Rekomendasi Kendaraan Roda 4 Berdasarkan Tweet Customer Menggunakan Word2Vec

Iwan Syarif[1], Rengga Asmara[2], Bagas Dewangkara[3]
[1,2,3]Departemen Teknik Informatika dan Komputer
Politeknik Elektronika Negeri Surabaya
Kampus PENS, Jalan Raya ITS, Keputih, Sukolilo
Surabaya, 60111, Indonesia
Email: iwanarif@pens.ac.id[1], rengga@pens.ac.id[2], bagas.dewangkara@gmail.com[3]

Abstrack - Utilizing twitter's fast flow of information, this research proposes a new algorithm to decide suitable automotive product to be promoted into the market. By analyzing each user behavior on twitter, automotive companies could get insight of what did the market interested right now. This will help them to boost their marketing efficiency and increase the probability to attract potential customers. We use Word2Vec as the main algorithm to create a word vector consisting of netizen's opinion, and calculating the connection of each word on each context. After that, we match the vector onto our automotive product dataset, which contains numerous car types, and create the correlations. From prior process, we could propose which product that is suitable with the market's demand and could be promoted further. This algorithm has been implemented on official car company account followers' tweet in Indonesia, and has processed more than 200.000 tweets

Keywords – 4 Wheeler Product Recommendation, Twitter, Word2Vec, Potential Customer.

Intisari – Penelitian ini mempersembahkan sebuah algoritma untuk menentukan rekomendasi kendarran roda 4 yang cocok untuk dijual ke masyarakat sesuai dengan keinginan pasar di sosial media. Melalui algoritma yang telah dibangun, produsen mobil dapat mengetahui apa yang sedang marak dibicarakan oleh masyarakat di sosial media Twitter, dan akan membantu sang produsen untuk menentukan produk mana yang lebih efektif untuk dipromosikan. Penulis menggunakan algoritma Word2Vec untuk membangun sebuah ruang vektor yang berisikan kata-kata yang diperbincangkan oleh warganet, lalu melihat koneksi dari setiap kata-kata yang ada. Setelah itu penulis mencari kecocokan antara beberapa dataset produk yang akan dipromosikan dengan tweet-tweet yang membahas produk tersebut. Dari hasil itu penulis dapat menentukan sekiranya produk manakah yang tengah hangat di mata warganet dan dapat dipromosikan lebih lanjut. Algoritma ini telah diimplementasikan menggunakan data Twitter pengikut akun produsen mobil yang ada di Indonesia, dan telah memproses lebih dari 200.000 tweet.

Kata Kunci - Rekomendasi Kendaraan Roda 4, Twitter, Word2Vec, Calon Pembeli

I. PENDAHULUAN

Memasuki era pertukaran informasi dengan kecepatan tinggi, kebutuhan untuk memproses informasi yang didapat semakin tinggi tiap harinya. Fenomena ini membuat banya perusahaan berlomba untuk mengambil keuntungan dari arus informasi. Salah satu platform pertukaran informasi secara cepat ialah *Twitter*. *Twitter*, sebagai media sosial berbasis teks memberikan penggunanya kebebasan berpendapat dan menyampaikannya menggunakan teks. Karakteristik ini memberikan peneliti peluang untuk menganalisa data yang disebarkan. Seperti penelitian yang telah dilakukan oleh Zeel Doshi [1], Twitter merupakan *platform* yang sangat mumpuni untuk digunakan sebagai sumber data penelitian. Dengan menganalisa data *tweet* dari berbagai pengguna, penulis dapat mengetahui kebiasaan seseorang, preferensi, dan berbagai sifat individu lainnya. Beberapa perusahaan mobil di Indonesia telah sudah mulai menyadari akan

potensi dari data media sosial. Dan mereka sangat tertarik untuk menggunakan data dari twitter sebagai salah satu strategi pemasaran mereka.

Permasalahan yang ingin penulis angkat ialah bagaimana cara untuk mengoptimalkan penggunaan data yang didapatkan dari Twitter sebagai sistem perekomendasian produk. Sebagai media sosial yang sangat luas dan hampir tak terbatas, pengguna Twitter tidak selalu membahas mengenai kendaraan roda 4. Oleh karena itu, penelitian ini bertujuan untuk menguji sebuah algoritma baru yang menggunakan data Twitter sebagai sumber sistem perekomendasian kendaraan roda 4.

