Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor

Jeremy Andre Septian, Tresna Maulana Fahrudin, Aryo Nugroho
Fakultas Ilmu Komputer, Universitas Narotama

***Abstrak*— Persepakbolaan Indonesia belakangan ini memiliki banyak polemik mulai dari kasus pengaturan skor, pergantian pelatih timnas senior hingga pergantian ketua umum Persatuan Sepak bola Seluruh Indonesia (PSSI). Polemik ini menimbulkan banyaknya opini maupun pendapat dari pengguna *twitter* terhadap persepakbolaan di Indonesia sehingga diperlukan sebuah sistem untuk memudahkan dalam mengetahui sentimen pada setiap kalimat. Tujuan dari penelitian ini adalah untuk menganalisis sentimen pada setiap kalimat dari pengguna *twitter* terhadap persepakbolaan Indonesia apakah memiliki sentimen negatif atau positif menggunakan *K-Nearest Neighbor*. Data yang digunakan dalam penelitian ini didapatkan dari hasil *crawling* dari media sosial *twitter* terkait persepakbolaan di Indonesia yang diambil dari akun *twitter* resmi PSSI. Setelah data dikumpulkan kemudian akan dilakukan beberapa tahapan yaitu *preprocessing* yang terdiri dari *cleansing, tokenizing, stopword removal,* dan *stemming.* Pembobotan kata menggunakan *Term Frequency-Invers Document Frequency (*TF-IDF*).* Pada tahap validasi data dilakukan pengujian silang sebanyak 10 kali menggunakan *k-fold cross validation,* kemudian diklasifikasikan dengan metode *K-Nearest Neighbor* dapat menghasilkan akurasi yang cukup baik. Dari 2000 data *tweet* berbahasa indonesia didapatkan hasil akurasi optimal pada nilai k=23 sejumlah 79.99%.**

Kata Kunci*— Persepakbolaan Indonesia, Analisis sentimen, *Preprocessing*, TF-IDF, *K-Nearest Neighbor

I. PENDAHULUAN

Persepakbolaan di Indonesia dalam beberapa tahun terakhir ini sedang mengalami banyak polemik seperti contohnya pada kasus pengaturan skor, pergantian pelatih timnas senior, dan pergantian ketua umum PSSI. Dengan adanya polemik yang terjadi dalam beberapa bulan terakhir ini banyak menimbulkan opini maupun pendapat dari para pengguna media sosial salah satunya adalah *twitter*. *Twitter* merupakan layanan media sosial yang berkategori *microblogging* yang paling populer, pengguna dapat membaca dan berbagi pesan singkat dengan jumlah maksimal 280 karakter [1],[2]. Berdasarkan permasalahan tersebut maka untuk memudahkan dalam mengetahui sentimen dari setiap opini maupun pendapat pada setiap kalimat perlu dibuatkan suatu sistem untuk menganalisis sentimen. Analisis sentimen adalah suatu proses untuk mengetahui pendapat atau opini pada sebuah kalimat maupun dokumen apakah memiliki sentimen negatif atau positif [3]. Dalam akun *twitter* resmi PSSI ini berisi banyak opini maupun pendapat dari pengguna *twitter* terkait tentang polemik yang terjadi pada persepakbolaan Indonesia sehingga dapat digunakan sebagai sumber data untuk melakukan penelitian terkait analisis sentimen [4]. Data yang digunakan dalam penelitian ini sejumlah 2000 data *tweet* berbahasa indonesia yang akan dilakukan pembobotan kata dengan TF-IDF dan diklasifikasi menggunakan *K-Nearest Neighbor*. Kinerja *K-Nearest Neighbor* sebagai algoritma untuk klasifikasi data teks cukup bagus [5].

Penelitian terdahulu terkait analisis sentimen dengan *K-Nearest Neighbor* pernah dilakukan untuk menganalisis sentimen dari data komentar akun *facebook* jasa ekspedisi barang J&T pada tahun 2018 dan mendapatkan akurasi tertinggi sebesar 79.21% [6]. Penelitian selanjutnya *K-Nearest Neighbor* digunakan untuk mengkasifikasikan dokumen komentar pada situs *youtube* didapatkan akurasi tertinggi sebesar 80.6% [7].

Kontribusi pada penelitian ini adalah untuk dapat memberikan gambaran umum terkait dengan adanya polemik yang terjadi pada persepakbolaan Indonesia, dimana data *tweet* diambil sebagai sampel untuk mewakili saran dan opini yang disampaikan oleh masyarakat.

