J M S K
Jurnal Matematika, Statistika, & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
Vol. 17, No. 1, 26-41 September, 2020
doi: 10.20956/jmsk.v%vi%i.9947

p-ISSN: 1858-1382
e-ISSN: 2614-8811

Sentiment Analysis of *Southeast Asian Games* (SEA Games) in Philippines 2019 Based on Opinion of Internet User of Social Media *Twitter* with *K-Nearest Neighbor* and *Support Vector Machine*

Analisis Sentimen terhadap *Southeast Asian Games* (SEA Games) di Filipina Tahun 2019 berdasarkan Opini Netizen dari Media Sosial *Twitter* dengan Metode *K-nearest Neighbor* dan *Support Vector Machine*

Muhammad Riefky[1*], Wara Pramesti[2*]

Abstract

Sports events are an activity that is in great demand, especially the people of Southeast Asia. One of the most prestigious sporting events in the Southeast Asian region is the Southeast Asian Games (SEA Games). SEA Games is one of the sporting events held in the Southeast Asia region and is only held every two years involving eleven member countries of the Association of South East Asian Nations (ASEAN). The most SEA Games issues occurred on Twitter with 20,600 tweets. This is because the 2019 SEA Games event in the Philippines experienced many irregularities, one of which is the Rizal Memorium stadium, which has not been renovated until now. The purpose of this study is to obtain and compare the results of the accuracy of the classification of Twitter users' sentiments towards the 2019 SEA Games in the Philippines using k-nearest neighbor and support vector machine. The data used in this study comes from data from Twitter social media users who often use the hashtag "SEA Games 2019" which has been done with text preprocessing of 2697 tweets with data partitions of 60% for training data and 40% for testing data. The conclusion that can be drawn from this research is that the best accuracy results in the k-nearest neighbor and support vector machine classification are the support vector machine classification with a polynomial kernel of 92.96% so that the predictions of the Support Vector Machine classification tend to be negative.

Keywords: *K-nearest Neighbor, SEA Games, Sentiment, Support Vector Machine*.

Abstrak

Acara olahraga merupakan suatu kegiatan yang banyak diminati terutama masyarakat Asia Tenggara. Salah satu acara olahraga yang paling bergengsi di kawasan Asia Tenggara yaitu *Southeast Asian Games* (SEA Games). SEA Games merupakan salah satu acara olahraga yang diselenggarakan di wilayah Asia Tenggara dan hanya dilaksanakan setiap dua tahun sekali yang melibatkan sebelas negara anggota *Association of South East Asian Nation* (ASEAN). Isu SEA Games paling banyak terjadi di *twitter* dengan sebesar 20.600 *tweets*. Hal ini disebabkan karena acara SEA Games di Filipina tahun 2019 yang banyak mengalami kejanggalan, salah satunya adalah stadion Rizal Memorium yang sampai saat ini belum direnovasi. Tujuan dari penelitian ini adalah mendapatkan dan membandingkan hasil ketepatan klasifikasi sentimen pengguna *twitter* terhadap SEA Games di Filipina tahun 2019 menggunakan *k-nearest neighbor* dan *support vector machine*. Data yang digunakan pada penelitian ini berasal dari data pengguna media sosial *Twitter* yang sering menggunakan

hashtag "SEA Games 2019" yang sudah dilakukan *text preprocessing* sebesar 2697 *tweets* dengan partisi data sebanyak 60% untuk data *training* dan 40% untuk data *testing*. Kesimpulan yang dapat diambil dari penelitian ini adalah hasil akurasi terbaik dalam klasifikasi *k-nearest neighbor* dan *support vector machine* adalah klasifikasi *support vector machine* dengan kernel *polynomial* sebesar 92.96% sehingga prediksi dari klasifikasi *Support Vector Machine* cenderung negatif.

Kata kunci: *K-nearest Neighbor*, SEA Games, Sentimen, *Support Vector Machine*.

1. Pendahuluan

Olahraga merupakan salah salatu unsur yang berpengaruh dalam kehidupan manusia yang ikut berperan dalam mengharumkan nama daerah dan bangsa baik melalui kompetisi regional, nasional maupun internasional. Setiap bangsa dari seluruh dunia berlomba-lomba menciptakan prestasi dalam kegiatan olahaga karena prestasi olahraga yang baik akan meningkatkan citra daerah, bangsa di dunia internasional [1]. Salah satu kegiatan olahraga yang menarik perhatian masyarakat kawasan Asia Tenggara yaitu *Southeast Asian Games* (SEA Games).

Southeast Asian Games (SEA Games) merupakan salah satu *event* olahraga yang diselenggarakan di wilayah Asia Tenggara saja. Acara ini dilaksanakan setiap dua tahun sekali dan melibatkan sebelas negara anggota *Association of South East Asian Nation* (ASEAN). SEA Games memiliki banyak tujuan yaitu untuk mengeratkan kerja sama antar negara angota ASEAN, menyatukan pemahaman dan mempererat hubungan antar negara angota ASEAN di kawasan semenanjung Asia Tenggara [2].

Acara SEA Games di Filipina tahun 2019 menjadi acara dengan persiapan yang kurang baik. Hal ini disebabkan ketika atlet sepak bola Timor Leste yang dikabarkan terlantar di bandara selama hampir tiga jam. Sementara itu, atlet Myanmar dan Kamboja juga merasakan ketidaknyamanan atas kelalaian Filipina dalam menyediakan transportasi dan akomodasi yang layak. Filipina juga mendapat sorotan dari berbagai media karena sejumlah *venue* masih belum siap dipakai, yaitu Stadion Rizal Memorium yang bakal digunakan untuk cabang olah raga sepak bola dan atletik belum tuntas direnovasi [3]. Buruknya acara SEA Games di Filipina tahun 2019 menjadi *hot topic trending* terutama di media sosial *twitter*.

Isu SEA Games paling banyak terjadi di *twitter* dengan sebesar 20.600 *tweets*. Netizen menggunakan *hashtag* #SEAGames2019fail di media sosial sebagai kritikan terhadap Filipina. *Hashtag* #SEAGames2019fail sudah dicuit sebanyak 20.600 *tweets*, baik dari netizen Indonesia maupun netizen dari kawasan Asia Tenggara, termasuk netizen Filipina sendiri. Mereka menyayangkan Filipina tidak mempersiapkan acara SEA Games dengan baik [4].

Twitter adalah sebuah situs *website* yang dimiliki dan dioperasikan oleh *Twitter* Inc. yang menawarkan jaringan sosial berupa mikroblog sehingga memungkinkan penggunanya untuk mengirim dan membaca pesan *tweet*. *Tweet* menampilkan beberapa spesifikasi seperti sebuah teks tulisan hingga 140 karakter yang ditampilkan pada halaman profil pengguna, serta dapat dilihat secara publik, namun pengirim dapat membatasi pengiriman pesan ke daftar teman-teman mereka saja [5]. Media sosial *twitter* digunakan untuk melakukan analisis sentimen terhadap topik yang sedang dibahas berdasarkan opini netizen dari *twitter*.

