ANALISIS SENTIMEN *BRAND AMBASSADOR* BTS TERHADAP TOKOPEDIA MENGGUNAKAN KLASIFIKASI *BAYESIAN NETWORK* DENGAN EKSTRAKSI FITUR TF-IDF

Regina[1], Triando Hamonangan Saragih[2], Dwi Kartini[3]

[1,2,3,4,5] Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat
[1]renareginana@gmail.com, [2]triando.saragih@ulm.ac.id, [3]dwikartini@ulm.ac.id

Abstrak

BTS (Bangtan *Boys)* adalah salah satu *boy* grup asal korea selatan yang ditunjuk oleh Tokopedia untuk menjadi *brand ambassador* Tokopedia di Indonesia, BTS merupakan salah satu *boy* grup yang sangat mendunia dalam bidang musik dan Tokopedia merupakan salah satu *E-commerce* terkenal yang banyak digunakan oleh masyarakat indonesia untuk melakukan jual beli online. Kerjasama ini tentu saja memberikan pengaruh terhadap Tokopedia serta memperoleh banyak respon berupa opini masyarakat terutama pada media sosial *twitter*, karena hal tersebut maka dilakukan penelitian analisis sentiment. Data yang digunakan yaitu 900 data *tweet* dan terbagi menjadi 3 kelas yaitu positif, negatif, dan netral. Tahapan penilitian terdiri dari pengambilan dan pengumpulan data, *preprocessing* data, ekstraksi fitur dengan *Term Frequency - Inverse Document Frequency* (TF-IDF), klasifikasi dengan *Bayesian network*, evaluasi kinerja menggunakan *K-fold cross validation* (K-10) dan *confution matrix*. Perbandingan terjadi pada tahap *preprocessing* data, yaitu saat menggunakan normalisasi data dan tidak menggunakan normalisasi data, dari hasil perbandingan tersebut diperoleh nilai akurasi jika tidak menggunakan normalisasi data sebesar 66,6667%, presisi sebesar 68,1%, dan recall sebesar 66,7%. Sedangkan hasil akurasi dengan menggunakan normalisasi data sebesar 76,5556%, *presisi* sebesar 77,4%, dan *recall* sebesar 76,6%. Selisih nilai akurasi dari kedua percobaan sebesar 9,8889 %, hal ini membuktikan bahwa menggunakan normalisasi data lebih baik.

Kata kunci : *Sentiment analysis, Data Mining, Brand ambassador BTS, Tokopedia, TF-IDF, Bayesian Network.*

1. Pendahuluan

Di Indonesia saat ini *e-commerce* sudah berkembang sangat pesat. *E-commerce* merupakan transaksi jual beli yang dilakukan melalui media internet. Dari penelitian yang dilakukan oleh (Lailiya, 2020) menyatakan bahwa *e-commerce* dapat mempermudah dalam hal pelayanan dan memasarkan berbagai produk di dalam maupun di luar negeri menggunakan media sosial (internet) atau sering disebut dengan *online store.* Contoh *e-commerce* yang paling banyak dikunjungi adalah Tokopedia.

Pada Oktober 2019, BTS secara resmi ditunjuk sebagai *brand ambassador* Tokopedia selama satu tahun, kemudian pada 25 Januari 2021 tokopedia kembali menunjuk BTS untuk menjadi *brand ambassador* Tokopedia hingga saat ini. BTS, singkatan dari Bangtan Sonyeondan atau "*Beyond the Scene*", adalah sebuah boy grup yang berasal dari Korea Selatan dan memiliki jutaan penggemar dari seluruh dunia sejak debut mereka pada Juni 2013 hingga saat ini. BTS merupakan salah satu artist yang memiliki banyak prestasi yang mendunia. Pengaruh personality dari seorang *brand ambassador* akan sangat mempengaruhi *personality* dari sebuah *brand*. *personality* dari *brand ambassador* inilah yang nantinya akan mempengaruhi persepsi serta pemikiran dari masyarakat yang akan meningkatkan daya tarik terhadap citra merek dan dapat menarik konsumen untuk melakukan transaksi jual beli (Amalia Probosini et al., 2021).

Dikutip dari webside tokopedia (tokopedia.com) menyatakan bahwa pada tahun 2019, Tokopedia menerima penghargaan yaitu 'Fastest Value Growth' dalam BrandZ Top 50 Most Valuable Indonesian Brands, berdasarkan riset dari lembaga WPP dan Kantar. Menduduki peringkat 10 besar untuk pertama kalinya. Tokopedia diakui sebagai *brand* dengan pertumbuhan tercepat dan kenaikan nilai brand sebesar 487%. Dan hal tersebut terjadi saat tokopedia bekerjasama oleh BTS sebagai brand ambassador Tokopedia pada tahun 2019. Berdasarkan pernyataan sebelumnya menyebabkan banyaknya tanggapan dari fans maupun masyarakat terhadap Tokopedia baik itu positif, negatif maupun netral terutama pada sosial media twitter, maka dilakukanlah analisis sentimen mengenai tanggapan yang begitu banyak tersebut menggunakan salah satu metode klasifikasi dengan bantuan salah satu ekstraksi fitur text tweet pada twitter.