II. TINJAUAN PUSTAKA

A. Analisis Sentimen

Analisis sentimen atau penambangan opini adalah suatu bidang studi untuk menganalisis pendapat orang terhadap entitas seperti produk, layanan,organisasi, individu, masalah, peristiwa, dan topik. Analisis sentimen ini berfokus pada pendapat seseorang yang mengekspresikan atau menyiratkan sentimen positif atau negatif, kebanyakan analis sentimen ini berkaitan dengan orang-orang di media sosial. Faktanya, analisis sentimen sekarang berada di pusat penelitian media sosial. Oleh karena itu, penelitian dalam analisis sentimen tidak hanya memiliki dampak penting pada NLP (*Natural Language Processing*), tetapi mungkin juga memiliki dampak mendalam pada ilmu manajemen, politik, ekonomi, dan ilmu sosial karena mereka semua dipengaruhi oleh pendapat orang. Memperoleh pendapat dari publik dan konsumen telah lama menjadi bisnis besar dalam bidang pemasaran, hubungan masyarakat, dan

Jeremy Andre Septian, *Fakultas Ilmu Komputer, Universitas Narotama*, Surabaya, Jawa Timur, Indonesia (e-mail: jeremyandreseptian@gmail.com)
Tresna Maulana Fahrudin, *Fakultas Ilmu Komputer, Universitas Narotama*, Surabaya, Jawa Timur, Indonesia (e-mail: tresna.maulana@narotama.ac.id)
Aryo Nugroho *Fakultas Ilmu Komputer, Universitas Narotama*, Surabaya, Jawa Timur, Indonesia (e-mail: aryo.nugroho@narotama.ac.id)

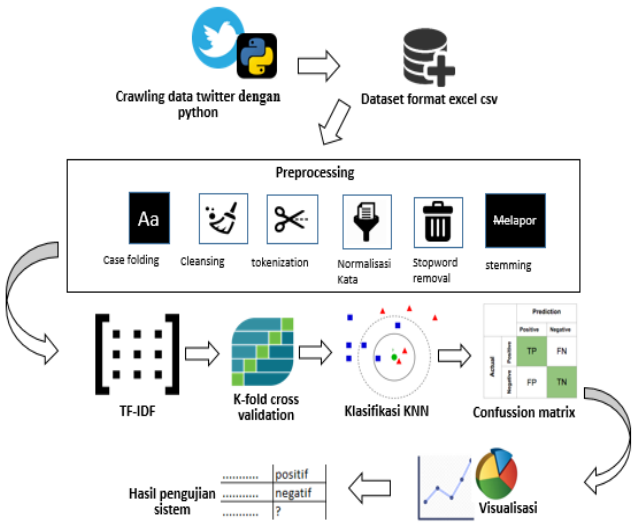
perusahaan kampanye politik Besarnya pengaruh dan manfaat dari analisis sentimen menyebabkan penelitian ataupun aplikasi mengenai analisis sentimen berkembang pesat, bahkan di Amerika kurang lebih 20-30 perusahaan yang memfokuskan pada layanan analisis sentimen [8].

B. *Pembobotan TF-IDF (Term Frequency-Inverse Document Frequency)*

Pembobotan TF-IDF (*Term Frequency-Inverse Document Frequency)* adalah suatu proses untuk melakukan transformasi data dari data tekstual ke dalam data numerik untuk dilakukan pembobotan pada tiap kata atau fitur. TF-IDF ini adalah sebuah ukuran statistik yang digunakan untuk mengevaluasi seberapa penting sebuah kata di dalam sebuah dokumen. TF adalah frekuensi kemunculan kata pada di tiap dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam tiap dokumen tersebut. DF adalah frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. IDF adalah *inverse* dari nilai DF. Hasil dari pembobtan kata menggunakan TF-IDF ini adalah hasil perkalian dari TF dikalikan dengan IDF. Bobot kata semakin besar jika sering muncul dalam suatu dokumen dan semakin kecil jika muncul dalam banyak dokumen [9].

C. *K-Nearest Neighbor*

K-Nearest Neighbor (KNN) adalah algoritma klasifikasi *supervised learning* atau biasa dikenal dengan metode berbasis jarak. Metode ini bekerja dengan hanya menghafal semua contoh pelatihan yang tersedia selama fase pelatihan. Selanjutnya pada fase pengujian, dokumen yang akan diklasifikasi dibandingkan dengan contoh-contoh berdasarkan ukuran jarak yang ditentukan sebelumnya. Dokumen yang paling mirip disebut "tetangga terdekat" untuk jumlah tetangga terdekat dapat ditentukan dari berapa jumlah nilai k. KNN ini merupakan metode yang sederhana untuk pengklasifikasian tanpa harus melakukan perhitungan secara kompleks, oleh sebab itu KNN ini juga biasa disebut *lazy learning.* Untuk menghitung jarak antar fitur KNN ini menggunakan perhitungan *Euclidean Distance* [10].



Gambar. 1. Desain sistem analisis sentimen terhadap peresepakbolaan Indonesia

III. METODE PENELITIAN

Dalam penelitian ini terdapat alur proses penelitian untuk dapat melakukan analisis sentimen yaitu dimulai dengan pengumpulan data, *preprocessing*, pembobotan kata dengan TF-IDF,validasi dengan *k-fold cross validation*, klasifikasi *K-Nearest Neighbor,* penghitungan akurasi dengan *confussion matrix*, visualisasi dan pengujian sistem. Seperti ditunjukan pada Gambar 1.