Analisis sentimen merupakan penilaian publik dari berbagai media sosial terkait suka atau tidak suka [6]. Salah satu tujuan dari analisis sentiment yaitu untuk mengetahui

seberapa banyaknya teks yang mengandung ekspresi positif, negatif, ataupun netral. Namun, apabila teks tersebut tidak bisa terdeteksi dalam kamus di salah satu *software*, maka teks tersebut dianggap netral sehingga kemungkinan besar peneliti mengalami kesulitan dalam mengambil keputusan yang tepat dalam melakukan analisis sentimen. Sehingga dengan melakukan analisis sentimen, peneliti dapat mengklasifikasikan dengan menggunakan metode *k-nearest neighbor* dan *support vector machine* pada penelitian ini.

Algoritma *K-nearest Neighbor* merupakan suatu metode *lazy learning* dimana tidak ada model yang dipelajari dari *data testing* sehingga hanya belajar dari contoh uji yang harus diklasifikasikan. Tujuan dari algoritma *K-nearest Neighbor* adalah untuk mengklasifikasikan objek berdasarkan atribut dari *data training* [7].

Support Vector Machine merupakan salah satu metode terbaik yang bisa dipakai dalam permasalahan klasifikasi. Konsep *Support Vector Machine* bermula dari masalah klasifikasi dua kelas sehingga membutuhkan *training set* positif dan negatif. *Support Vector Machine* berusaha menemukan *hyperplane* (pemisah) terbaik untuk memisahkan ke dalam dua kelas dan memaksimalkan margin antara dua kelas tersebut. Pada beberapa kasus, data tidak bisa diklasifikasi menggunakan metode *linear* SVM, sehingga dikembangkan fungsi *kernel* untuk mengklasifikasikan data dalam bentuk non *linear* [8].

Penelitian sebelumnya yang berkaitan dengan analisis sentimen dan metode *k-nearest neighbor* pernah dilakukan oleh Retno Sari pada tahun 2020 dengan judul penelitian "Analisis Sentimen pada Review Objek Wisata Dunia Fantasi Menggunakan Algoritma *K-nearest Neighbor*". Variabel yang digunakan oleh peneliti berasal dari media sosial *twitter* tentang opini masyarakat mengenai objek wisata dunia fantasi sebanyak 100 data yang dibagi menjadi dua jenis yaitu 50 data *review* positif dan 50 data *review* negatif [9].

Penelitian sebelumnya yang berkaitan dengan analisis sentimen dan metode *support vector machine* pernah dilakukan oleh Umi Rofiqoh, Rizal Setya Perdana dan M. Ali Fauzi pada tahun 2017 dengan judul penelitian "Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia pada *Twitter* dengan Metode *Support Vector Machine* dan *Lexicon Based Features*". Variabel yang digunakan oleh peneliti berasal dari media sosial *twitter* tentang opini masyarakat mengenai penyedia layanan telekomunikasi seluler sebanyak 300 data yang dibagi menjadi dua jenis yaitu data dengan perbandingan 70% untuk data latih dan 30% untuk data uji [10].

Berdasarkan latar belakang yang sudah dipaparkan, maka peneliti ingin meneliti tentang komentar netizen *twitter* terhadap *Southeast Asian Games* (SEA Games) di Filipina tahun 2019 menggunakan *k-nearest neighbor* dan *support vector machine*. Tujuan dari penelitian ini adalah mendapatkan dan membandingkan hasil ketepatan klasifikasi sentimen pengguna *twitter* terhadap *Southeast Asian Games* (SEA Games) di Filipina tahun 2019 menggunakan *k-nearest neighbor* dan *support vector machine*.

2. Tinjauan Pustaka

2.1 *Text Mining*

Text mining merupakan bagian dari *data mining*, namun proses *text mining* memerlukan lebih banyak tahapan dibandingkan dengan *data mining* karena data teks memiliki karakteristik yang lebih kompleks dari pada data biasa bahkan data yang sudah terstruktur. Berdasarkan ketidakteraturan struktur data teks, maka proses *text mining*

memerlukan beberapa tahap awal yang pada intinya adalah mempersiapkan agar teks dapat diubah menjadi lebih terstruktur [11].

2.2 Analisis Sentimen

Analisis sentimen merupakan salah satu cabang ilmu dari *text mining*, *natural language program*, dan *artificial intelligence*. Proses yang dilakukan oleh analisis sentimen untuk memahami, mengekstrak, dan mengolah data teks secara otomatis sehingga menjadi suatu informasi yang bermanfaat [12].

2.3 *Wordcloud*

Wordcloud merupakan suatu dari output dalam *sentiment analysis* yang menggambarkan karakteristik dari teks. Karakteristik dari *wordcloud* yaitu berupa kumpulan kata-kata yang ukuran besar hurufnya. Semakin besar tampilan suatu kata dalam *wordcloud* maka semakin besar juga frekuensi kemunculan dari kata tersebut [13].

2.4 Pembobot Kata

Pembobot kata dapat digunakan untuk memberikan bobot pada fitur kata pada teks berdasarkan modus kata. Fitur kata pada teks tersebut dapat digunakan untuk melakukan klasifikasi *k-nearest neighbor* dan *support vector machine*. Berikut adalah langkah-langkah pembobot kata [10].

a. *Term Frequency* (TF)

Term Frequency merupakan jumlah kemunculan atau frekuensi kata pada suatu *tweet*. *Term Frequency* (tf_{pn}) didefinisikan sebagai jumlah frekuensi kata ke-p pada *tweet* ke-n. Berikut ini adalah persamaan dari *Term Frequency* yang dapat dilihat pada Persamaan (2.4.1).

$$W_{tf_{(pn)}} = \frac{n_{pn}}{n_n} \tag{2.4.1}$$

Keterangan :
$W_{tf(pn)}$: pembobot *Term Frequency* pada kata ke-p di dalam *tweet* ke-n.
n_{pn} : jumlah kata ke-p di dalam *tweet* ke-n.
n_n : jumlah semua kata di dalam *tweet* ke-n.

b. *Inverse Document Frequency* (IDF)

Inverse Document Frequency merupakan bobot kebalikan dari bobot *document frequency*. Kata yang jarang muncul pada *tweet* memiliki bobot *inverse document frequency* yang tinggi. *Inverse Document Frequency* (idf_p) didefinisikan sebagai invers pembobot frekuensi *tweet* pada kata ke-n. Berikut ini adalah persamaan dari *Inverse Document Frequency* yang dapat dilihat pada Persamaan (2.4.2).

$$idf_p = \log_{10}\left(\frac{N}{df_{(p)}}\right) \tag{2.4.2}$$

Keterangan :
idf_p : invers pembobot frekuensi *tweet* pada kata ke-p.
N : jumlah *tweet*.
$df_{(p)}$: jumlah *tweet* yang mengandung kata ke-p.

c. *Term Frequency - Inverse Document Frequency* (TF - IDF)

Pembobot *Term Frequency-Inverse Document Frequency* merupakan hasil perkalian dari *term frequency* dan *inverse document frequency*. Pembobot *Term Frequency-Inverse Document Frequency* (W_{pn}) didefinisikan sebagai pembobot kata ke-p pada *tweet* ke-n. Berikut ini adalah persamaan dari pembobot *Term Frequency-Inverse Document Frequency* yang dapat dilihat pada Persamaan (2.4.3).

$$W_{pn} = W_{tf_{(pn)}} * idf_p \tag{2.4.3}$$

W_{pn} : pembobot kata ke-p di dalam *tweet* ke-n.
$W_{tf(pn)}$: pembobot *Term Frequency* pada kata ke-p di dalam *tweet* ke-n.
idf_p : invers pembobot frekuensi *tweet* pada kata ke-p.