Pada saat proses pengambilan data melalui media sosial twitter setiap orang pasti menuliskan berbagai kata dengan penulisan yang unik, menggunakan singkatan kata, bahkan tidak jarang melakukan kesalahan penulisan pada kata (*typo*), berdasarkan pada fenomena tersebut maka peneliti melakukan tahap preprocessing data yang bertujuan untuk mengubah data awal menjadi data yang lebih baik dari sebelumnya. Tahapan *preprocessing* yang digunakan adalah *cleaning, case folding*, normalisasi data, *stopword removal* atau *filtering, stemming*, dan *tokenizing*. Pada penelitian ini akan dilakukan perbandingan pada tahapan normalisasi data yaitu saat menggunakan normalisasi data dan tanpa menggunakan normalisasi data.

Analisis Sentimen (opinion mining) adalah proses menganalisis sebuah text berbentuk digital yang prosesnya adalah mengekstrak, memahami, dan mengolah data secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam sebuah kalimat opini pada suatu dokumen. Penelitian analisis sentimen dilakukan untuk mengetahui apakah opini pada setiap masalah atau objek dari seseorang yang didapatkan cenderung positif atau negatif (Nur et al., 2019). Dengan adanya sistem analisis sentimen ini diharapkan dapat membantu suatu perusahaan untuk mengetahui dan mendapatkan umpan balik terhadap merk dagangnya dari masyarakat dalam menilai sebuah produk berdasarkan opini dan review yang telah didapatkan (Gunawan et al., 2018).

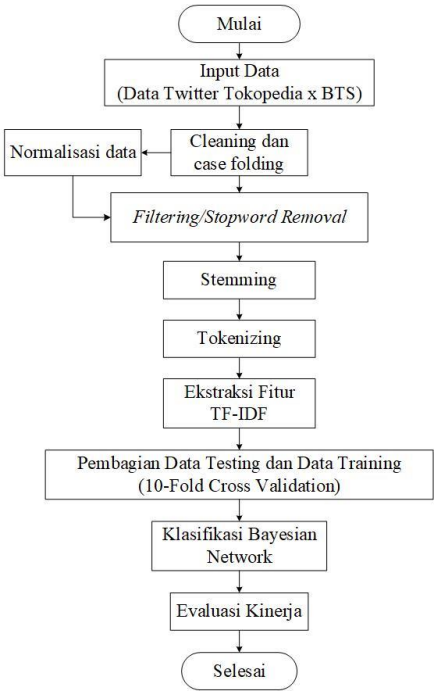
Penelitian analisis sentimen dapat dilakukan menggunakan beberapa metode klasifikasi salah satunya adalah *Bayesian Network*. Pada penelitian (Hasniati et al., 2019) menyatakan bahwa Bayesian Network dapat menunjukkan probabilitas hubungan antara beberapa kejadian yang saling berhubungan atau yang tidak berhubungan. Generalisasi Bayesian Network dapat mewakili dan memecahkan keputusan dibawah ketidakpastian yang disebut diagram pengaruh. Bayesian Network juga merupakan model grafis yang mengodekan hubungan probabilistik antara variabel - variabel yang menarik. Pada penelitian (Windarti, 2018) juga menyatakan secara umum bahwa metode *Bayesian Network* memiliki kinerja yang lebih baik dibandingkan dengan *naive bayes* yang berdasarkan nilai akurasi, presisi dan recall. Untuk kondisi *percentage split* 90, kedua algoritma memiliki nilai akurasi yang sama sebesar 80%, dan pada *percentage split* 80, *Bayesian Network* lebih baik dengan akurasi sebesar 75% sedangkan Naïve Bayes sebesar 70%.

Pada penelitian analisis sentimen selain menggunakan metode klasifikasi peneliti juga menggunakan ekstraksi fitur pada penelitian. Ekstraksi fitur merupakan metode yang digunakan untuk mengidentifikasi fitur yang relevan dengan klasifikasi. Pada penelitian (Mahfud, 2020) menyatakan keunggulan TF-IDF adalah memiliki hasil yang efisien, mudah dan akurat, serta hasil yang maksimal bila digunakan pada klasifikasi dengan beberapa kelas, pada penelitian ini TF-IDF dapat menghasilkan kinerja klasifikasi yang lebih tinggi dari yang lain karena dengan TF-IDF pembobotan lebih menggambarkan dokumen yang berisi teks dibandingkan dengan TFbinary.

Berdasarkan paparan pada latar belakang di atas, penelitian ini berfokus pada penggunaan tahapan preprocessing data terutama proses normalisasi data, metode klasifikasi menggunakan *Bayesian Networks* dengan ekstraksi fitur TF-IDF, sehingga tujuan dari penelitian ini yaitu dapat mengetahui berapa tingkat akurasi yang didapatkan dari hasil kombinasi klasifikasi *Bayesian Networks* dengan ekstraksi fitur TF-IDF saat menggunakan normalisasi data atau tanpa menggunakan normalisasi data, untuk evaluasi kinerja pada penelitian ini menggunakan *confution matrix* 3x3 karena memiliki 3 label kelas pada data, yaitu label positif, negatif, dan netral.