A. *Pengumpulan data*

Pada penelitian ini data diambil dari media sosial *twitter* dengan program *crawling* menggunakan bahasa pemrograman *python* dengan *library tweepy* yang memanfaatkan *API Twitter*. Langkah awal dalam tahap pengumpulan data ini adalah membuat aplikasi *API Twitter* melalui akun *twitter* pribadi untuk mendapatkan *consumer key*, *consumer secret*, *access token*, dan *acess token secret* yang digunakan untuk syarat kunci pengambilan data. Kemudian memasukan *query* '@pssi' nantinya pada *query* ini memuat semua *tweet* yang ditujukan kepada akun *twitter* PSSI dan menggunakan filter tambahan *no-retweet*. Periode pengambilan data pada *twitter* dalam peneltian adalah 21 februari – 04 Mei 2019, seluruh data hasil *crawling* disimpan kedalam format excel csv (*comma separated value*). Dataset yang digunakan dalam penelitian ini adalah *tweet* dan sentimen. 0 untuk merepresentasikan sentimen negatif dan 1 untuk sentimen positif. Untuk lebih jelas, dapat diperhatikan Tabel 1.

TABEL I
DATA TWEET PENGGUNA TWITTER TERKAIT PERSEPAKBOLAAN INDONESIA

Tweet	Sentimen
@PSSI Semangat garudaku,,, jangan minder lihat lawan, jangan anggap remeh juga dengan lawan.. KAMU PASTI BISA...DOAKU UNTUK KALIAN	1
@PSSI Indonesia mesti menang lawan Mafia Bola Garuda Pancasila. Akulah pendukungmu#TimnasIndonesiaMaju @KEMENPORA_RI @PSSI #AFFU22 https://t.co/9WloaWpfD5	1
Bila kelak Indonesia lolos ke final AFF U22 2019, itu tandanya kita sedang mulai menuju prestasi sepakbola yang kita dambakan. #TimnasIndonesiaMaju @KEMENPORA_RI @PSSI	1
@PSSI bukan hari yang kurang baik tapi emnk Kualitas kalian cuman segitu!!! nyadar diri lah BGST!!!'	0
@PSSI Berat utk lolos	0
@PSSI mencium aroma kekalahan	0
@PSSI Federasi BUSUK,timnas jeblok\n#busuk\n#busuk\n#busuk'	0
@PSSI Organisasinya induknya aja asal2an ya maen bola ya asal2an aja'	0
@PSSI timnas bobrok karena federasi korup n tidak becus'	0

B. *Preprocessing*

Data yang sudah dikumpulkan kemudian dilakukan *preprocessing* untuk menghindari data yang belum siap untuk diolah seperti terdapat gangguan (*noise*) dan data yang tidak konsisten [11]. Adapun tahapan pada *preprocessing* ini adalah sebagai berikut:

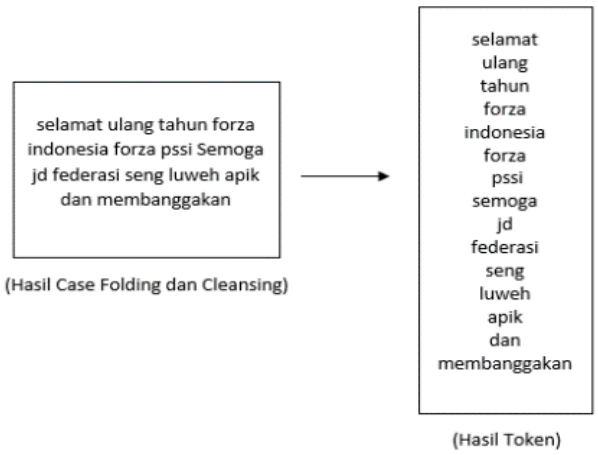
1. Case Folding dan Cleansing

Pada tahapan ini dilakukan penyeragaman seluruh teks menjadi huruf kecil (*lowercase*) dan pembersihan atau penghapusan pada pada semua dokumen yang berisi angka, url (http://) ,*username* (@) ,tanda pagar (#) ,delimiter seperti koma (,) dan titik (.) dan juga tanda baca lainya. Pada Gambar 2 diilustrasikan proses *case folding* dan *cleansing*.

Gambar. 2. Proses *Case Folding* dan *Cleansing*

2. Tokenizing

Pada tahapan ini dilakukan proses pemotongan pada sebuah dokumen ataupun kalimat menjadi kata atau biasa disebut token. Pada Gambar 3 diilustrasikan proses *tokenizing*.



Gambar. 3. Proses *Tokenizing*

3. Normalisasi Kata

Pada tahapan ini dilakukan proses untuk menormalisasi kata terhadap setiap kata-kata yang mengandung kata tidak baku maupun *noise* menjadi kata yang baku dan siap diolah. Kata tidak baku dan adanya *noise* yang dimaksud adalah kata yang mengandung unsur bahasa daerah yang tidak sesuai dengan Kamus Besar Bahasa Indoesia (KBBI) dan kata singkatan dalam media sosial. Seperti contoh kata singkatan yangs ering muncul pada media sosial ini adalah kata "tdk" dengan proses normalisasi kata ini diubah menjadi kata "tidak", kata yang mengandung unsur bahasa daerah seperti kata "apik" dengan proses normalisasi kata ini diubah menjadi "baik", kata tidak baku seperti kata "endonesah" dengan proses normalisasi kata ini diubah menjadi kata "Indonesia". Untuk lebiih jelas dapat diperhatikan Gambar 4 ilustrasi proses Normalisasi kata.

4. Stopword Removal

Stopword removal adalah suatu proses untuk menghapus kata yang dianggap tidak penting seperti contoh kata 'di',

Gambar. 4. Proses Normalisasi Kata

'yang', 'dan', 'ke', dan semua kata yang terdapat dalam kamus *stopword* yang sudah dibuat. Tujuan dari proses ini adalah untuk mengurangi jumlah kata yang disimpan dalam daftar token yang nantinya akan dilakukan proses selanjutnya. Pada Gambar 5 diilustrasikan proses *stopword removal*.