2.5 *K-nearest Neighbor*

Algoritma *K-nearest neighbor* merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data *training* yang menggunakan jarak terdekat atau kemiripan pada objek tertentu. *K-nearest neighbor* bekerja pada jarak terpendek dari *query distance* ke data *training sample* untuk menentukan *K-nearest neighbor* [14]. Untuk menghitung jarak kedekatan antar tetangga dengan berjumlah sebesar k (kelompok) berdasarkan *Euclidean Distance* yang dapat dilihat pada Persamaan (2.5.1) sebagai berikut.

$$d_{(x_1,x_{2i})} = \sqrt{\sum_{i=1}^{n}(x_1 - x_{2i})^2} \tag{2.5.1}$$

Nilai x_1 merupakan plot data *training* yang belum terprediksi mendapat klasifikasi positif atau negatif, x_{2i} merupakan plot data *training* yang sudah terprediksi mendapat klasifikasi positif atau negatif, i merupakan variabel dari suatu data *training* (i = 1,2,..., n), serta $d_{(x1, x2i)}$ merupakan jarak Euclidian.

2.6 *Support Vector Machine*

Support Vector Machine merupakan suatu teknik yang relatif baru untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi. *Support Vector Machine* adalah seperangkat metode pembelajaran terbimbing yang menganalisis data dan mengenali pola yang digunakan untuk klasifikasi dan analisis regresi. Algoritma SVM pertama kali diciptakan oleh Vladimir Vapnik dan turunan saat ini (margin lunak) diusulkan oleh Corinna Cortes dan Vladimir Vapnik [15].

Tahapan untuk metode *support vector machine* dimulai dari mengubah data teks ke dalam bentuk *vector* data dan dikombinasikan dengan nilai TF IDF untuk pembobotan [16].

Fungsi deskriptif dari metode *support vector machine* dapat diketahui berdasarkan kedua kelas tersebut diasumsikan masuk kelas (-1) dan (+1) yang dapat terpisah secara sempurna oleh *hyperplane* yang berdimensi *d* yang dapat ditulis seperti Persamaan (2.6.1) sebagai berikut.

$$h(x) = \vec{w}' \vec{x} + b = 0 \tag{2.6.1}$$

Dimana $\bar{w}$ adalah vektor bobot untuk dimensi ke *d* dan *b* adalah suatu konstanta yang biasa disebut bias (skalar).

Pattern $\overline{x_i}$ yang termasuk kelas sampel positif (+1) dapat dirumuskan sebagai *pattern* yang memenuhi Pertidaksamaan (2.6.2) sebagai berikut.

$$h(x) = \vec{w}' \vec{x} + b \leq -1 \tag{2.6.2}$$

Sedangkan *pattern* $\overline{x_i}$ yang termasuk kelas sampel positif (+1) dapat dirumuskan sebagai *pattern* yang memenuhi Pertidaksamaan (2.6.3) sebagai berikut.

$$h(x) = \vec{w}' \vec{x} + b \geq +1 \tag{2.6.3}$$

2.7 *Confusion Matrix*

Confusion Matrix merupakan Teknik yang digunakan untuk mengevaluasi klasifikasi model untuk memperkirakan objek yang benar atau salah. Sebuah matriks dari prediksi akan dibandingkan dengan kelas asli yang berisi informasi actual dan prediksi nilai klasifikasi. Setelah sistem berhasil mengklasifikasikan *tweet*, dibutuhkan ukuran untuk menentukan seberapa valid atau tepat dari klasifikasi yang telah dibuat oleh sitem. Tabel 2.1 menunjukkan *confusion matrix* yang digunakan untuk membantu dalam perhitungan sistem evaluasi [17]. Berikut adalah struktur *confusion matrix* yang dapat dilihat pada Tabel 2.1 serta rumus perhitungan nilai *accuracy, sensitivity* dan *specificity* berdasarkan Persamaan (2.7.1), Persamaan (2.7.2), dan Persamaan (2.7.3).

Tabel 2.1 *Confusion Matrix*

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{2.7.1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.7.2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.7.3}$$

2.8 SEA Games

Southeast Asian Games (SEA Games) merupakan salah satu *event* olahraga yang diselenggarakan di wilayah Asia Tenggara dan dilaksanakan setiap dua tahun sekali yang melibatkan sebelas negara anggota *Association of South East Asian Nation* (ASEAN). Negara-negara anggota *Association of South East Asian Nation* (ASEAN) diantaranya adalah Indonesia, Malaysia, Brunei Darussalam, Singapura, Timor Leste, Filipina, Thailand, Laos, Myanmar, Kamboja, dan Vietnam. *Southeast Asian Games* (SEA Games) memiliki banyak tujuan, yaitu untuk mengeratkan kerja sama antar negara angota *Association of South East Asian Nation* (ASEAN), menyatukan pemahaman dan mempererat hubungan antar negara angota *Association of South East Asian Nation* (ASEAN) di kawasan semenanjung Asia Tenggara [2].

3. Metodologi Penelitian

3.1 Sumber Data dan Variabel Penelitian

Data yang digunakan pada penelitian ini berasal dari data pengguna media sosial *Twitter* yang sering menggunakan *hashtag* "SEA Games 2019" pada tanggal 20 November 2019 hingga 25 November 2019. Data didapat dari *Twitter* API (*Application Programming Interface)* sebanyak 5000 *tweets*. Setelah dilakukan proses *text preprocessing* dan *feature selection*, maka data yang didapat telah tereduksi menjadi 2697 *tweets* dengan klasifikasi sentimen positif dan negatif.

Data yang berjumlah 2697 *tweets* dibagi menjadi data *training* dan data *testing* dengan persentase menjadi 60% : 40%. Data *training* digunakan untuk membuat prediksi dari *k-nearest neighbor* dan *support vector machine*, sedangkan data *testing* digunakan unuk melihat nilai akurasi dari *k-nearest neighbor* dan *support vector machine*. Persentase data *training* dan data *testing* yaitu 60% : 40% mengikuti peneliti sebelumnya yang menggunakan persentase data *training* dan data *testing* yang sama [18]. Sehingga data *training* yang didapat pada penelitian ini adalah 1618 *tweets*, sedangkan data *testing* nya adalah 1079 *tweets*. Berikut adalah variabel penelitian yang digunakan dalam penelitian ini yang dapat dilihat pada Tabel 3.1.