2. Metode

Pada penelitiani ini memiliki alur penelitian yang dimulai dengan melakukan pengambilan data pada media sosial twitter yang berhubungan dengan *brand ambassador* BTS (*event* tertentu) pada Tokopedia serta pelayanan pada perusahaan Tokopedia. Proses selanjutnya adalah preprocessing data, terdapat 6 proses yaitu *cleaning, case folding*, normalisasi data, *stopword removal* atau *filtering, stemming*, dan *tokenizing*. Perbandingan terjadi pada tahapan normalisasi data yaitu saat menggunakan normalisasi data dan tanpa menggunakan normalisasi data. Selanjutnya adalah ekstraksi fitur menggunakan metode TF-IDF (*Term Frequency- Inverse Document Frequency*), proses pembagian data *testing* dan *training* menggunakan *K-Fold cross validation* dengan K : 10, selanjutnya melakukan klasifikasi menggunakan metode *Bayesian network*, proses terakhir merupakan evaluasi kinerja menggunakan *confution matrix*. Untuk lebih jelasnya dapat dilihat pada gambar 1 :



Gambar 1. Alur penelitian

2.1 Pengumpulan Data

Data dikumpulkan secara manual dengan mengamati setiap komentar dari retweetan yang berhubungan dengan TokopediaXBTS, kemudian dilakukan pelabelan data, sehingga data dapat dibagi menjadi tiga kelas yaitu kelas positif, negatif dan netral. Setelah melakukan pengumpulan data maka dilakukan pelabelan data secara manual, setiap label memiliki ciri khusus untuk membedakan setiap label. Data berlabel positif memiliki ciri terdapat ungkapan persetujuan, pujian, kebahagiaan serta respon yang baik data berlabel negatif memiliki ciri terdapat ungkapan penolakan, amarah, kekecewaan, hujatan, respon yang tidak baik, sedangkan data netral memiliki ciri kalimat dapat mengandung kalimat tanya, atau mengandung dua makna positif dan negatif. Untuk contoh dataset awal lebih jelasnya dapat dilihat pada Tabel 1 :

Tabel 1. Contoh dataset awal

No	Tweet	Label
1	Yeayy kebagian!! Belanja di @tokopedia dapat Photocard BTS dan kebagian yang versi ✨ HOLOGRAM ✨ Terimakasih minto 😭 💜	Positif
2	Bismillah rezeki ya Allah mau pc ganteng taehyung plisss	Positif
3	Udh terlanjur kecewa ikutan rules pc Kya gini, co tepat waktu aja ga dapat dahlaa...	Negatif
4	Males ah min. Namjoon ama jin ga dapat kmaren. Mnding beli di toko oren aja dah.	Negatif
5	jd ini cepet cepetan klaim kan ya bukan cekout?	Netral
6	Inii beratii ga ada minimal harga barang yaa ?	Netral

2.2 *Preprocessing Data*

Preprocessing data merupakan proses merubah data mentah menjadi data yang dapat lebih mudah dipahami, singkatnya *preprocessing* data dilakukan untuk melakukan pembersihan serta perbaikan pada data yang telah dikumpulkan melalui media sosial twitter agar menjadi data yang lebih baik, dalam *preprocessing* data memiliki beberapa proses. Pada penelitian (Shevira et al., 2022) menyatakan bahwa tahapan *preprocessing* data sangat beragam tergantung pada kebutuhan penelitian yang dilakukan, namun secara umum diantaranya adalah cleaning tweet, lowering case (*case folding*), normalisasi, *stop-word*, *stemming*, dan *tokenizing* :

1. *cleaning* berfungsi untuk membersihkan dan menghilangkan karakter selain huruf, seperti berupa angka, username (@, RT, retweet, link, hastag (#), HTML, emoticon dan tanda baca lainnya seperti “,!$%&*”.
2. *case folding* berfungsi untuk mengubah semua kata (term) menjadi huruf kecil.
3. Normalisasi berfungsi untuk mengubah kata yang tidak tepat atau berlebihan menjadi kata yang lebih baik.
4. *stopword removal* atau *filtering* berfungsi untuk menghapus kata yang terlalu umum dan kurang penting.
5. *stimming*, berfungsi untuk mengubah kata imbuhan kembali menjadi kata dasar, bertujuan untuk memperkecil sebuah data.
6. *Tokenizing* berfungsi untuk pemenggal kalimat menjadi kata perkata.

2.3 TF-IDF

TF-IDF merupakan ekstraksi fitur yang sangat populer, metode ini bekerja menggunakan cara menghitung bobot setiap kata pada kalimat yang umum digunakan. Untuk menghitung bobot setiap kata dalam dokumen metode ini akan menghitung kemunculan sebuah kata dalam dokumen tersebut (Pratomo et al., 2021).

Nilai yang didapatkan dari proses pembobotan menggunakan TF-IDF akan menjadi semakin kecil karena pembobotan ini berdasarkan frekuensi dokumen terbalik, yang berarti adalah semakin banyak frekuensi kemunculan suatu kata (term) maka nilai pembobotannya akan semakin rendah, sedangkan jika frekuensi kemunculan kata (term) sedikit maka nilai pembobotannya akan semakin tinggi (Fadillah Grandis & Arumsari, 2021).