Gambar. 5. Proses *Stopword Removal*

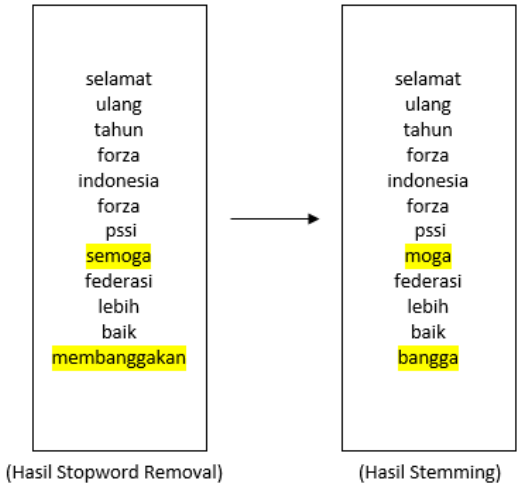
5. Stemming

Pada tahapan terakhir pada *preproceesing* ini dilakukan proses untuk mengubah semua kata-kata pada dokumen menjadi kata dasar dengan menghilangkan semua kata imbuhan. Kata imbuhan yang dihilangkan terdiri dari awalan (prefix), akhiran (suffix), sisipan (infix), dan gabungan awalan-akhiran (confix). Pada penelitian ini menggunakan *library python* sastrawi untuk proses *stemming*. Pada Gambar 6 diilustrasikan proses *stemming*.

C. Pembobotan Kata dengan TF-IDF

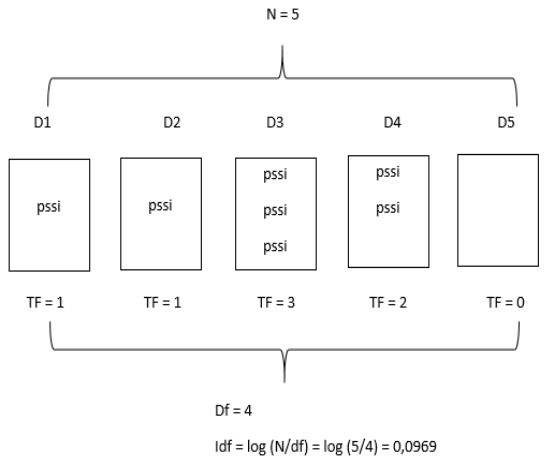
Data hasil *preprocessing* yang berupa kata akan diubah ke dalam bentuk angka dengan dilakukan proses pembobotan kata yang bertujuan untuk menghitung bobot pada masing-masing kata yang akan digunakan sebagai fitur, semakin banyak dokumen yang akan diproses maka semakin banyak fitur. Pada tahapan ini terdapat dua bagian proses yaitu TF (*Term Frequency*) dan IDF (*Inverse Document Frequency*), TF adalah jumlah kemunculan tiap

kata pada sebuah dokumen semakin banyak kata muncul pada tiap dokumen maka semakin besar nilai TF. IDF



Gambar. 6. Proses *Stemming*

adalah jumlah nilai dokumen pada tiap kata yang berbanding terbalik yaitu apabila suatu kata jarang muncul pada sebuah dokumen maka nilai IDF lebih besar daripada kata yang sering muncul [12],[13].



Gambar. 7. Ilustrasi Algoritma TF-IDF

Keterangan:
D1,..,D5 = Dokumen
TF = Jumlah kata pada tiap dokumen
N = Total dokumen
Df= Jumlah dokumen pada kata yang dicari

Adapun rumus dari pembobotan kata TD-IDF adalah:

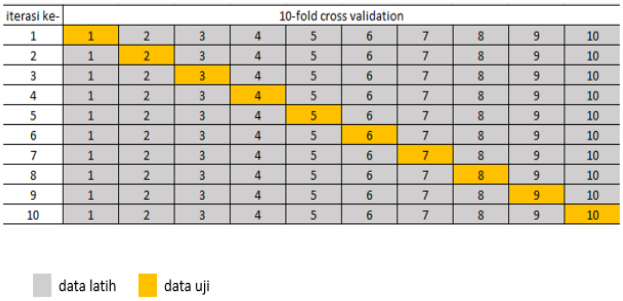
$$W_{t,d} = tf_{t,d} \text{ x } idf_t = tf_{t,d} \text{ x } \log \frac{N}{df_t} \quad (1)$$

Keterangan (1):
$W_{t,d}$ = Bobot TF-IDF
$tf_{t,d}$ = Jumlah frekuensi kata
idf_t = Jumlah *inverse* frekuensi dokumen tiap kata
df_t = Jumlah frekuensi dokumen tiap kata
N = Jumlah total dokumen

Hasil dari pembobotan kata dengan TF-IDF ini adalah perkalian dari nilai TF dan IDF yang akan menghasilkan bobot lebih kecil apabila kata tersebut sering muncul pada setiap dokumen dalam koleksi, sebaliknya bobot TF-IDF akan lebih besar apabila kata tersebut jarang muncul pada setiap dokumen dalam koleksi. Dalam penelitian ini pembobotan TF-IDF yang digunakan adalah TF-IDF tanpa normalisasi.