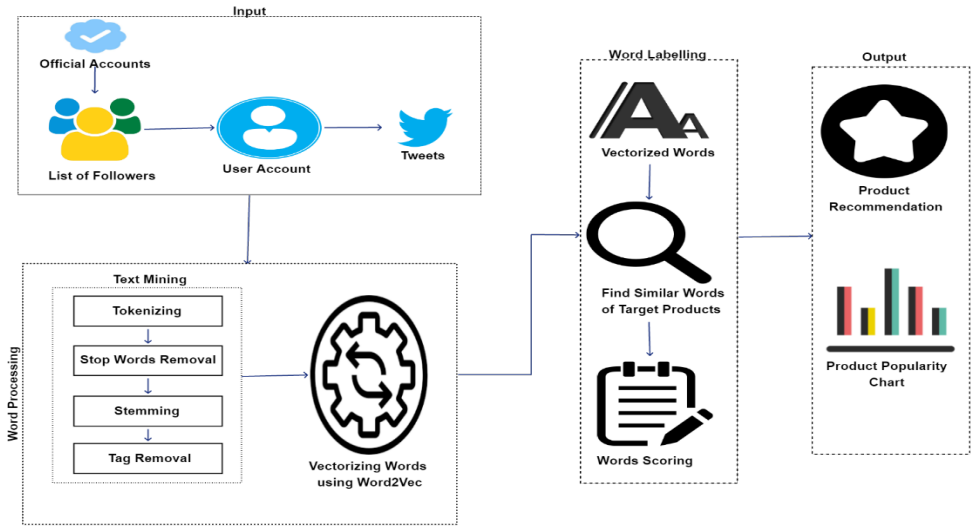
II. SIGNIFIKANSI STUDI

Savitha B Japali [2] menggunakan *fuzzy c-means* dan *association rule generator in* KNIME untuk membuat rekomendasi produk untuk toko retail. Dia menggali data dari aktivitas toko ritel dan menunjukkan status situs web toko saat ini. Termasuk produk yang dijual dan dilihat oleh pelanggan. Vanshee Krishna Kiran M, Archanaa R, dkk [3] membangun rekomendasi produk dan sistem penilaian dengan melakukan analisis sentimen pada ulasan produk. Mereka mendapatkan ulasan produk tertentu dan menganalisis sentimen dari produk tersebut. Ulasan memainkan peran penting dalam pengalaman pengguna membeli suatu produk. Jadi mereka membuat sistem untuk merekomendasikan produk berdasarkan pada sentimen dan mencapai akurasi 88,33%.

Dalam penelitian lain, beberapa tim telah menciptakan sistem rekomendasi menggunakan gabungan analisis sentimen dan model *Fuzzy Kano* [4], dan algoritma genetik [5]. Ini berarti bahwa menciptakan sistem rekomendasi produk adalah pencarian yang tidak pernah berakhir. Dengan setiap iterasi membawa metode dan pendekatan baru untuk menciptakan sistem.

Dalam penelitian ini, penulis membuat algoritma untuk menentukan produk apa yang lebih disukai pengguna tertentu. Berdasarkan tweet mereka di *twitter*. Penulis menggunakan Word2Vec sebagai metode utama untuk menentukan kesamaan *tweet* dengan basis data produk penulis. Kata-kata yang cocok kemudian akan dikalkulasi dan akan menjadi suati rekomendasi produk. Selain itu, penulis juga akan membuat bagan popularitas produk berdasarkan jumlah *tweet* terkait produk yang muncul. Untuk menjelaskannya lebih lanjut, Gambar 1 berisi desain sistem dari algoritma ini.

Metodologi yang digunakan dalam penelitian ini ialah metode kuantitatif. Metode ini dipilih mengingat besarnya jumlah data yang diproses, dan data yang didapatkan.



Gambar 1 Desain Sistem

Dari Gambar. 1 di atas, algoritma ini dipisahkan menjadi empat proses. Input, Word *Processing*, *Word Labelling*, dan *output*.

A. Input

1) Official Account

Akun ofisial yang dimaksud dalam penelitian ini ialah akun resmi dari perusahaan mobil Indonesia seperti Daihatsu, Honda, Mitsubishi, dll.

2) List of Followers

Dari akun resmi, penulis mengumpulkan daftar pengikut akunresmi tersebut. Daftar ini akan menjadi input untuk proses selanjutnya.

3) User Account

Dari daftar pengikut, penulis mengambil data dari setiap akun individu yang ada. Akun yang diambil ditentukan secara acak.

4) Tweets

Terakhir, dari setiap akun individu yang didapat, penulis mengambil 600 *tweet* terbaru yang ada, dean menyimpannya ke dalam database.

B. Word Processing

Langkah ini akan membersihkan data yang dikumpulkan dari langkah sebelumnya, dan kemudian membangun vektor kata dari itu. Proses ini dapat dipecah menjadi dua proses terpisah. Proses pertama ialah text mining, dan proses selanjutnya ialah vektorisasi kata menggunakan Word2Vec.