Berikut merupakan alur proses pembobotan dengan algoritma TF-IDF (Putra & Rochmawati, 2021)

1. Tahap pertama menghitung nilai TF yaitu total kemunculan kata dalam dokumen.

2. Tahap kedua menghitung nilai DF yaitu seluruh dokumen yang mengandung kata tersebut. Jumlah dokumen yang digunakan akan sangat memengaruhi nilai DF yang akan didapatkan.
3. Tahap ketiga menghitung nilai IDF yang merupakan nilai hasil kebalikan atau inverse dari nilai DF. Nilai IDF diperoleh menggunakan persamaan (1). Dimana N merupakan total keseluruhan dokumen yang digunakan pada penelitian.

$$IDF_{(t)} = Log2\,(\frac{N}{df_t}) \quad (1)$$

4. Tahap keempat menghitung TF-IDF, pada tahapan ini tinggal mengalikan nilai TF dan IDF menggunakan persamaan (2). Sehingga pada tahapan ini akan didapatkan nilai bobot untuk setiap kata (term) yang diambil dari hasil perkalian nilai TF dan IDF.

$$TF - IDF = tf \times Log2\,(\frac{N}{df_t}) \quad (2)$$

2.4 K-fold Cross Validation

K-fold cross validation biasanya digunakan untuk mengestimasi sebuah kesalahan prediksi dalam mengevaluasi kinerja suatu model. Data dibagi menjadi beberapa kelompok dan berjumlah hampir sama. Model dalam klasifikasi sebagai data latih (*training*) dan data uji (*testing*) sebanyak ”K”. Disetiap pengulangan, salah satu himpunan bagian akan digunakan sebagai data latih (*training*) dan data uji (*testing*) (Mardiana et al., 2022).

Pelatihan dan pengujian dilakukan sebanyak K kali. Pada proses pertama, kelompok (subset) S1 menjadi data uji (*testing*) dan kelompk (subset) lainnya menjadi data latih (*training*), pada percobaan kedua kelompok (subset) S1, S3,...Sk akan menjadi data latih dan S2 menjadi data uji, begitu setererusnya sampai kelompok (subset) ke Sk (Fitrianah et al., 2021).

2.5 Bayesian Network

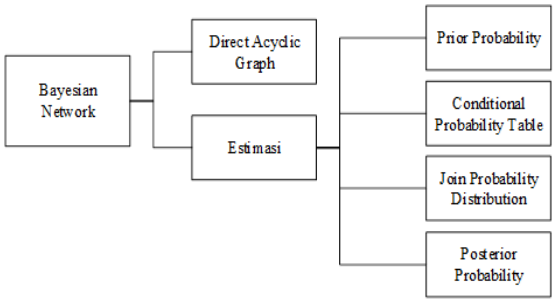
Bayesian Network merupakan metode pemodelan data yang berbasis probabilitas serta merepresentasikan suatu himpunan variabel dan atribut yang saling berhubungan atau berkorespondensi menggunakan *Directed Acyclic Graph* (DAG). *Bayesian Network* memiliki dua tugas, tugas pertama adalah pembelajaran melalui *Directed Acyclic Graph* (DAG) dan kedua struktur dari *Bayesian Network* merupakan jaringan. (Windarti & Suradi, 2019). Bayesian network didasari oleh teorema bayes. Dapat dilihat pada persamaan (3).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (3)$$

Bayesian network digambarkan sebagai graf yang terdiri dari busur (arc) dan simpul (*node*). Simpul (*node*) menunjukkan atribut atau variabel beserta nilai probabilitasnya dan busur menunjukkan hubungan antar simpul, di bawah ini merupakan langkah-langkah penerapan metode *Bayesian Network* (Windarti, 2018) :

1. Membuat struktur *Bayesian Network* atau *Directed Acyclic Graph* atau DAG
2. Mencari estimasi parameter yang pertama adalah mencari nilai *Perior Probability*
3. Mencari nilai *Conditional Probability Table* atau CPT
4. Mencari nilai *Joint Probability Distribution* atau JPD, untuk mendapatkan nilai *Joint Probability Distribution* adalah mengalikan nilai dari *Conditional Probability* dengan nilai *Prior Probability*.
5. Proses selanjutnya mencari nilai *Posterior Probabilistik* yang didapatkan dari hasil *Joint Probability Distribution* yang telah diperoleh.

Tahapan *Bayesian network* dapat dilihat pada gambar 2 (Suryana et al., 2018) :



Gambar 2. Struktur Bayesian Network

2.6 Evaluasi dengan Confusion Matrix

Pada penelitian (Ardiyansyah et al., 2018) menyatakan bahwa fungsi confusion matrix yaitu untuk mengukur tingkat nilai akurasi, presisi, dan recall, dari suatu model algoritma yang dievaluasi. Nilai akurasi adalah tingkat ketepatan presentase antara nilai prediksi dan nilai sebenarnya, sedangkan nilai presisi adalah nilai akurasi dengan kelas yang sudah diprediksi. Sedangkan nilai recall adalah presentase nilai kinerja keberhasilan algoritma yang digunakan.