Tabel 3.1 Variabel Penelitian

Variabel	Indikator	Skala	Definisi Operasional
Y	Klasifikasi Sentimen	Nominal	Sentimen merupakan keputusan pendapat dari seseorang terhadap suatu topik atau target [19].
X_1	Nilai pembobot *Term Frequency-Inverse Document Frequency* pada Kata ke-1	Rasio	Pembobot *Term Frequency-Inverse Document Frequency* (W_{1n}) merupakan hasil perkalian dari *term frequency* dan *inverse document frequency* sehingga didefinisikan sebagai pembobot kata ke-1 pada *tweet* ke-n [10].
:	:	:	
X_p	Nilai pembobot *Term Frequency-Inverse Document Frequency* pada Kata ke-p	Rasio	Pembobot *Term Frequency-Inverse Document Frequency* (W_{pn}) merupakan hasil perkalian dari *term frequency* dan *inverse document frequency* sehingga didefinisikan sebagai pembobot kata ke-p pada *tweet* ke-n [10].

3.2 Langkah Penelitian

Langkah-langkah dalam penelitian ini adalah sebagai berikut.

1. Mengambil data *tweet* dengan *Twitter* API.
2. *Text Preprocessing*.

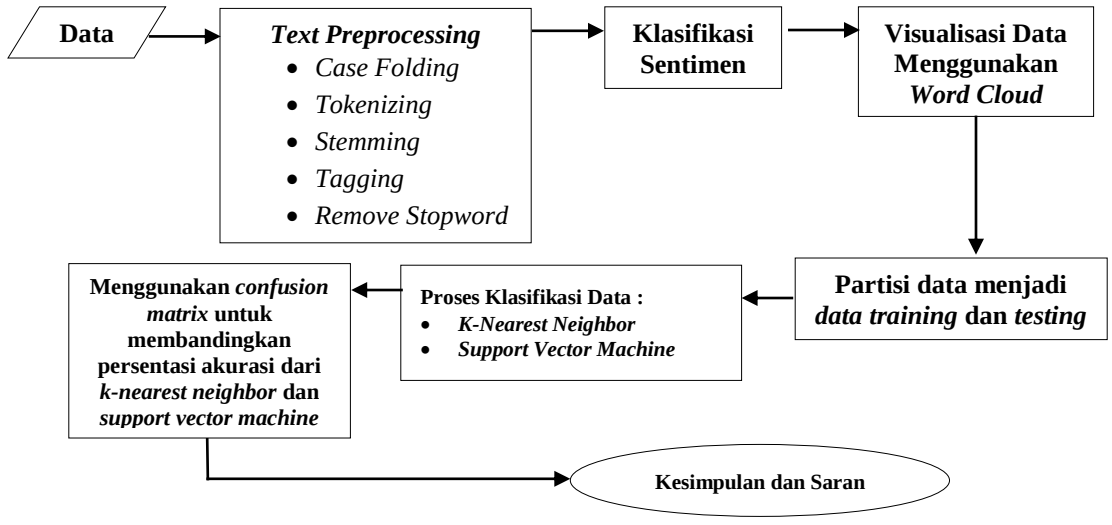
Text preprocessing memiliki lima proses, yaitu :

- *Case Folding* : Menghapus karakter huruf yang tidak valid seperti angka, tanda baca dan *Uniform Resources Locator* (URL).
- *Tokenizing* : Mengubah semua karakter huruf menjadi huruf kecil
- *Stemming* : Mengubah kata yang berimbuhan menjadi kata dasar.
- *Tagging* : Mengubah kata dalam bentuk lampau menjadi kata awalnya.
- *Remove Stopword* : Menghapus kata penghubung.

3. Pengklasifikasikan analisis sentimen (komentar positif dan negatif) yang diambil dari sebuah *tweet* yang sudah dilakukan *text preprocessing*.
4. Melakukan visualisasi *tweet* dengan *word cloud*.
5. Melakukan partisi data menjadi 60% : 40% dengan syarat data sudah terklasifikasi sentimen positif dan negatif serta mendapat nilai pembobot TF-IDF pada setiap kata.
6. Klasifikasi data menggunakan *k-nearest neighbor* dan *support vector machine* untuk masing-masing komentar netizen terhadap *Southeast Asian Games* (SEA Games) di Filipina tahun 2019.
7. Menggunakan *confusion matrix* untuk membandingkan hasil akurasi dari *k-nearest neighbor* dan *support vector machine* untuk masing-masing komentar netizen terhadap *Southeast Asian Games* (SEA Games) di Filipina tahun 2019.
8. Menarik Kesimpulan dan Saran.

3.3 Diagram Alir

Berdasarkan langkah penelitian maka dilakukan melalui beberapa tahapan seperti pada Gambar 3.1.



Gambar 3.1 Diagram Alir

4. Hasil dan Pembahasan

4.1 *Text Preprocessing* dan Karakteristik Data SEA Games 2019 di Filipina

Data *tweet* mengenai SEA Games di Filipina tahun 2019 yang telah terkumpul dilakukan *text preprocessing*, seperti *case folding*, *tagging*, *stemming*, dan *tokenizing*.

Berikut adalah struktur data *tweet* pada SEA Games di Filipina tahun 2019 sebelum dilakukan *text preprocessing* :

Tabel 4.1 Struktur Data SEA Games di Filipina Tahun 2019 Sebelum *Text Preprocessing*

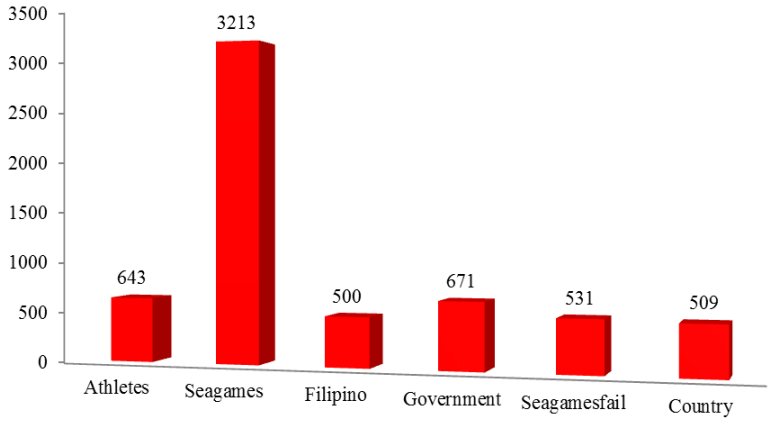
No.	*Tweet*
1.	RT @rapplerdotcom: Senator Risa Hontiveros calls it 'laughable' that Vice President Leni Robredo got fired from her anti-drug post first, i…
2.	RT @iskolarspeaks: Before #SEAGames2019 officially starts, I just want to wish all competing athletes, especially our own, the best of luck…
:	:
5000.	RT @jomszj: Cayetano: I'm fucked. Duterte: *hold my beer* #SEAGames2019 #SEAGamesfail

Data *tweet* yang belum dilakukan *text preprocessing* masih tersusun dalam satu kolom seperti pada Tabel 4.1. Data tersebut masih memuat *username, URL,* kata-kata yang dianggap bukan merupakan kata penting dalam *tweet* (*stopword*) dan simbol-simbol lainnya yang tidak menggambarkan isi tweet seperti simbol *retweet* (RT) dan tanda baca, sehingga perlu dilakukan *text preprocessing* untuk mendapatkan data *tweet* yang tidak memuat hal-hal tersebut, serta meningkatkan ketepatan klasifikasi dari segi akurasi dan mengurangi tingkat kesalahan dalam mengklasifikasikan data. Berikut adalah struktur data SEA Games di Filipina tahun 2019 yang sudah dilakukan *text preprocessing* :

Tabel 4.2 Struktur Data SEA Games di Filipina Tahun 2019 Setelah *Text Preprocessing*

No.	*Tweet*
1.	senator risa hontiveros calls laughable vice president leni robredo fired antidrug post …
2.	seagames officially starts wish competing athletes especially own luck…
:	:
5000.	cayetano im fucked duterte hold beer seagames seagamesfail

Tabel 4.2 dapat dilakukan perhitungan frekuensi kemunculan kata tertinggi dengan minimal frekuensi kemunculan kata di dalam semua *tweet* dalam penelitian ini adalah 100 kali, seperti pada Gambar 4.1 sebagai berikut..