1) Text Mining

Proses ini akan memisah dan membersihkan *tweet* yang didapatkan. Proses ini terbagi lagi menjadi beberapa langkah, yaitu:

a. *Tokenizing*

Tokenizing akan memsiahkan *tweet* menjadi per kata, dan akan disebut sebagai token untuk sealnjutnya.

b. *Stop Words Removal*

Dari token yang didapat, dialkukan pembersihan *stop words*, atau kata-kata yang tidak memiliki arti dan tidak berguna dalam algoritma ini.

c. *Stemming*

Stemming akan mengembalikan token yang ada menjadi bentuk dasarnya. Dengan menghilangkan imbuhan yang ada.

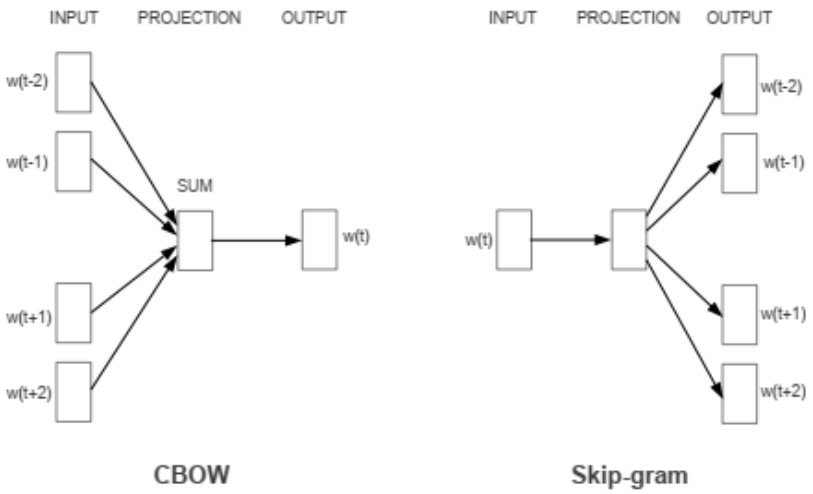
d. *Tag Removal*

Tag removal akan menghilangkan tag bawaan dari *Twitter* yang tidak diperlukan. Contohnya ialah tag retweet (RT), tag http (http://), hashtag (#) dan lainnya.

2) Vectorizing Words using Word2Vec

Memasuki langkah selanjutnya. Dari token yang dibersihkan, penulis membuat vektor kata menggunakan metode yang disebut *Word2Vec* [6]. *Word2Vec* awalnya dicetuskan oleh Thomas Mikolov dan kawan-kawan [7], dan kemudian dijelaskan lebih lanjut oleh Goldberg [8] dan Thu Anh Le [9].

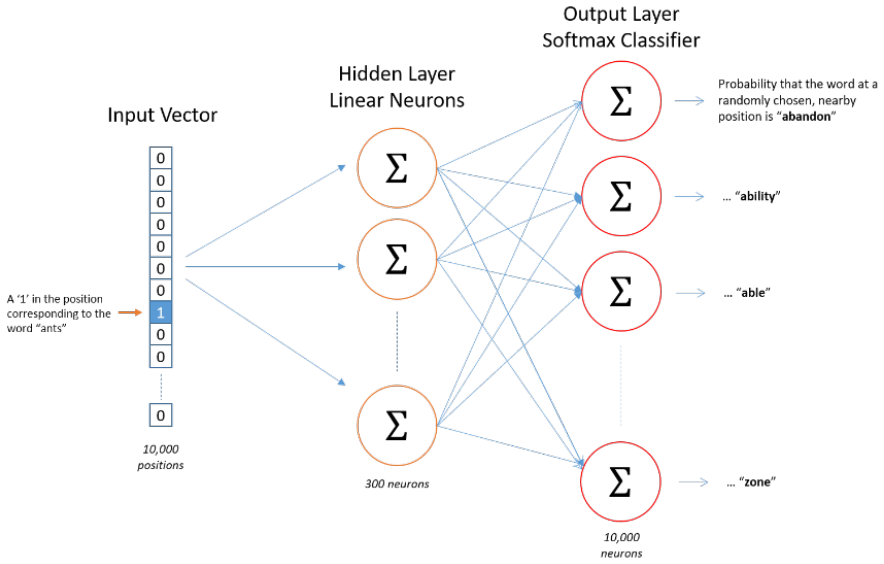
Nama Word2Vec adalah akronim dari "*Word to Vector*", yang mana memiliki arti algoritma untuk mengubah kata menjadi vektor. Algoritma ini adalah *two-layer neural network* yang dilatih untuk memahami konteks kalimat. *Word2Vec* menerima input *corpus*, kemudian membuat ruang vektor dari kata-kata yang ada di dalam corpus, biasanya terdiri dari ratusan ribu dimensi. Setiap titik vektor adalah kata yang memiliki nilai spesifik sehingga kata-kata yang memiliki konteks yang sama akan memiliki nilai yang sama. *Word2Vec* sendiri bukan algoritma monolitik, melainkan memiliki beberapa model dan algoritma yang berbeda. *Word2Vec* bisa model *Skip-Gram*, atau model CBOW (*Continous Bag of Words*).