D. Validasi dengan k-fold cross validation

Data yang sudah dilakukan pembobotan kata dengan TF-IDF akan dilakukan validasi menggunakan *k-fold cross validation*. Cara kerjanya adalah dilakukan pengelompokan antara data latih dan data uji kemudian dilakukan pengujian yang diulang sebanyak jumlah k. Dalam penelitian ini k yang digunakan adalah 10-*fold* yang berarti akan dilakukan 10 kali pengujian pada seluruh isi dokumen secara acak. Dimana nantinya hasil akurasi didapatkan dari rata-rata akurasi pada 10 kali hasil pengujian [14],[15].



Gambar. 8. Ilustrasi tabel 10-*fold cross validation*

E. Klasifikasi K-Nearest Neighbor

K-Nearest Neighbor (KNN) adalah salah satu algoritma klasifikasi *supervised learning* yang digunakan untuk mengkasifikasikan objek berdasarkan atribut kelas dan data *training* [16]. Cara kerja KNN ini adalah melakukan klasifikasi berdasarkan data latih yang jaraknya paling dekat dengan objek data uji. Setiap data pada data latih memiliki atribut kelas, selanjutnya adalah dengan menguji model dengan data uji yang tidak memilki atribut kelas.Adapun tahapan proses pada KNN ini adalah sebagai berikut:

1. Tahapan awal pada KNN ini adalah menentukan nilai K, misal k=23 artinya 23 dokumen terdekat dengan dokumen uji yang akan diambil.
2. Menghitung jarak antara data baru di setiap label data (jarak *euclidean*) dengan jarak semua data *training*. Untuk menghitung tingkat kesamaan dalam dokumen ini menggunakan *Euclidean distance*. Cara perhitungan *euclidean distance* adalah dengan rumus berikut.

$$D(X,Y) = \sqrt{\sum_{k=1}^{n}(Xk - Yk)^2} \quad (2)$$

Keterangan (2):
D = Jarak antara dua titik x dan y
X = Data uji
Y = Sampel data
n = Dimensi data

3. Kemudian urutkan hasil euclidean distance berdasarkan jarak K yang ditentukan jika K=3 artinya akan dipilih 3 jarak terkecil dari hasil euclidean distance.

4. Selanjutnya gunakan mayoritas atribut kelas pada 3 tetangga terdekat yang sudah dipilih untuk menentukan prediksi kelas pada data baru tersebut. Misalkan pada 3 tetangga terdekat memilki 2 atribut kelas positif dan 1 atribut kelas negatif maka kelas pada data baru teresebut adalah positif.

F. Penghitungan akurasi menggunakan Confusion Matrix

Setelah dilakukan proses klasifikasi dengan menggunakan *K-Nearest Neighbor* selanjutnya akan dilakukan proses penghitungan akurasi menggunakan *confusion matrix*. Langkah pengujian untuk menghitung akurasi dengan *confusion matrix* adalah sebagai berikut [6]:

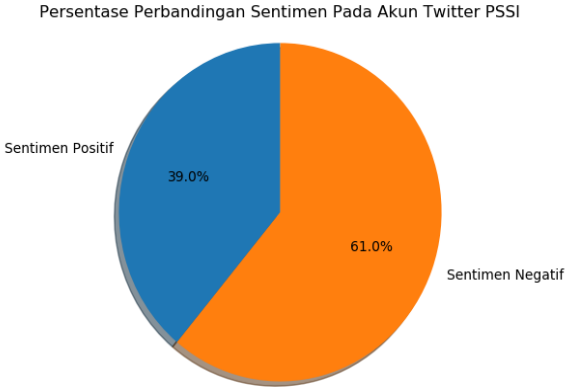
1. Menghitung jumlah hasil data asli positif dan data klasifikasi positif (TP).
2. Menghitung jumlah hasil data asli positif dan data klasifikasi negatif (FN).
3. Menghitung jumlah hasil data asli negatif dan data klasifikasi negatif (TN).
4. Menghitung jumlah hasil data asli negatif dan data klasifikasi positif (FP).
5. Setelah diketahui jumlah pada masing-masing langkah 1-4, kemudian jumlahkan TP dan TN
6. Selanjutnya jumlahkan semua TP, FN, TN dan FP.
7. Lakukan pembagian dari langkah 5 dan 6 kemudian hasil bagi dikalikan 100.

Adapun rumus untuk perhitungan *confusion matrix* untuk mencari akurasi adalah sebagai berikut :

$$\text{Akurasi} = \frac{TP + TN}{TP+TN+FP+FN} \qquad (3)$$

IV. HASIL DAN PEMBAHASAN

Data pada penelitian ini diambil dari media sosial *twitter* memanfaatkan *Twitter API* dengan menggunakan *query* '@PSSI'. Periode pengumpulan data dilakukan pada 21 februari – 04 Mei 2019 kemudian data dilakukan sorting untuk setiap *tweet* yang memiliki sentimen yang akan ditambahkan ke dalam dataset yang digunakan untuk penelitian analisis sentimen ini yaitu sejumlah 2000 *tweet*. Kemudian data tersebut diberi label sentimen negatif dan positif, data diharuskan mempunyai label dikarenakan dalam penelitian analisis sentimen dengan metode *K-Nearest Neighbor* ini merupakan *supervised learning*. Dari 2000 *tweet* terdapat 790 *tweet* positif dan 1210 *tweet* negatif , sehingga dapat disimpulkan bahwa dari data yang diambil dari akun *twitter* pssi ini cenderung negatif. Adapun proporsi data yang digunakan pada penelitian ini seperti ditunjukan pada Gambar 9.