Tabel pada *confusion matrix* menunjukkan jumlah data uji yang diklasifikasikan dengan benar dan jumlah data uji yang salah diklasifikasikan (Fikri et al., 2020). Seperti yang di sebutkan sebelumnya bahwa tabel confution matrix dapat membantu dalam proses pencarian nilai akurasi, presisi, dan recall dari sebuah penelitian. Pada jenis klasifikasi *multiclass* yang menggunakan 3 kelas, *confusion matrix* disajikan seperti pada tabel 1 :

Tabel 2. Confution Matrix Multiclass

		Aktual		
		Positif	Negatif	Netral
Prediksi	Positif	PP	PN	PR
	Negatif	NP	NN	NR
	Netral	RP	RN	RR

Dimana nilai Positif Positif (PP), Positif Negatif (PN), Positif Netral (PR), Negatif Positif (NP), Negatif Negatif (NN) Negatif Netral (NR), Netral Positif (RP), Netral Negatif (RN), Netral Netral (RR) dari tabel confution matrix ini dapat diperoleh nilai akurasi, presisi, dan recall. Nilai tersebut dihitung menggunakan persamaan (4), (5), dan (6).

$$A = \frac{PP + NN + RR}{PP + PN + PR + NP + NN + NR + RP + RN + RR} \times 100\% \quad (4)$$

$$P = \frac{\frac{PP}{PP + NP + RP} + \frac{NN}{PN + NN + RN} + \frac{RR}{PR + NR + RR}}{3} \times 100\% \quad (5)$$

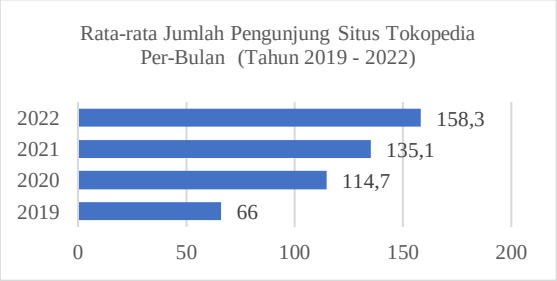
$$R = \frac{\frac{PP}{PP + PN + PR} + \frac{NN}{NP + NN + NR} + \frac{RR}{RP + RN + RR}}{3} \times 100\% \quad (6)$$

Berdasarkan persamaan di atas A merupakan akurasi, P merupakan presisi, R merupakan recall.

3. Hasil dan Pembahasan

Data yang digunakan pada penelitian ini berjumlah 900 data komentar maupun opini dari pengguna *twitter* mengenai pengaruh *brand ambassador* BTS terhadap Tokopedia. Data yang di ambil hanya dari tahun 2020 sampai dengan tahun 2021. Jumlah dataset yang digunakan pada setiap kelas bernilai sama (seimbang) yaitu 300 data positif, 300 data negatif, dan 300 data netral.

Dilansir dari databoks.katadata.co.id dapat dilihat Pengaruh *brand ambassador* BTS terhadap Tokopedia dari pertumbuhan pengunjung dari tahun 2019 hingga 2022. BTS memberikan pengaruh yang cukup besar terhadap Tokopedia berupa peningkatan pengunjung Tokopedia dari tahun 2019 sampai 2022. Tahun 2019 tokopedia memiliki pengunjung sebesar 66 juta orang, 2020 sebesar 114,7 juta orang, 2021 sebesar 135,1 juta orang, dan 2022 sebesar 158,3 juta orang. Untuk lebih jelasnya dapat dilihat pada gambar 3.



Gambar 3. pengunjung Tokopedia 2019-2022

Setelah melakukan pengumpulan data, tahapan selanjutnya adalah *preprocessing* data Tahap *preprocessing* memiliki beberapa proses. Pada penelitian ini, proses *preprocessing* yang digunakan adalah tahap *cleaning, case folding*, normalisasi, *stopword removal* atau *filtering, stimming,* dan *tokenizing*. Pada tahapan ini dilakukan perbandingan pada bagian normalisasi data yaitu ketika menggunakan normalisasi dan tanpa menggunakan normalisasi.

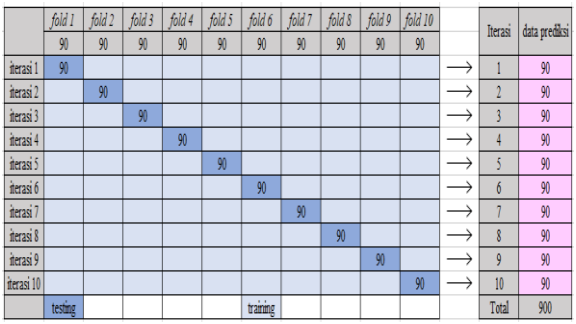
Proses normalisasi merupakan tahapan untuk mengubah kata non-baku menjadi baku, contoh kata "pengen" akan menjadi "ingin", kata "dgn" akan menjadi "dengan", kata "smuanya" akan menjadi "semuanya". Intinya adalah memperbaiki sebuah kata yang salah menjadi kata yang sebenarnya. Pada proses ini peneliti menggunakan kamus "slang" yang di buat khusus untuk memperbaiki data yang telah di dapatkan sebelumnya.