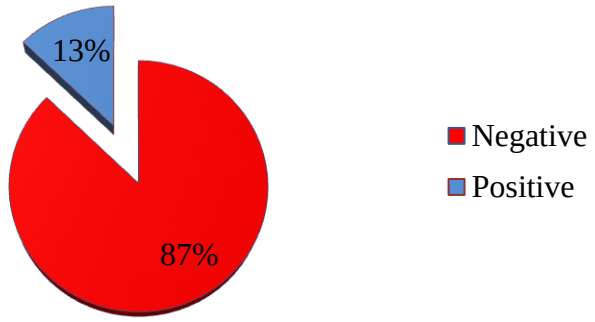
Gambar 4.1 Frekuensi Kemunculan Kata pada Data SEA Games di Filipina Tahun 2019 Setelah *Text Preprocessing*

Daftar kata dalam frekuensi kemunculan tertinggi pada SEA Games 2019 dalam Gambar 4.1 dijelaskan bahwa kata-kata tersebut merupakan kata-kata yang memiliki pengaruh signifikan dalam pembangunan model klasifikasi, seperti yang ditunjukkan pada Tabel 4.3 berisi data kategori sentimen beserta data nilai TF-IDF untuk setiap satu kata dalam satu *tweet*.

Tabel 4.3 Struktur Data SEA Games di Filipina Tahun 2019 Setelah *Text Preprocessing* dengan Klasifikasi Sentimen dan Nilai TF-IDF

No.	athletes	seagames	Seagamesfail	Filipino	country	goverment	Sentiment
1.	0	0	0	0	0	0	Negative
2.	0.1005	0	0	0	0	0.1072	Negative
:	:	:	:	:	:	:	:
2697.	0	0.0198	0	0	0.0871	0.0858	Negative

Pada Tabel 4.3 terdapat kolom sentiment merupakan variabel respon, sedangkan nilai TF-IDF untuk setiap satu kata dalam satu *tweet* sebagai variabel prediktor. Apabila kategori sentimen disajikan dalam bentuk *pie chart* dapat dilihat perbandingan kategori sentiment positif dan negatif.



Gambar 4.2 *Pie Chart* Kategori Sentimen pada Data SEA Games di Filipina Tahun 2019

Frekuensi kategori sentimen pada Gambar 4.2 menujukkan bahwa sebanyak 87% netizen *twitter* lebih banyak berkomentar buruk tentang acara SEA Games di Filipina tahun 2019. Seperti dalam Gambar 1, kata "seagames" memiliki frekuensi kemunculan kata tertinggi sebesar 3213 kali. Hal ini disebabkan oleh isu-isu yang beredar di *twitter* tentang SEA Games di Filipina tahun 2019 yang memburuk.

4.2 Visualisasi Data SEA Games 2019 di Filipina Menggunakan *Wordcloud*

Visualisasi data teks menggunakan *wordcloud* digunakan untuk mengetahui kata-kata yang paling sering muncul pada suatu data. Pada penelitian ini, *wordcloud* digunakan untuk visualisasi *tweet* berdasarkan kategori sentimennya sehingga dapat diketahui kata-kata yang sering muncul pada setiap sentimen. Ukuran *font* pada *wordcloud* menunjukkan frekuensi kemunculan kata. Semakin besar ukuran *font* berarti semakin besar frekuensi kemunculan kata tersebut. Berikut merupakan *wordcloud* data pada SEA Games di Filipina tahun 2019 dengan kata yang sering muncul adalah "seagames" yang ternyata sesuai dengan Gambar 4.1.

Gambar 4.3 *Wordcloud* pada Data SEA Games di Filipina Tahun 2019

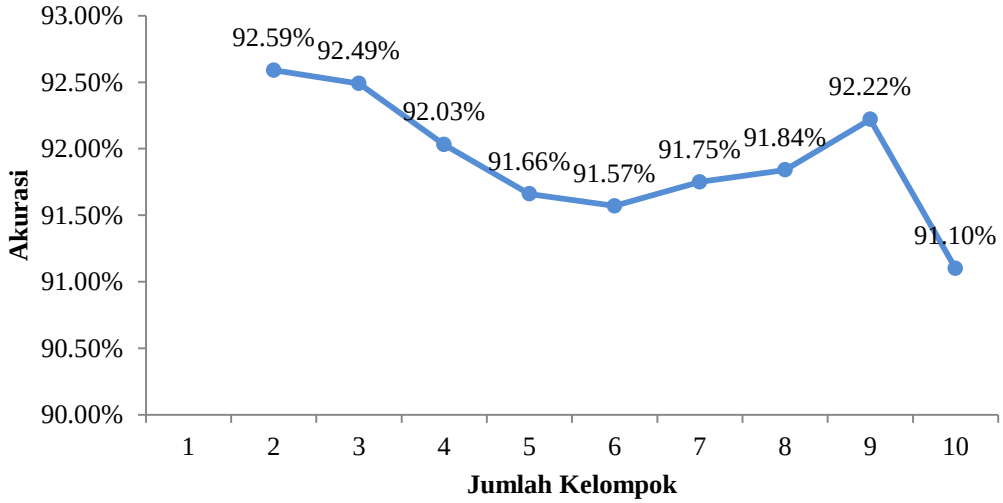
Wordcloud pada data SEA Games di Filipina tahun 2019 dalam Gambar 4.3 menunjukkan kata-kata yang paling sering muncul pada pembahasan mengenai SEA Games di Filipina tahun 2019. Kata yang paling sering muncul adalah kata "seagames", "goverment", "filipino", "athletes", "country" dan "seagamesfail" tanpa memandang klasifikasi sentimen itu positif maupun negatif.