Gambar 2 Struktur model CBOW dan Skip-gram
(sumber : https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314)

Sebagaimana terlihat pada Gambar. 2, CBOW dan *Skip-gram* memiliki konsep yang sama, tetapi terbalik. CBOW akan menghitung bobot beberapa kata input untuk menemukan konteks, lalu menyarankan kata yang sesuai untuk konteks. Sedangkan *Skip-gram*, alih-alih memprediksi kata saat ini berdasarkan konteks, ia mencoba untuk memprediksi konteks menggunakan kata-kata yang muncul dalam kalimat yang sama.

Dalam penelitian ini, penulis menggunakan model *skip-gram* sebagai algoritma untuk *Word2Vec*. Jadi penulis akan memasukkan kata, dan dapatkan beberapa sebagai balasannya. Seperti yang dinyatakan sebelumnya, *Word2Vec* adalah jaringan saraf, sehingga memiliki beberapa lapisan yang diproyeksikan di dalamnya.



Gambar 3 Model Skip-gram

(sumber: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/)

Gambar 3 menjelaskan proses aliran model *Skip-gram* [10]. Di sana penulis memiliki tiga *layer*. Input *layer*, *hidden layer*, dan *output layer*. *Input layer* akan menerima kata yang ingin kita cari. Ini adalah matriks yang berisi semua kata yang berbeda di dalam *corpus*. Masing-masing bernilai nol, sedangkan kata yang dimasukkan bernilai satu. Kemudian di lapisan

tersembunyi, ada matriks yang berisi bobot untuk setiap kata pada setiap *neuron*. Secara *default*, ia memiliki 300 *neuron*. Matriks input akan dikalikan dengan *hidden layer* matrix untuk mendapatkan nilai representasi vektor dalam format matriks untuk selanjutnya akan diberikan ke *layer output*. Pada *layer output*, nilai matriks yang diterima akan diproses menggunakan *Softmax Regression Classifier*. Di mana setiap *neuron* akan memberikan nilai 0 hingga 1. Dan jumlah setiap neuron adalah 1. Nilai ini akan menjadi hasil dari jarak titik vektor dan kesamaan. Semakin tinggi nilainya, semakin dekat mereka dalam konteks yang sama.

C. Word Labelling

Dalam proses word labelling, kata-kata vektor dari hasil *Word2Vec* akan dihitung dan dinilai. Penilaian dilakukan dengan mencari kata kesamaan terlebih dahulu, kemudian menghitung skor total kata-kata serupa yang muncul dalam *tweet*.

1) *Finding Similar Words*
   Dalam proses ini, penulis memasukkan nama produk yang ingin penulis temukan, dan menerima kata-kata yang sama dari setiap produk. Kemudian, penulis mencari setiap *tweet*, jika *tweet* tersebut berisi kata-kata yang sama, dari *tweet* itu akan mendapatkan skor kata-kata yang sama terhadap produk
2) *Scoring Words*
   Kemudian, setelah setiap *tweet* dinilai, Skor dijumlahkan. Dan skor tertinggi dari suatu produk dalam *tweet* berarti bahwa *tweet* tersebut berbicara tentang produk. Lalu, setiap *tweet* pengguna dihitung lagi, dan semakin banyak *tweet* yang berbicara tentang produk yang sama, berarti pengguna itu tertarik dengan produk tersebut. Karena itu penulis dapat merekomendasikannya.

D. Output

Langkah Output merupakan langkah terakhir dalam system ini. Dimana penulis akan membuat chart dan rekomendasi dari hasil pada proses sebelumnya.

1) *Product Recommendation*
   Dari data minat pengguna, penulis kemudian akan membuat bagan yang menunjukkan pengguna mana yang tertarik pada produk mana. Dan di samping itu, penulis juga akan membuat bagan yang menunjukkan produk mana yang paling diminati pengguna yang memiliki lebih banyak pengikut. Pada dasarnya, *influencer* masing-masing produk.
2) *Product Popularity Chart*
   Selain grafik minat pengguna, dari langkah sebelumnya penulis dapat memberikan informasi tentang produk paling populer yang dibicarakan setiap akun. Dengan menghitung produk yang paling sering muncul di dalam *corpus*, penulis membuat grafik popularitas produk. Oleh karena itu, ini akan menjadi wawasan bagi perusahaan untuk menyesuaikan strategi pemasaran mereka.