Tahapan selanjutnya setelah melakukan *preprocessing* data adalah melakukan ekstraksi fitur atau pembobotan setiap kata pada dataset, pada penelitian ini tahapan pembobotan kata menggunakan TF-IDF atau *Term Frequency - Inverse Document Frequency*. TF-IDF merupakan metode yang digunakan untuk menentukan nilai frekuensi pada setiap kata (*term*) dalam sebuah *document*. Rumjs yang digunakan yaitu rumus 1 dan 2. Pada proses ini, data yang tidak menggunakan normalisasi data menghasilkan 1.970 fitur sedangkan data yang menggunakan normalisasi data menghasilkan 1.171 fitur, untuk contoh hasil dari TF-IDF dapat dilihat pada Tabel 3 :

Tabel 3. Hasil perhitungan TF-IDF

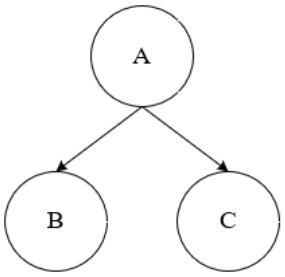
No	admin	…	sistem	…	youtube	label
1	0.2068	…	0.4992	…	0	Positif
…	…	…	…	…	…	…
301	0	…	0	…	0	Negatif
…	…	…	…	…	…	…
900	0.2844	…	0	…	0	Netral

Sebelum melakukan proses klasifikasi, data yang sudah melewati proses ini akan dibagi menjadi data *testing* dan data *training* dengan menggunakan *K-fold Cross Validation (K=10)*. Untuk lebih jelasnya dapat dilihat pada gambar 4.



Gambar 4. Ilustrasi 10-Fold Cross Validation

Tahapan selanjutnya yaitu melakukan klasifikasi pada data. Pada penelitia ini, metode klasifikasi yang digunakan adalah *Bayesian network*, metode ini didasari oleh metode *bayes.* Tahapan *Bayesian network* dibagi menjadi dua yaitu membuat *Diagram Acyclic Graph* (DAG) dan estimasi parameter. Untuk contoh DAG dapat dilihat pada gambar 5.



Gambar 5. Ilustrasi DAG (*Diagram Acyclic Graph*)

Setiap node mewakili sebuah variabel dan setiap garis menggambarkan hubungan ketergantungan antara 2 variabel. Berdasarkan gambar 5 kita dapat menyimpulkan bahwa terdapat garis A yang mengarah ke B, yang berarti A merupakan orang tua (*parents)* dari B sedangkan B merupakan anak (*child)* dari A. hal ini juga berlaku untuk A ke C yang berarti A merupakan oramgtua (*parents*) dari C sedangkan C merupakan anak (*child*) dari A. Tahapan selanjutnya adalah melakukan stimasi parameter. Pada tahapan ini, memiliki berapa proses yang harus dilakukan pertama mencari nilai *perior probability* dari setiap kelas (*class*), *conditional probability*, *join probability*, dan *posterior probability*.

Prior probability adalah nilai probabilitas yang di anggap benar sebelum melakukan penelitian terhadap data yang digunakan, apabila saat melakukan penelitian terjadi adanya perubahan atau perbaikan terhadap nilai probabilitas awal. Nilai *perior probability* di dapatkan dengan menggunakan persamaan (7).

$$P(A|B) = \frac{Jumlah\ data\ setiap\ class}{jumlah\ seluruh\ data} \quad (7)$$

Dibawah ini merupakan tabel hasil perhitungan *prior probability :*

Tabel 4. Hasil perhitungan *prior probability*

Positif	Negatif	Netral
0,33333333	0,33333333	0,33333333

Conditional probability adalah peluang bersyarat, atau nilai probabilitas suatu kejadian A apabila kejadian B sudah terjadi. Nilai conditional probability di dapatkan dengan menggunakan persamaan (8).

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \quad (8)$$

Dibawah ini merupakan tabel contoh hasil perhitungan *Conditional probability :*

Tabel 5. Hasil perhitungan *conditional probability*

No	Term	Positif	Negatif	Netral
1	Suga	0.5333333	0.3333333	0.1
2	Taehyung	0.6333333	0.3	0.2
3	Alhamdulillah	0.3333333	0.0333333	0.0333333
4	Pulang	2.2333333	0.2666667	0.1
5	Admin	4.233333	3.7	5.8
6	Terimakasih	2.1	0.0333333	0.1666667

Joint probability adalah probabilitas semua kejadian terjadi secara bersamaan. Cara mencari nilai *join probability* adalah mengalikan nilai *conditional probability* dengan nilai *prior probability* seperti persamaan 9.

$$P(X \cap Y) = P(X|Y)P(Y) \quad (9)$$

Dibawah ini merupakan tabel contoh hasil perhitungan *Joint probability :*

Tabel 6. Hasil perhitungan *prior probability*

No	Term	Positif	Negatif	Netral
1	Suga	0.1777778	0.1111111	0.0333333
2	Taehyung	0.2111111	0.1	0.0666667
3	Alhamdulillah	0.1111111	0.0111111	0.0111111
4	Pulang	0.7444444	0.0888889	0.0333333
5	Admin	1.4111111	1.2333333	1.9333333
6	Terimakasih	0.7	0.0111111	0.0555556

