 *tweet*, negatif sebanyak 101, dan netral sejumlah 124 dari total 364 *tweet*. Komposisi sentimen *tweet* mengenai DC Universe dapat dilihat pada Gambar 4 berikut.



Gambar 4. *Komposisi sentimen mengenai DC Universe*

2. Pada penelitian ini hasil dari akurasi sistem dengan 6 kali percobaan menggunakan berbagai batasan minimum bobot TF-IDF, dengan hasil uji coba pada penelitian ini yaitu batas minimum 10 & 15 meningkatkan akurasi sedangkan batas minimum 25 menurunkan nilai akurasi.

3. Pada penelitian ini menghasilkan 10 *terms* teratas yang sering muncul pada teks *tweet* mengenai DC Universe dengan *terms* 'dcuniverse' sebagai pengecualian adalah : 'film', 'joker', 'dc', 'marvel', 'nonton', 'batman', 'vs', 'universe', 'kalo'. Berdasarkan data yang ada *terms* 'batman', 'superman', 'flash', 'titans', 'swamp', 'doom' yang merupakan jajaran karakter dalam DC Universe mempunyai kecenderungan sentimen positif. Kemudian *tweet* dengan *terms* 'joker', 'harley', yang merupakan karakter dari DC Universe mempunyai kecenderungan bersentimen negatif.

Berdasarkan data yang ada *terms* yang sering dibicarakan dalam teks *tweet* mengenai DC Universe dimana dapat dimanfaatkan informasinya sebagai acuan untuk memprioritaskan karakter mana yang mempunyai nilai jual lebih dalam memulai bisnis bertema superhero karena karakter tersebut mempunyai kecenderungan sentimen positif, sedangkan untuk karakter yang mempunyai kecenderungan sentimen negatif dapat mempunyai nilai jual yang rendah dalam memulai bisnis bertema pahlawan super, dapat dilihat pada tabel 5 dan tabel 6 sebagai berikut.

Tabel 5. Tabel Acuan Bisnis Positif

‘batman’, ‘superman’, ‘flash’, ‘titans’, ‘swamp’, ‘doom’

Tabel 6. Tabel Acuan Bisnis Negatif

‘joker’, ‘harley’

5.2 Saran

Beberapa saran untuk pengembangan penelitian dimasa akan datang adalah sebagai berikut :

1. Pada penelitian ini, peneliti menemukan beberapa masalah yang disebabkan *TwitterOAuth* dalam melakukan *crawling* diharapkan akan ada perbaikan dimasa yang akan datang atau penggunaan *library* lain yang lebih baik.
2. Pada penelitian selanjutnya peneliti menyarankan menambahkan metode *Information Gain* dan *Adaboost* untuk meningkatkan akurasi klasifikasi.
3. Menambahkan lagi kata untuk koleksi kamus tipografi (typo), kata gaul dan slang.
4. Pada penelitian selanjutnya peneliti menyarankan agar mengganti metode pembobotan TF-IDF dengan metode TF-IDF-CF untuk meningkatkan akurasi.

Daftar Pustaka

[1] Wikipedia contributors, “DC Universe,” *Wikipedia, The Free Encyclopedia.*, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=DC_Universe&oldid=950558951. [Accessed: 31-Mar-2020].

[2] Tim DeForest, “DC Comics,” *Encyclopædia Britannica, inc.*, 2020. [Online]. Available: https://www.britannica.com/topic/DC-Comics. [Accessed: 31-Mar-2020].

[3] Putriana Cahya, “Demam The Avengers, 6 Bisnis Ini Dijamin Bakal Laris Manis,” 2018. [Online]. Available: https://www.idntimes.com/business/economy/putriana-cahya/demam-avengers-6-bisnis-ini-laris-manis-1/full. [Accessed: 31-Mar-2020].

[4] F. Nurhuda, S. Widya Sihwi, and A. Doewes, “Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier,” *J. Teknol. Inf. ITSmart*, vol. 2, no. 2, p. 35, 2016, doi: 10.20961/its.v2i2.630.

[5] G. Berliana, M. T. Shaufiah, S.T., and M. T. Siti Sa’adah, S.T., “Klasifikasi Posting Tweet mengenai Kebijakan Pemerintah Menggunakan Naive Bayesian Classification,” *e-Proceeding Eng.*, vol. 5, no. 1, pp. 1562–1569, 2018.

[6] I. F. Rozi, E. N. Hamdana, M. Balya, and I. Alfahmi, “Pengembangan Aplikasi Analisis Sentimen Twitter (Studi Kasus Samsat Kota Malang),” *Inform. Polinema*, vol. 1, no. 1, pp. 149–154, 2018, doi: 10.33795/jip.v4i2.164.

[7] A. Novantirani, M. K. Sabariah, and V. Effendy, “Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine,” *e-Proceeeding Eng.*, vol. 2, no. 1, pp. 1–7, 2015.

[8] A. (UNIVERSITAS I. N. M. M. I. M. SYAKURO, “PADA MEDIA SOSIAL MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER (NBC) DENGAN SELEKSI FITUR INFORMATION GAIN (IG) HALAMAN JUDUL SKRIPSI Oleh : ABDAN SYAKURO,” *Anal. sentimen Masy. terhadap e-commerce pada media Sos. menggunakan Metod. naive bayes Classif. dengan Sel. fitur Inf. gain*, pp. 1–89, 2017.

[9] D. H. Wahid and A. SN, “Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity,” *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 10, no. 2, p. 207, 2016, doi: 10.22146/ijccs.16625.

[10] M. Liu and J. Yang, “An improvement of TFIDF weighting in text categorization,” *Int. Conf. Comput. Technol. Sci.*, vol. 47, no. Iccts, pp. 44–47, 2012, doi: 10.7763/IPCSIT.2012.V47.9.

[11] S. F. Rodiyansyah and Edi Winarko, “Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification,” *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 6, no. 1, pp. 91–100, 2012, doi: 10.1163/ej.9789004182127.i-302.6.

Biodata Penulis

Andrew Patrick de Fretes, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Informatika Universitas AMIKOM Yogyakarta, lulus tahun 2020.

Anggit Dwi Hartanto, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika STMIK AMIKOM Yogyakarta, lulus tahun 2009. Memperoleh gelar Magister Komputer (M.Kom) Program Pasca Sarjana Magister Teknik Informatika STMIK AMIKOM Yogyakarta, Lulus tahun 2011. Saat ini menjadi Dosen di Universitas AMIKOM Yogyakarta, pada Program StudiInformatika..