4.3 Klasifikasi Data SEA Games di Filipina Tahun 2019 Menggunakan *K-nearest Neighbor* dan *Support Vector Machine*

Klasifikasi *K-nearest Neighbor* dan *Support Vector Machine* dapat mengeluarkan nilai akurasi dari *confusion matrix*, serta memberikan prediksi dari klasifikasi sentimen tersebut.

a. Klasifikasi *K-nearest Neighbor*

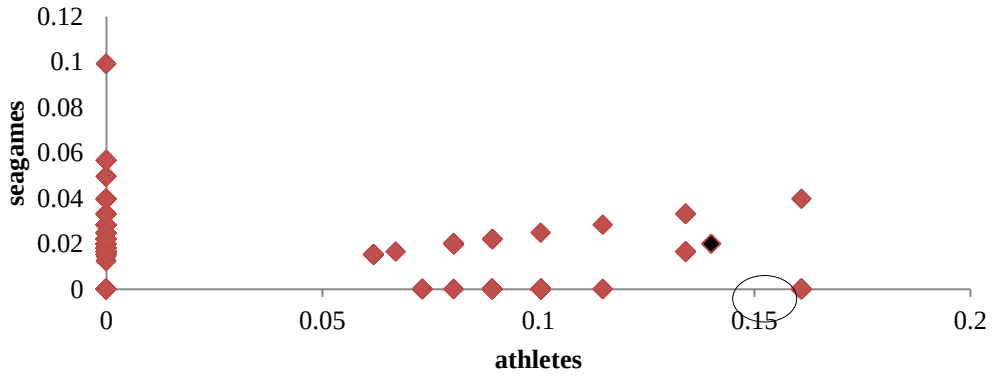
K-nearest Neighbor merupakan suatu klasifikasi yang berguna untuk mengelompokkan kedekatan antar plot. Berikut ini adalah hasil akurasi dengan menggunakan *K-nearest Neighbor* yang ditujukan pada Gambar 4.4.



Gambar 4.4 Klasifikasi Data SEA Games di Filipina Tahun 2019 Menggunakan *K-nearest Neighbor*

Hasil akurasi terbaik pada Gambar 4.4 berada pada jumlah kelompok (k) sebanyak 2 dengan sebesar 92.59%, kemudian dapat dilanjutkan dengan melihat hasil plot dari klasifikasi *k-nearest neighbor* dengan nilai k yang optimum dari segi akurasi.

Hasil akurasi dengan menggunakan *k-nearest neighbor* pada Gambar 4.5 memunculkan hasil kedekatan antar plot sebagai berikut.



Gambar 4.5 Plot Data SEA Games di Filipina Tahun 2019 Menggunakan *K-nearest Neighbor*

Gambar 4.5 dapat digunakan untuk melihat kedekatan antar plot dengan menggunakan *k-nearest neighbor* dan diperoleh salah satu plot yang belum terprediksi mendapat sentimen positif maupun negatif. Plot berwarna hitam yaitu nilai pembobot *Term Frequency-Inverse Document Frequency* pada *athletes* dan *seagames* yaitu 0.14 dan 0.02. Perhitungan manual dapat dilihat sebagai berikut.

$$d_{(x_1,x_{2,[1]})} = \sqrt{(0.14-0)^2+(0.02-0)^2} = 0.1414$$

$$d_{(x_1,x_{2[2]})} = \sqrt{(0.14-0.1006)^2+(0.02-0)^2} = 0.0442$$

$$\vdots$$

$$d_{(x_1,x_{2,[1618]})} = \sqrt{(0.14-0)^2+(0.02-0)^2} = 0.1414$$

Berdasarkan perhitungan manual diperoleh jarak Euclidian yang di*rank* dengan urutan dari nilai terkecil hingga terbesar dapat disajikan seperti di Tabel 4.6.

Tabel 4.6 Perhitungan Manual pada Klasifikasi *K-nearest Neighbor* (Sebelum dan Sesudah Diurutkan Berdasarkan Jarak Euclidian)

Perhitungan Manual *K-nearest Neighbor* (Sebelum diurut berdasarkan Jarak Euclidian)			
Tweet	Jarak Euclidian	*Rank*	Prediksi
1.	0.1414	918	Negative
2.	0.0442	60	Negative
:	:	:	:
1618.	0.1414	918	Negative
Perhitungan Manual *K-nearest Neighbor* (Sesudah diurut berdasarkan Jarak Euclidian)			
Tweet	Jarak Euclidian	*Rank*	Prediksi
277.	0.0068	1	Positive
475.	0.0068	1	Positive

:	:	:	:
1429.	0.1609	1618	Negative

Dari Tabel 4.6 setelah di*rank*, *tweet* ke 277 dan 475 memiliki jarak Euclidian terpendek dikarenakan (*tweet* ke 277 dan 475 berdekatan dengan plot hitam). *Tweet* ke 277 dan 475 diambil dari *data training* yang mendapat klasifikasi sentimen positif, sehingga plot hitam yang belum terprediksi mendapat klasifikasi sentimen berdekatan dengan plot merah (*tweet* ke 277 dan 475) dengan mayoritas mendapat sentimen positif, sehingga plot hitam dipastikan mendapat prediksi klasifikasi sentimen positif.

Setelah melihat hasil plot data SEA Games di Filipina tahun 2019, maka dilakukan perhitungan ketepatan klasifikasi pada data SEA Games di Filipina tahun 2019 dengan klasifikasi *k-nearest neighbor*. Perhitungan ketepatan klasifikasi sangat berguna bagi peneliti untuk mengukur apakah klasifikasi tersebut merupakan klasifikasi dengan nilai akurasi yang tertinggi dibandingkan dengan klasifikasi yang lainnya, sehingga hasil klasifikasi *k-nearest neighbor* (nilai k optimum berada di k = 2) pada data SEA Games di Filipina tahun 2019 dapat dilihat pada Tabel 4.7 sebagai berikut.

Tabel 4.7 Ketepatan Klasifikasi pada Data SEA Games di Filipina Tahun 2019 Menggunakan *K-nearest Neighbor*

Prediction	*Reference*	
	Negative	*Positive*
Negative	924	63
Positive	17	75

Pengukuran ketepatan klasifikasi pada data SEA Games di Filipina tahun 2019 diketahui berdasarkan *confusion matrix* adalah sebagai berikut.

$$Accuracy = \frac{924+75}{924+63+17+75} = 0.9259 = 92.59\%$$

$$Sensitivity = \frac{75}{75+63} = 0.5434 = 54.34\%$$

$$Specificity = \frac{924}{924+17} = 0.9819 = 98.19\%$$

Perhitungan diatas dapat diambil kesimpulan bahwa nilai akurasi dari klasifikasi *k-nearest neighbor* adalah sebesar 92.59% yang artinya klasifikasi *k-nearest neighbor* memiliki ketepatan akurasi dalam memprediksi data SEA Games di Filipina tahun 2019 sebesar 92.59%. Nilai sensitivitas (*sensitivity*) dari klasifikasi *k-nearest neighbor* adalah sebesar 54.34% yang artinya proporsi komentar netizen tentang SEA Games di Filipina tahun 2019 yang mendapat sentimen positif adalah sebesar 54.34%. Sedangkan nilai spesifisitas (*specificity*) dari klasifikasi *k-nearest neighbor* adalah sebesar 98.19% yang artinya proporsi komentar netizen tentang SEA Games di Filipina tahun 2019 yang mendapat sentimen negatif adalah sebesar 98.19% .

b. Klasifikasi *Support Vector Machine*

Support vector machine merupakan suatu klasifikasi yang memperhatikan nilai *cost* (C) sebagai penentuan parameter untuk hasil akurasi terbaiknya, baik itu menggunakan kernel *linear*, *polynomial*, *radial basis* maupun *sigmoid*. Perbandingan antar keempat kernel pada klasifikasi *support vector machine* bertujuan untuk menentukan apakah salah