Posterior probability merupakan probabilitas yang direvisi atau diperbaiki dengan menggunakan informasi tambahan. Pada proses ini, peneliti dapat mengetahui setiap kalimat benar bernilai positif, negatif, atau netral. Nilai *posterior probability* dapat dihasilkan dengan menggunakan nilai hasil dari *joint probability* yang sudah diperoleh sebelumnya. Nilai ini didapatkan dengan menggunakan persamaan 10.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X|Y)P(Y) + P(X|Y^C)P(Y^C)} \quad (10)$$

Dibawah ini merupakan tabel contoh hasil perhitungan *posterior probability :*

Tabel 7. Hasil perhitungan *posterior probability*

No	Term	Positif	Negatif	Netral

1	Suga	0.5517241	0.3448275	0.1034482
2	Taehyung	0.5588235	0.2647058	0.1764705
3	Alhamdulillah	0.8333333	0.0833333	0.0833333
4	Pulang	0.8589743	0.1025641	0.0384615
5	Admin	0.3082524	0.2694174	0.4223300
6	Terimakasih	0.9130434	0.0144927	0.0724637
Hasil Kali Probabilitas		0.0621145	3.04619E-06	1.79066E-06

Dari hasil total perkalian setiap *class* di atas dapat diketahui bahwa hasil akhir perkalian pada *class* negatif lebih besar dari positif dan netral. Maka kalimat "suga taehyung alhamdulillah pulang admin terimakasih" berlabel negatif dan dengan prediksi di awal sentimen yaitu positif. Maka kalimat ini bernilai Positif Negatif (PN).

Untuk menghitung akurasi, *presisi*, *recall* dan diperlukan tabel *confusion matrix*, untuk lebih jelasnya dapat dilihat pada tabel 8 yang merupakan hasil *confution matrix* tanpa menggunakan normalisasi, sedangkan tabel 9 hasil *confution matrix* dengan menggunakan normalisasi data.

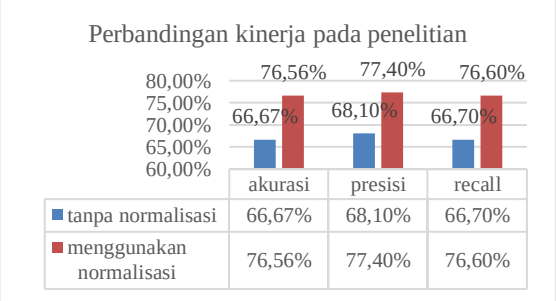
Tabel 8. Hasil *Confution Matrix* Tanpa Normalisasi Data

		Aktual		
		Positif	Negatif	Netral
Prediksi	Positif	180	51	69
	Negatif	28	191	81
	Netral	22	49	229

Tabel 9. Hasil *Confution Matrix* Menggunakan Normalisasi Data

		Aktual		
		Positif	Negatif	Netral
Prediksi	Positif	211	34	55
	Negatif	24	222	54
	Netral	17	27	256

Untuk mengetahui nilai perbandingan dari penggunaan normalisasi data dan tanpa menggunakan normalisasi data, maka diperlukan nilai akurasi, presisi, dan recall dengan menghitung nilai yang sudah didapatkan dari kedua tabel confution matrix sebelumnya, nilai didapatkan menggunakan persamaan (4), (5), dan (6). Setelah dilakukan perhitungan maka akan didapatkan hasil perbandingan evalusi kinerja seperti pada gambar 6.



Gambar 6. Perbandingan Hasil Kinerja

Dari penelitian ini dapat diketahui bahwa dataset yang melewati tahapan *preprocessing* dengan normalisasi data dan menggunakan kombinasi klasifikasi *Bayesian network* dengan metode pembobotan *Term frequency-inverse document frequency* (TF-IDF) menghasilkan hasil kinerja yang lebih baik dibandingkan tanpa menggunakan proses normalisais data. Dari gambar sebelumnya dapat diartikan bahwa perbedaan nilai akurasi cukup besar yaitu 9,89%. Sehingga dapat diketahui bahwa menggunakan normalisasi data pada tahap *preprocessing* lebih baik dibandingkan tanpa menggunakan normalisasi data. Untuk lebih jelasnya dapat dilihat pada tabel 10.

Tabel 10. Hasil Perbandingan Evaluasi Kinerja

Hasil Kinerja	Akurasi	Presisi	Recall
Tanpa Normalisasi Data	66,6667%	68,1%.	66,7%.
Mengunakan Normalisasi Data	76,5556%	77,4%.	76,6%.

4. Kesimpulan

Pada penelitian ini, membuktikan bahwa brand ambassador BTS memiliki pengaruh yang besar terhadap Tokopedia berupa peningkatan pengunjung semenjak BTS diangkat menjadi *brand ambassador* Tokopedia dari tahun 2019 sampai 2022 dan membuktikan bahwa kombinasi metode klasifikasi *Bayesian network* dengan ekstraksi fitur *Term frequency-inverse document frequency* (TF-IDF) dengan menggunakan normalisasi data pada tahapan *preprocessing* menghasilkan nilai akurasi sebesar 76,5556%, *presisi* sebesar 77,4%, dan *recall* sebesar 76,6%, sehingga selisih nilai akurasi dari kedua percobaan sebesar 9,8889 %, hal ini jelas membuktikan bahwa menggunakan normalisasi data lebih baik.