Gambar. 9. Diagram pie persentase sentimen positif dan negatif dari dataset

Berdasarkan Gambar 9, data dengan sentimen negatif lebih banyak dibandingkan data dengan sentimen positif dikarenakan pada periode pengumpulan data 21 februari – 04 Mei 2019 ditemukan lebih banyak opini dar pengguna twitter terhadap polemik yang terjadi pada persepakbolaan Indonesia. Setelah data dilakukan *preprocessing* dari 2000 *tweet* ditemukan 5843 kata yang siap untuk diolah. Setelah kata hasil dari *preprocessing* dikumpulkan selanjutnya akan dilakukan penghitungan untuk pembobotan kata dengan TF-IDF seperti pada Tabel 2.

TABEL II
PEMBOBOTAN KATA DENGAN TF-IDF

D1 = selamat ulang tahun forza indonesia forza pssi moga federasi lebih baik bangga						
D2 = selamat pssi menang						
Term	D1	D2	DF	IDF	**TFIDF**	
					D1	D2
selamat	1	1	2	0,1761	0,1761	0,1761
ulang	1	0	1	0,4771	0,4771	0
tahun	1	0	1	0,4771	0,4771	0
forza	2	0	2	0,1761	0,3522	0
indonesia	1	0	1	0,4771	0,4771	0
pssi	1	1	2	0,1761	0,1761	0,1761
moga	1	0	1	0,4771	0,4771	0
federasi	1	0	1	0,4771	0,4771	0
lebih	1	0	1	0,4771	0,4771	0
baik	1	0	1	0,4771	0,4771	0
bangga	1	0	1	0,4771	0,4771	0
menang	0	1	1	0,4771	0	0,4771

Pada Tabel 2 dapat diperhatikan cara penghitungan untuk mencari bobot pada tiap dokumen, dengan menghitung jumlah kata yang muncul pada sebuah dokumen (TF) kemudian jumlah dokumen yang memiliki kata (DF). Setelah nilai DF didapatkan kemudian hitung nilai IDF dengan rumus log=N/df, dimana N merupakan jumlah seluruh dokumen yang ada.

TABEL III
AKURASI NILAI K

Nilai K	Akurasi %	Nilai K	Akurasi %
23	79.99	15	78.65
19	79.89	11	77.39
21	79.75	9	76.89
25	79.75	7	76.20
27	79.75	5	74.34
29	79.6	1	67.00
17	79.54	3	65.75
13	79.00		

Setelah pembobotan kata kemudian dilakukan validasi dengan *k-fold cross validation* menggunakan 10 *fold*, klasifikasi KNN, dan penghitungan akurasi dengan *confussion matrix*. Dilakukan pengujian untuk mencari akurasi terbaik pada nilai k=1, k=3, k=5, k=7, k=9, k=11, k=13, k=15, k=17, k=19, k=21, dan k=23, k=25, k=27, k=29. Untuk lebiih jelas, dapat dperhatikan Tabel 3.

Berdasarkan Tabel 3, akurasi nilai K didapatkan hasil optimal nilai K pada K=23 dengan akurasi yang didapatkan adalah 79.99%. Dari hasil nilai akurasi mulai dari K=1 hingga K=19 semakin banyak nilai k maka semakin tinggi juga akurasinya, namun pada K=21 hingga K=29 tidak selalu semakin tinggi nilai K semakin tinggi akurasinya. Jumlah akurasi dan nilai K juga dipengaruhi dari banyaknya dokumen jika semakin banyak dokumen maka untuk mendapatkan nilai akurasi K optimal perlu semakin banyak nilai K nya. Hal tersebut dikarenakan semakin banyak dokumen akan semakin banyak pula fitur atau jumlah kata. Gambar 10 dapat diperhatikan untuk proses dari validasi menggunakan *10-fold cross validation* dan evaluasi *confusion matrix* pada nilai K=23.

```
K = 23
   Accuration  Error Rate      Confussion Matrix
0       0.635       0.365    [[34, 6], [67, 93]]
1       0.735       0.265   [[78, 16], [37, 69]]
2       0.775       0.225  [[109, 20], [25, 46]]
3       0.835       0.165  [[150, 19], [14, 17]]
4       0.835       0.165   [[156, 5], [28, 11]]
5       0.865       0.135   [[123, 6], [21, 50]]
6       0.880       0.120  [[158, 11], [13, 18]]
7       0.850       0.150    [[91, 8], [22, 79]]
8       0.795       0.205    [[76, 9], [32, 83]]
9       0.795       0.205  [[120, 15], [26, 39]]
Accuracy : 0.7999999999999999
Average error rate : 0.2
```

Gambar. 10. Hasil Validasi dan Evaluasi K=23

Gambar 10 menunjukan proses validasi menggunakan *10-fold cross validation* pada nilai K=23 dimana dilakukan proses validasi silang dengan 10 kali iterasi. Setiap iterasi atau *fold* berisi 1800 *tweet* yang digunakan sebagai data latih dan 200 *tweet* yang digunakan sebagai data uji/validasi yang diklasifikasikan dengan KNN. Kemudian untuk menghitung hasil akurasi dari klasifikasi KNN maka akan dilakukan penghitungan pada tabel *confusion matrix*.