III. HASIL DAN PEMBAHASAN

Bagian ini terbagi menjadi 2, yaitu pengkoleksian data dan pelabelan data. Setiap bagian akan dibahas masing-masing.

A. Pengoleksian Data

Menggunakan *crawler NodeJS*, penulis mengakses API *Twitter* dan mengambil *tweet* yang penulis butuhkan seperti yang dikatakan sebelumnya. Penulis mengumpulkan 600 *tweet* terbaru dari setiap akun yang mengikuti akun perusahaan mobil besar. Hasil sampel seperti yang ditunjukkan pada tabel I.

TABEL I
SAMPEL TWEET

No	User ID	Tweet
1	1016551316	RT @kompascom: Berikut cara dan syarat melaporkan kendaraan bermotor yang sudah dijual. https://t.co/R5IkfBIItr
2	895663296556056576	Jangan Kaget, Beli Mitsubishi Xpander di Tangerang Lebih Mahal Ketimbang di Jakarta https://t.co/s8gRNxmywI #nitnot… https://t.co/I8c9mIN6us
3	60518255	Suzuki All New Ertiga 2019 Upgrade Fitur, Seolah Jawab Kritik atas Kekurangan Ertiga Gen-2. Baca selengkapnya di… https://t.co/6yvE5jP2q6
4	514054021	Indonesia tidak kekurangan Orang Pintar ! Mungkin Tetapi Kekurangan Orang JUJUR !
5	78529472	Urinoir yang pake sensor adalah kutukan bagi Indonesia. Itu bersihin pake apaan? Daun kering?
6	1432575872	Kembali TOYOTA ASTRA MOTOR mempersembahkan NEW TOYOTA VELOZ 2019, yang merupakah salah satu mobil andalan TOYOTA ya… https://t.co/A3q5lZHDBJ
7	169048705	RT @CampusBoys1976: @fauzanoddie Oke sepakat, kami tidak mengajak debat. Hanya sama2 membuka pikiran. Betul, setuju sekali pak. Tapi menur…

Seperti yang terlihat pada tabel I, data dari *tweet* adalah acak, dan tidak selalu berbicara tentang mobil. Jadi penulis masih perlu memprosesnya secara menyeluruh. Dengan memproses data sebelumnya, penulis berhasil membersihkan data, sehingga memungkinkan untuk memprosesnya.

B. Pelabelan Data

Setelah mendapatkan data yang telah dibersihkan, penulis membuat vector kata menggunakan *Word2Vec*. Dan penulis menemukan 10 kata yang paling mirip dari dataset produk penulis. Kata-kata yang sama akan bertindak sebagai perwakilan produk. jadi ketika seseorang mengatakan kata yang sama, penulis menyimpulkan bahwa orang itu mengatakan tentang produk penulis. Hasil pembangunan vektor kata dapat dilihat pada gambar. 4

```
tweet_w2v.wv.most_similar("ayla")
```

```
C:\Users\ACER\Anaconda3\lib\site-
dtype from `int` to `np.signedint
  if np.issubdtype(vec.dtype, np.
```

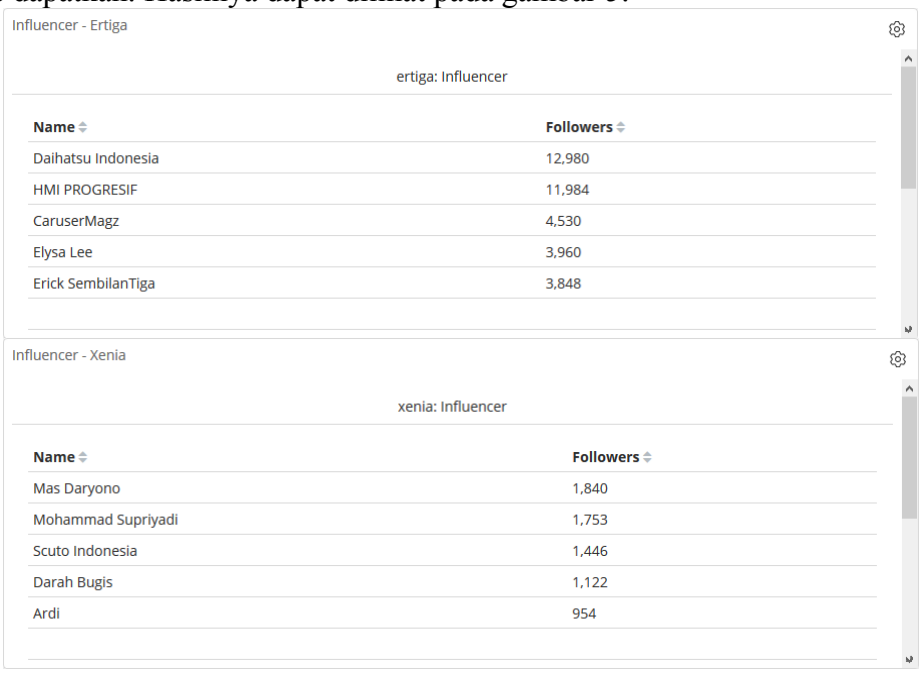
```
[('matic', 0.832058846950531),
 ('agya', 0.7822976112365723),
 ('benz', 0.7598838210105896),
 ('toyota', 0.7569708824157715),
 ('avanza', 0.7447090148925781),
 ('innova', 0.7393487095832825),
 ('xenia', 0.7262661457061768),
 ('dealer', 0.7255107760429382),
 ('trail', 0.720403790473938),
 ('yamaha', 0.7153176665306091)]
```