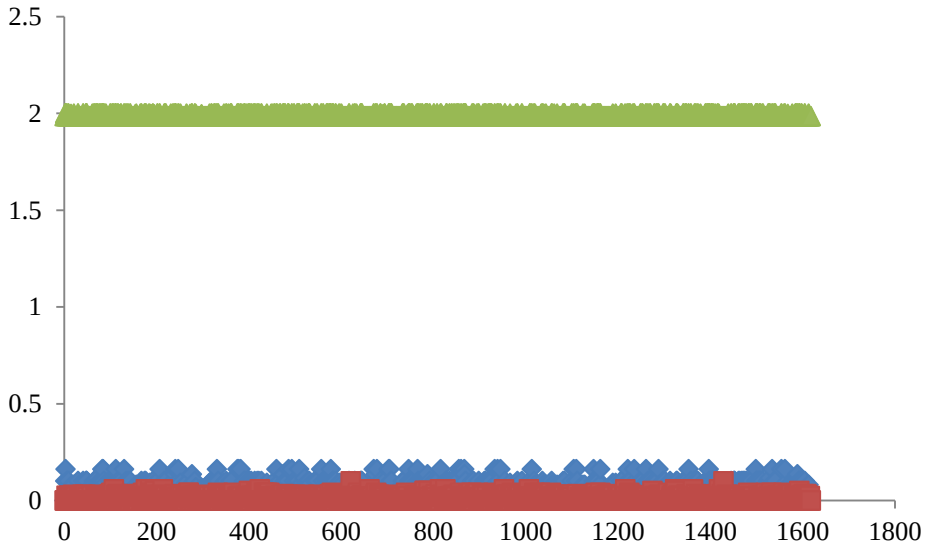
satu dari keempat kernel pada klasifikasi *support vector machine* memperoleh nilai akurasi yang terbaik. Untuk mengetahui hasil perbandingan antar keempat kernel pada klasifikasi *support vector machine* dapat dilihat pada Tabel 4.8.

Tabel 4.8 Perbandingan Keempat Kernel pada Klasifikasi Data SEA Games di Filipina Tahun 2019 Menggunakan *Support Vector Machine*

Klasifikasi	Kernel	*Cost* (C) (0.01, 0.1, 1, 10, 100)	*Degree* (p) (4, 5, 6)	Gamma (γ) (0.01, 0.1, 1, 10, 100)	Coef0 (r) (0, 1, 2, 3)	Akurasi
Support Vector Machine	*Linear*	v	-	-	-	87.21%
	Polynomial	**v**	**v**	**v**	**v**	**92.96%**
	Radial Basis Function	v	-	v	-	92.31%
	Sigmoid	v	-	v	v	89.06%

Perbandingan keempat kernel pada klasifikasi *support vector machine* pada Tabel 4.8 disimpulkan bahwa klasifikasi *support vector machine* dengan kernel *polynomial* merupakan kernel dengan mendapat nilai akurasi terbaik dengan sebesar 92.96%, sehingga klasifikasi *support vector machine* dengan kernel *polynomial* telah merepresentasikan dari semua kernel yang digunakan dari klasifikasi *support vector machine*.

Klasifikasi *support vector machine* dengan kernel *polynomial* merupakan suatu kernel yang mendapat nilai akurasi terbaik, sehingga dapat melihat plot *support vector machine* dengan kernel *polynomial* yang memperhatikan nilai parameter *cost* (C), *degree* (p), gamma (γ) dan coef0 (r) sebagaimana digambarkan pada Gambar 4.8 sebagai berikut.



Gambar 4.8 Plot Data SEA Games di Filipina Tahun 2019 Menggunakan *Support Vector Machine*

Plot data SEA Games di Filipina tahun 2019 dengan menggunakan *support vector machine* menggunakan kernel *polynomial* sudah ditentukan. Dengan parameter *cost* (C) sebesar 100, *degree* (p) sebesar 6, gamma (γ) sebesar 100 dan coef0 (r) sebesar 2, maka pada Gambar 3.8 menandakan bahwa plot data tersebut tidak terpisah secara linier

dikarenakan fungsi *hyperplane* yang berwarna hijau berada di luar plot variabel *athletes* dan *seagames* yaitu sebesar 2.

Setelah melihat hasil plot data SEA Games di Filipina tahun 2019, maka dilakukan perhitungan ketepatan klasifikasi pada data SEA Games di Filipina tahun 2019 dengan klasifikasi *support vector machine*. Perhitungan ketepatan klasifikasi digunakan untuk mengukur apakah klasifikasi tersebut merupakan klasifikasi dengan nilai akurasi yang tertinggi dibandingkan dengan klasifikasi yang lainnya, sehingga hasil klasifikasi *support vector machine* (kernel optimum berada di kernel *polynomial*) pada data SEA Games di Filipina tahun 2019 dapat dilihat pada Tabel 4.9 sebagai berikut.

Tabel 4.9 Ketepatan Klasifikasi pada Data SEA Games di Filipina Tahun 2019 Menggunakan *Support Vector Machine*

Prediction	*Reference*	
	Negative	*Positive*
Negative	934	69
Positive	7	69

Pengukuran ketepatan klasifikasi pada data SEA Games di Filipina tahun 2019 diketahui bahwa perhitungan dari ketepatan klasifikasi pada Tabel 4.9 dengan adalah sebagai berikut.

$$Accuracy = \frac{934 + 69}{934 + 69 + 7 + 69} = 0.9296 = 92.96\%$$

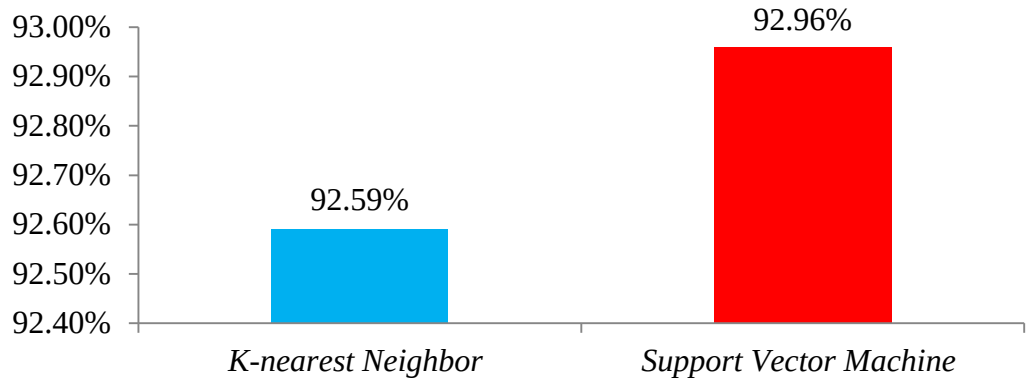
$$Sensitivity = \frac{69}{69 + 69} = 0.5000 = 50\%$$

$$Specificity = \frac{934}{934 + 7} = 0.9925 = 99.25\%$$

Perhitungan diatas dapat diambil kesimpulan bahwa nilai akurasi dari klasifikasi *support vector machine* adalah sebesar 92.96% yang artinya klasifikasi *support vector machine* memiliki ketepatan akurasi dalam memprediksi data SEA Games di Filipina tahun 2019 sebesar 92.96%. Nilai sensitivitas (*sensitivity*) dari klasifikasi *support vector machine* adalah sebesar 50% yang artinya proporsi komentar netizen tentang SEA Games di Filipina tahun 2019 yang mendapat sentimen positif sebesar 50%. Sedangkan nilai spesifisitas (*specificity*) dari klasifikasi *support vector machine* adalah sebesar 99.25% yang artinya proporsi komentar netizen tentang SEA Games di Filipina tahun 2019 yang mendapat sentimen negatif sebesar 99.25%.

c. Perbandingan Klasifikasi *K-nearest Neighbor* dan *Support Vector Machine*

Perbandingan antara kedua klasifikasi sangat berguna untuk menentukan apakah dengan menggunakan metode klasifikasi yang dipakai benar-benar mendapatkan nilai akurasi terbaiknya. Berikut ini adalah hasil perbandingan klasifikasi data SEA Games di Filipina tahun 2019 menggunakan *k-nearest neighbor* dan *support vector machine* yang dapat dilihat pada Gambar 4.9.