Adapun saran-saran yang dapat diberikan berdasaran penelitian ini adalah yaitu pada penelitian selanjutnya disarankan untuk menggunakan data yang tidak seimbang dengan jumlah data setiap class lebih besar seperti 500 data text dan menggunakan metode ekstraksi fitur lain seperti Bag Of Word.

Selain itu disarankan untuk melakukan perbandingan metode klasifikasi maupun ektraksi fiturnya. Peneliti dapat menggunakan 1 ekstraksi fitur yang dikombinasikan dengan 2 metode klasifikasi atau sebaliknya. Seperti ekstraksi fitur menggunakan Term frequency-inverse document frequency (TF-IDF) dikombinasikan dengan 2 metode klasifikasi Bayesian network dengan Random forest.

DAFTAR PUSTAKA:

Amalia Probosini, D., Hidayat, N., & Yusuf, M. (2021), *Pengaruh Promosi dan Brand Ambassador terhadap Keputusan Pembelian Pengguna Market Place X dengan Brand Image sebagai Variabel Intervening*, *Jurnal Bisnis, Manajemen, Dan Keuangan*, *2*(2), 445–458.

Ardiyansyah, Rahayuningsih, P. A., & Maulana, R. (2018), *Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner*, *Jurnal Khatulistiwa Informatika*, *VI*(1), 20–28.

Fadillah Grandis, G., & Arumsari, Y. (2021), *Seleksi Fitur Gain Ratio pada Analisis Sentimen Kebijakan Pemerintah Mengenai Pembelajaran Jarak Jauh dengan K-Nearest Neighbor* (Vol. 5, Issue 8).

Fikri, M. I., Sabrila, T. S., Azhar, Y., & Malang, U. M. (2020), *Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter*.

Fitrianah, D., Dwiasnati, S., H, H. H., & Baihaqi, K. A. (2021), P*enerapan Metode Machine Learning untuk Prediksi Nasabah Potensial menggunakan Algoritma Klasifikasi Naïve Bayes*, *Faktor Exacta*, *14*(2), 92.

Gunawan, B., Pratiwi, H. S., & Pratama, E. E. (2018), *Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes*, *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *4*(2), 113.

Hasniati, H., Arianti, A., & Philip, W. (2019), *Penerapan Metode Bayesian Network Model Pada Sistem Diagnosa Penyakit Sesak Nafas Bayi*, *IKRA-ITH INFORMATIKA : Jurnal Komputer Dan Informatika*, *3*(2), 19–26.

Lailiya, N. (2020), *IQTISHADequity Prodi S1 Manajemen, Fakultas Ekonomi dan Bisnis PENGARUH BRAND AMBASSADOR DAN KEPERCAYAAN TERHADAP KEPUTUSAN PEMBELIAN DI TOKOPEDIA* (Vol. 2, Issue 2). ONLINE.

Mahfud, F. K. R. (2020), *Sentiment Analysis of Perpustakaan Nasional Republik Indonesia Through Social Media Twitter*, *MATICS*, *12*(1), 90.

Mardiana, L., Kusnandar, D., & Satyahadewi, N. (2022), *ANALISIS DISKRIMINAN DENGAN K FOLD CROSS VALIDATION UNTUK KLASIFIKASI KUALITAS AIR DI KOTA PONTIANAK*, In *Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster)* (Vol. 11, Issue 1).

Nur, M., Utomo, Y., Negeri, P., & Pandang, U. (2019), *Analisis Sentimen pada Twitter terhadap Pelayanan Pemerintah Kota Makassar*.

Pratomo, S. A., Al Faraby, S., & Purbolaksono, M. D. (2021), *Analisis Sentimen Pengaruh Kombinasi Ekstraksi Fitur TF-IDF dan Lexicon Pada Ulasan Film Menggunakan Metode KNN*.

Putra, B. G., & Rochmawati, N. (2021), *Klasifikasi Berdasarkan Question Dalam Stack Overflow Menggunakan Algoritma Naïve Bayes*, *Journal of Informatics and Computer Science*, *02*.

Shevira, S., Made, I., Suarjaya, A. D., & Wira Buana, P. (2022), *Pengaruh Kombinasi dan Urutan Pre-Processing pada Tweets Bahasa Indonesia*, In *JITTER-Jurnal Ilmiah Teknologi dan Komputer* (Vol. 3, Issue 2).

Suryana, I., Suryani, M., Paulus, E., Rosadi, R., & Syahfitri, D. (2018), Jurnal Euclid. In *Jurnal Euclid* (Vol. 5, Issue 2).

Windarti, M. (2018), *Perbandingan Kinerja Algoritma Naïve Bayes Dan Bayesian Network Dalam Klasifikasi Masa Studi Mahasiswa*, *Prosiding Seminar Nasional Aplikasi Sains & Teknologi*, *September*, 249–261.

Windarti, M., & Suradi, A. (2019), *PERBANDINGAN KINERJA 6 ALGORITME KLASIFIKASI DATA MINING UNTUK PREDIKSI MASA STUDI MAHASISWA* (Vol. 1, Issue 1).