Gambar 4 Hasil Vektor Kata

Setelah mengetahui kata-kata yang serupa, langkah selanjutnya adalah menghitung skor setiap *tweet*, dan mencari produk yang paling sering dibicarakan. Hasil perhitungan akan menjadi dasar dari dua grafik berikutnya.

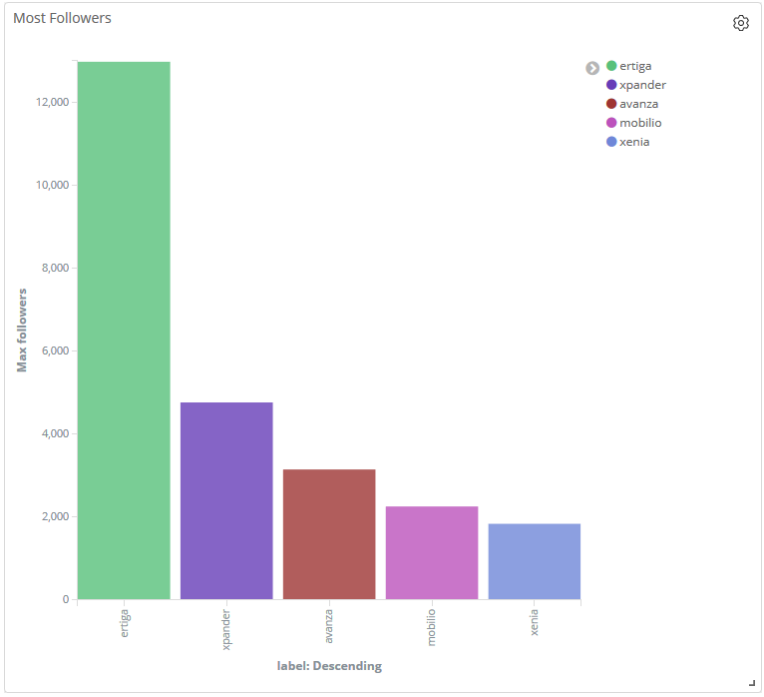
1) C*har*t Minat Pengguna

Bagan ini kana menampilkan informasi mengenai minat masing-masing pengguna yang penulis dapatkan. Hasilnya dapat dilihat pada gambar 5.



Gambar 5 Chart Minat Pengguna

Seperti yang ditunjukkan di atas, penulis memberi label pada masing-masing akun dengan jenis produk yang mereka sukai, dan mengurutkannya berdasarkan pengikut mereka. Semakin tinggi pengikut, berarti mereka lebih populer di *Twitter*. Oleh karena itu, preferensi mereka dapat menjadi peluang yang sangat baik bagi mereka untuk menjadi *influencer* produk pilihan mereka.