Gambar 4.9 Perbandingan Klasifikasi Data SEA Games di Filipina Tahun 2019 Menggunakan *K-nearest Neighbor* dan *Support Vector Machine*

Hasil akurasi terbaik pada Gambar 4.9 adalah dengan menggunakan klasifikasi *support vector machine* dengan nilai akurasi sebesar 92.96%. Sehingga dengan menggunakan klasifikasi *support vector machine* sebagai klasifikasi terbaik dari segi akurasi, maka dilakukan perbandingan antara *specificity* dan *sensitivity* sebagai penentuan prediksi dari klasifikasi *support vector machine* pada Tabel 4.9 yang menandakan bahwa nilai *specificity* lebih besar dari pada *sensitivity* sehingga prediksi dari klasifikasi *support vector machine* cenderung negatif yang artinya kemungkinan terbesar acara SEA Games di Filipina tahun 2019 mendapat sentimen negatif dari netizen di *twitter*.

4. Simpulan dan Saran

Kesimpulan yang dapat diambil pada penelitian ini adalah hasil akurasi terbaik dalam klasifikasi *k-nearest neighbor* dan *support vector machine* adalah klasifikasi *support vector machine* dengan kernel *polynomial* sebesar 92.96%, maka jika dilakukan perbandingan antara *specificity* dan *sensitivity* dari klasifikasi *support vector machine* bahwa nilai *specificity* lebih besar dari pada *sensitivity* sehingga prediksi dari klasifikasi *support vector machine* cenderung negatif yang artinya kemungkinan terbesar acara SEA Games di Filipina tahun 2019 mendapat sentimen negatif dari netizen di *twitter*, sehingga saran yang dikemukakan pada penelitian ini adalah pada peneliti perlu menambah data sebagai percobaan agar memperoleh hasil akurasi yang diharapkan serta menambah beberapa klasifikasi agar penelitian menjadi lebih menarik.

Daftar Pustaka

[1] Latimbang, F. I., 2017. Studi Kelayakan Sarana Prasarana Penunjang Pembinaan Olahraga Prestasi di Pusat Pendidikan dan Latihan Olahraga Pelajar (PPLP) Provinsi Gorontalo. Universitas Negeri Gorontalo, Gorontalo.

[2] Isna, Y., 2015. *Sea Games Ajang Pemersatu Asia Tenggara*. http://www.kompasiana.com. [1 Desember 2019]

[3] Tsalis, A., 2019. *Terungkap Sosok Penyebab Kebobrokan Penyelenggaraan SEA Games 2019 Filipina*. http://www.SportFEAT.com. [1 Desember 2019]

[4] Arifin, Y., 2019. *#SEAGames2019fail, Netizen Kritik Penyelenggaraan yang Kacau*. http://sport.detik.com. [1 Desember 2019]

[5] Twitter., 2020. *Pusat Bantuan*. http://help.twitter.com. [23 Maret 2020]

[6] Romadloni, N. T., Santoso, I., & Budilaksono, S., 2019. Perbandingan Metode Naïve Bayes, KNN dan Decision Tree Terhadap Analisis Sentimen Transportasi KRL. *Jurnal IKRA-ITH Informatika*, Vol 3, No. 2, 1-9.

[7] Salam, A., Zeniarja, J., Septiyan, R., Khasanah, U., 2018. Analisis Sentimen Data Komentar Sosial Media Facebook dengan K-nearest Neighbor (Studi Kasus pada Akun Jasa). *SINTAK 2018*, 480-486.

[8] Pratama, A., Wihandika, R. C., & Ratnawati, D. E., 2018. Implementasi Algoritma Support Vector Machine (SVM) untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, Vol 2, No. 4, 1704-1708.

[9] Sari, R., 2020. Analisis Sentimen pada Review Objek Wisata Dunia Fantasi Menggunakan Algoritma K-nearest Neighbor (K-NN). *Evolusi : Jurnal Sains dan Manajemen*, Vol 8, No. 1, 10-17.

[10] Rofiqoh, U., Perdana, R. S., & Fauzi, M. A., 2017. Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia pada Twitter dengan Metode Support Vector Machine dan Lexicon Based Features. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 1725-1732.

[11] Indriati., Ridok, A., 2016. Sentiment Analysis for Review Mobile Applications using Neighbor Method Weighted K-nearest Neighbor. *Journal of Enviromental Engineering & Sustainable Technology*, Vol 3, No. 1, 23-32.

[12] Luqyana, W. A., 2018. Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya, Malang.

[13] Latifah, E. F., 2018. Perbandingan Kinerja Machine Learning Berbasis Algoritma Support Vector Machine dan Naïve Bayes. Universitas Islam Indonesia, Yogyakarta.

[14] Riany, J., Fajar, M., & Lukman, M. P., 2016. Penerapan Deep Learning Analysis pada Angket Penilaian Terbuka Menggunakan K-nearest Neighbor. *Jurnal Sisfo 6*, 147-156.

[15] Buntoro, G. A., 2017. Analisis Sentimen Calon Gubernur DKI Jakarta Tahun 2017 di Twitter. *Integer Journal Maret*, Vol 1, No. 1, 32-41.

[16] Yunita, N., 2016. Analisis Sentimen Berita Artis dengan Menggunakan Algoritma Support Vector Machine dan Particle Swarm Optimization. *J. Sist. Inf. STMIK Antar Bangsa*, Vol V, No. 2, 104-112.

[17] Tiara., Sabariah, M. K., & Effendy, V., 2015. Sentiment Analysis on Twitter using The Combination of Lexicon-Based and Support Vector Machine for Assessing The Performance of a Television Program. *2015 3rd International Conference on Information and Communication Technology (ICoICT 2015)*, 386-390.

[18] Rachimawan, A. F., & Utama, B. S., 2016. Ads Filtering Menggunakan Jaringan Syaraf Tiruan Perceptron. Naïve Bayes Classifier, dan Regresi Logistik. *Jurnal Sains dan Seni ITS*, Vol 5, No 1, 83-89.

[19] Haryanto, D. J, Muflikhah, L., & Fauzi, M. A., 2018. Analisis Sentimen Review Barang Berbahasa Indonesia dengan Metode Support Vector Machine dan Query Expansion. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, Vol 2, No. 9, 2909-2916.