TABEL IV
CONFUSION MATRIX

Aktual	Klasifikasi KNN	
	Positif	Negatif
Positif	1095	115
Negatif	285	505

Tabel 4 menunjukan tabel *confusion matrix* dari nilai K=23 yang sudah dilakukan validasi menggunakan *10-fold cross validation* dengan porsi setiap *fold* sebanyak 200 data sehingga menghasilkan 2000 data uji validasi untuk *10-fold*. Adapun uraian hasil proses pembentukan *confusion matrix* sebagai berikut:

1. Hasil aktual positif yang diprediksi positif oleh KNN (TP) sebanyak data 1095 data.
2. Hasil aktual positif yang diprediksi negatif oleh KNN (FN) sebanyak 115 data.
3. Hasil aktual negatif yang diprediksi positif oleh KNN (FP) sebanyak 285 data.
4. Hasil aktual negatif yang diprediksi negatif oleh KNN (TN) sebanyak 505 data.

Penghitungan dari tabel *Confussion Matrix* total dari *10-fold cross validation* untuk mencari akurasi dan *error rate* adalah.

$$\text{Akurasi} = \frac{\text{TP+ TN}}{\text{TP+TN+FP+FN}} = \frac{1095 + 505}{1095+505+285+115} = 0{,}80$$

$$\textit{Error Rate} = \frac{\text{FP+ FN}}{\text{TP+TN+FP+FN}} = \frac{285 + 115}{1095+505+285+115} = 0{,}20$$

Akurasi adalah tingkat keberhasilan klasifikasi untuk memprediksi data aktual yang terdiri dari data dengan label positif dan negatif. *Error rate* adalah tingkat kesalahan klasifikasidalam memprediksi data aktual.

Dilakukan pengujian sistem dengan nilai K=23, pengujian ini diperlukan untuk membuktikan bahwa sistem dapat menganalisis sentimen dengan baik. Dengan dimasukannya tweet baru hasil *crawling* dari *twitter* pada akaun *twitter* resmi PSSI terkait topik persepakbolaan Indonesia untuk dilakukan analisis sentimen secara otomatis adapun hasil pengujian dapat diperhatikan pada Tabel 5.

TABEL V
HASIL PENGUJIAN SISTEM

No.	Tweet	Aktual	Prediksi
1	Congratss utk semua pemain,coach @indra_sjafri Dan staff dan utk @PSSI Juga kalian pantas di apresiasi https://t.co/rKCpfkxMuo'	Positif	Positif
2	Semoga semakin Jaya dan Berprestasi.\n\n#PSSI https://t.co/0ZTEIUdBv'	Positif	Positif
3	@PSSI Optimis Indonesia mendapat pelajaran berharga dari Pertandingan semalam	Positif	Positif
4	walau kalah kalian tetap membanggakan @PSSI	Positif	Negatif
5	@PSSI udahlah mau ada mafia mau enggak, emang pemaen timnas indo jelek jelek. percuma juga kalo ngarep prestasi cuma gara gara berantas mafia wkwkwk.	Negatif	Negatif
6	paling kalah lagi wkwk udah jadi tradisi tiap tahunnya kalah @PSSI	Negatif	Negatif
7	@PSSI Pengurus pssi bobrok prestasi timnas amburadul'	Negatif	Negatif
8	Banyak kasus, organisasi gajelas, gapunya uang. Sosoan pgn ngehibur. Nyadar ga sih masyarakat udah jijik liat kelakuan kalian? @PSSI https://t.co/ITbmPDLI8y'	Negatif	Negatif
9	@PSSI kita lemahnya di finising,crosing,taktikal dll.kita hrs banyk latihan suting krn bnyk melebar, melambung ,crosing bnyk tdk tpt sasaran maka dari itu kt hrs latihan yg keras semuaitu pasti ada ilmu nya , knp tdk tpt,melambung	Negatif	Negatif
10	Clear bisa bersihin ketombe sama bisa bersihin orang" busuk di @PSSI gak??!!	Negatif	Negatif

Berdasarkan Tabel 5, dapat dibuktikan bahwa sistem terbukti dapat menganalisis sentimen dengan baik. Dari 10 *tweet* baru tanpa label yang dilakukan pengujian terdapat 1 *tweet* yang tidak sesuai prediksinya.

V. KESIMPULAN DAN SARAN

Berdasarkan penelitian yang telah dilakukan, maka diperoleh kesimpulan bahwa:

1. Dari 2000 data tweet terkait polemik persepakbolaan Indonesia yang diambil dari akun twitter PSSI dengan proporsi 790 tweet positif dan 1210 tweet negatif, dilakukan percobaan untuk mencari model KNN dengan akurasi terbaik menggunakan range nilai k=1 hingga k=30 yang adalah bilangan ganjil, didapatkan akurasi optimal pada k=23 dengan akurasi sebesar 79,99% dan error rate sebesar 20,01%.
2. Analisis sentimen pengguna twitter terhadap topik sepak bola Indonesia menggunakan pembobotan TF-IDF dan metode KNN berhasil dilakukan. Dengan dibuktikan saat pengujian model KNN dengan diberikan nilai K=23, dari 10 *tweet* baru yang dilakukan pengujian untuk mendapatkan sentimen hanya 1 *tweet* yang tidak sesuai prediksinya.