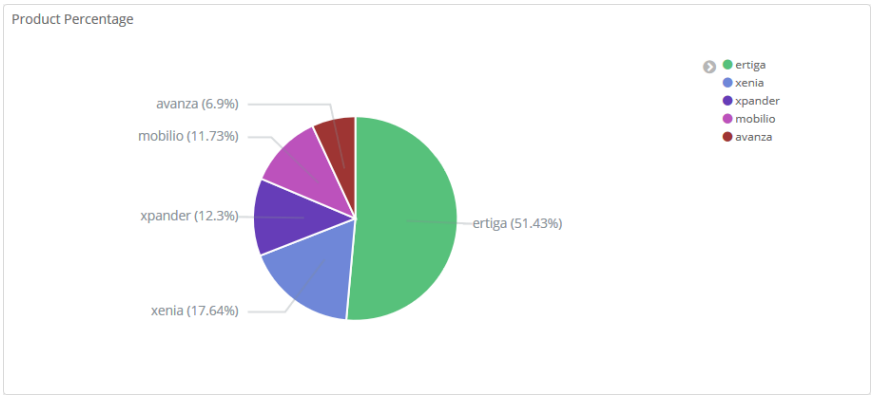


Gambar 6 Pengikut Terbanyak masing-masing Produk

Gambar 6 menunjukkan bagan yang menampilkan total pengikut terbanyak dari masing-masing produk. Nilai ini didapatkan dari total jumlah pengikut masing-masing pengguna yang memiliki preferensi terhadap produk yang sama.

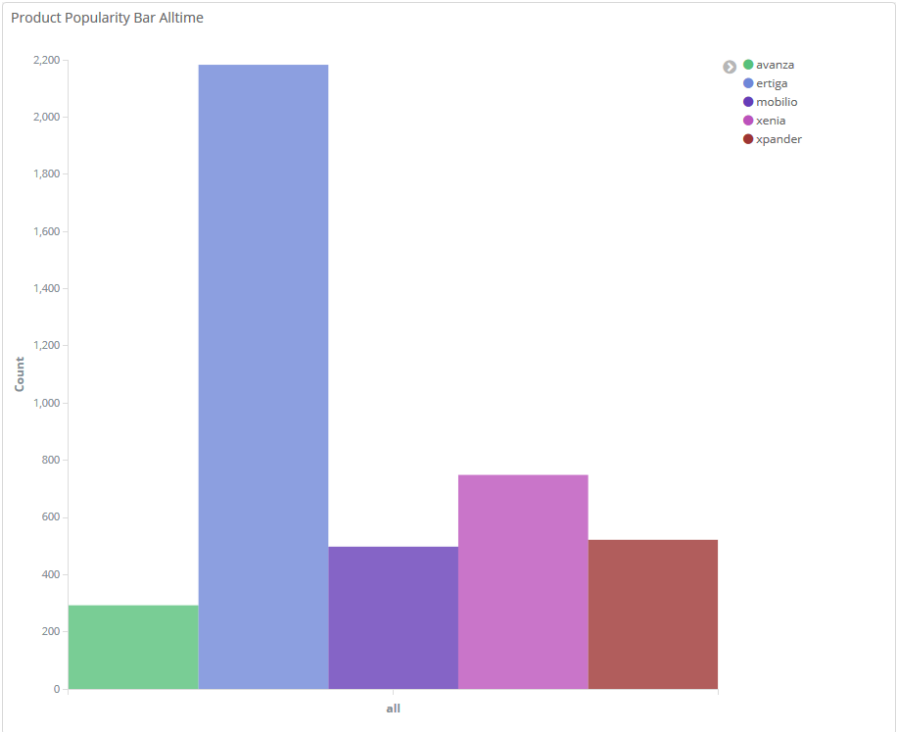
2) *Chart* Kepopuleran Produk

Selain *chart* minat pengguna, penulis juga membuat chart yang menampilkan tingkat kepopuleran untuk masing-masing produk yang ada. Hasilnya dapat dilihat pada gambar 7.