DAFTAR PUSTAKA

[1] R. Hidayatillah, M. Mirwan, M. Hakam, and A. Nugroho, "Levels of Political Participation Based on Naive Bayes Classifier," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 13, no. 1, p. 73, 2019.

[2] R. A. Setiawan and D. B. Setyohadi, "Analisis Komunikasi Sosial Media Twitter sebagai Saluran Layanan Pelanggan Provider Internet dan Seluler di Indonesia," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 3, no. 1, p. 16, 2017.

[3] B. J. M. Putra, A. Helen, and A. R. Barakbah, "Rule-based Sentiment Degree Measurement of Opinion Mining of Community Participatory in the Government of Surabaya," *Emit. Int. J. Eng. Technol.*, vol. 6, no. 2, p. 200, 2018.

[4] C. R. . Vinodhini, G, "Sentiment Analysis and Opinion Mining: A Survey," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 2, no. 6, 2012.

[5] R. R. Sani, J. Zeniarza, and A. Luthfiarta, "Pengembangan Aplikasi Penentuan Tema Tugas Akhir Berdasarkan Data Abstrak Menggunakan Algoritma K-nearest Neighbor," *Proceeding SENDI_U*, no. 207, pp. 103–111, 2016.

[6] A. Salam, J. Zeniarja, and R. S. U. Khasanah, "Analisis Sentimen Data Komentar Sosial Media Facebook Dengan K-Nearest Neighbor (Studi Kasus Pada Akun Jasa," *Pros. SINTAK*, pp. 480–486, 2018.

[7] N. MOH, "Klasifikasi Dokumen Komentar Pada Situs Youtube Menggunakan Algoritma K-Nearest Neighbor (K-Nn)," *Univ. Dian Nuswantoro*, no. 5, 2016.

[8] B. Liu, "Sentiment Analysis and Opinion Mining Morgan & Claypool Publishers," *Lang. Arts Discip.*, no. May, p. 167, 2012.

[9] V. Amrizal, "Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim)," *J. Tek. Inform.*, vol. 11, no. 2, pp. 149–164, 2019.

[10] A. Lüschow and C. Wartena, "Classifying medical literature using k-nearest-neighbours algorithm," *CEUR Workshop Proc.*, vol. 1937, no. Ddc, pp. 26–38, 2017.

[11] I. Hemalatha, G. . S. Varma, and A. Govardhan, "Preprocessing the Informal Text for Efficient Sentiment Analysis," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 1, no. 2, pp. 58–61, 2012.

[12] T. Winarti, J. Kerami, and S. Arief, "Determining Term on Text Document Clustering using Algorithm of Enhanced Confix Stripping Stemming," *Int. J. Comput. Appl.*, vol. 157, no. 9, pp. 8–13, 2017.

[13] M. Subhan, A. Sudarsono, and A. R. Barakbah, "Classification of Radical Web Content in Indonesia using Web Content Mining and k-Nearest Neighbor Algorithm," *Emit. Int. J. Eng. Technol.*, vol. 5, no. 2, p. 328, 2018.

[14] S. E. Syahfitri Kartika Lidya, Opim Salim Sitompul, "Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (Svm)," *Semin. Nas. Apl. Teknol. Inf. 2015 (SNATI 2015)*, vol., no., pp. 1–8, 2015.

[15] C. K. Park and D. G. Kim, "Historical background," *Curr. Futur. Manag. Brain Metastasis*, vol. 25, pp. 1–12, 2012.

[16] S. Ernawati and R. Wati, "Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel," *J. Khatulistiwa Inform.*, vol. VI, no. 1, 2018.

Jeremy Andre Septian. Mahasiswa semester 8 jurusan Teknik Informatika, Universitas Narotama, Surabaya. Dia bekerja sebagai admin di salah satu perusahaan *trading* besi beton di Surabaya. Minat penelitian di bidang *text minning* dan *data minning*.

Tresna Maulana Fahrudin. Menyelesaikan jenjang D4 dan S2 di Politeknik Elektronika Negeri Surabaya di Program Studi Teknik Informatika. Saat ini berkarir sebagai Dosen Tetap Fakultas Ilmu Komputer, Universitas Narotama Surabaya. Minat bidang penelitiannya adalah *Data Mining*, *Machine Learning*, *Metaheuristic* dan *Text Mining.*

Aryo Nugroho. Menyelesaikan jenjang S1 dibidang Teknik Sipil dan Teknik Informatika kemudian menyelesaikan S2 di ITS. Berkarir sebagai Dosen Tetap di Fakultas Ilmu Komputer, Universitas Narotama Surabaya. Sejak 2014, telah mengambil Program Doktor di Pascasarjana Teknik Elektro ITS. Minat penelitiannya di bidang *Text Mining, IT Strategic, Artificial Intelligent,* dan *Business Intelligence.*