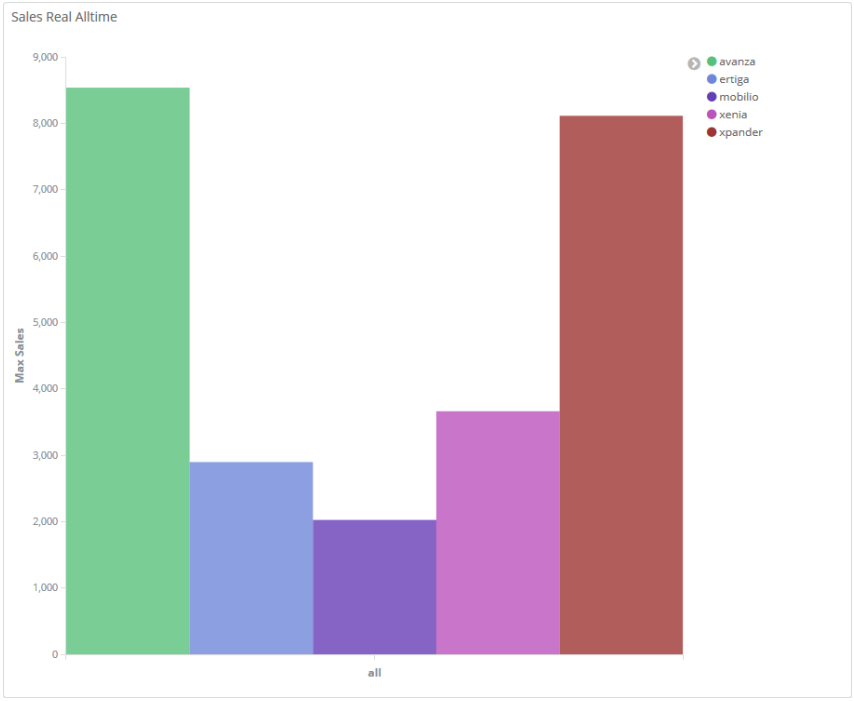


Gambar 7 Chart Kepopuleran Produk

Setelah mengkompilasi data yang memiliki label, penulis mendapati bahwa Suzuki Ertiga memimpin dalam popularitas media sosial, dengan 51,43% dari tweet membicarakannya. Sementara itu, Toyota Avanza adalah produk yang paling tidak populer di Twitter dengan hanya 6,9% dari *tweet* yang berbicara tentang Avanza. Namun, data ini tidak dapat diandalkan 100%, karena bagian terpenting dari pemasaran adalah penjualan. Jadi penulis membuat bagan yang membandingkan popularitas di media sosial dan laporan penjualan mereka, mencoba menemukan korelasi di antaranya. Hasilnya dapat dilihat pada gambar 8 dan 9.



Gambar 8 Tingkat Popularitas Produk



Gambar 9 Statsitik Penjualan Real

Dari dua grafik di atas, penulis menyimpulkan bahwa tidak ada korelasi antara popularitas media sosial dan penjualan pasar. Karena tidak semua orang yang berbicara tentang mobil mampu membeli sendiri. Jadi popularitas media sosial tidak selalu berarti bahwa penjualan akan mengikuti.

IV. KESIMPULAN

Setelah eksperimen selesai, kita dapat melihat bahwa algoritma ini telah berhasil menentukan produk mana yang diminati oleh sebuah akun, tetapi validitasnya belum valid karena data mentah yang penulis ambil begitu acak sehingga kata-kata yang serupa bahkan tidak dekat dengan artinya. Jadi masih perlu lebih banyak optimasi.

Dari data yang dikumpulkan, di media sosial Suzuki Ertiga memiliki popularitas terbesar, sementara Toyota Avanza adalah produk yang paling tidak populer. Sedangkan dalam statistik penjualan, Toyota Avanza adalah produk yang paling populer. Sementara Honda Mobilio adalah produk yang paling tidak populer. Ini berarti bahwa popularitas di media sosial tidak linier dengan statistik penjualan.

Selain validitas data, penelitian penulis juga menemukan bahwa Twitter bukanlah platform yang tepat untuk berbicara dan menjual mobil. Karena di Indonesia, kecenderungan seseorang mengulas masih rendah. Jadi hampir tidak ada tweet yang secara eksplisit berbicara tentang mobil. Dan akun perusahaan mobilsangat pasif, tanpa keterlibatan penting antara perusahaan dan para pengikutnya

Saran penulis untuk pengembangan ke deapnnya ialah berikan data yang baik dan mendukung sebagai data training dari algoritma ini, agar hasil ruang vektor yang didapat menjadiu lebih akurat.

REFERENSI

[1] Doshi, Zeel, et al, “TweetAnalyzer: Twitter Trend Detection and Visualization”, 2017.
[2] S.B.Japali, "Product Recommendation for the Day using Fuzzy c-means and Association Rule Generator in KNIME", 2017.
[3] V. M, "User Specific Product Recommendation and Rating System by Performing Sentiment Analysis on Product Reviews," 2017.
[4] Zhang, Jinming et al, “Combining Sentiment Analysis with a Fuzzy Kano Model for Product Aspect Preference Recommendation”, 2017
[5] Janjarassuk,Udon, “Product recommendation based on genetic algorithm”, 2019
[6] Mikolov, Thomas, et al, “Efficient Estimation of Word Representations in Vector Space”, 2013
[7] Mikolov, Thomas, et al, “Distributed Representations of Words and Phrases and their Compositionality”, 2013
[8] Y. Goldberg and O. Levy, "Word2Vec Explained: Deriving Mikolov et al's Negative Sampling Word Embedding Method," 2014.
[9] Le, Thu Anh, “An Exploration of the Word2Vec Algorithm: Creating a Vector Representation of a Language Vocabulary that Encodes Meaning and Usage Patterns in the Vector Space Structure”, 2016
[10] McCormick, Chris, “Word2Vec Tutorial – The Skip-Gram Model”, 2016