Journal of Innovation Information Technology and Application (JINITA)
Vol. 2, No.01, Juni 2020, pp.9-18
DOI: doi.org/10.35970/jinita.v2i01.119

Analisis Sentimen Kinerja KPU Pemilu 2019 Menggunakan Algoritma K-Means Dengan Algoritma Confix Stripping Stemmer

Sentiment Analysis KPU Performance 2019 Elections Using K-Means Algorithm with Confix Stripping Stemmer

A Sidang Amirul Haj[1], Victor Amrizal[2], Arini*[3]
[1,2,3]Teknik Informatika, Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta
e-mail: agungsidang1@gmail.com[1], victor.amrizal@uinjkt.ac.id[2], arini@uinjkt.ac.id[3]

Abstrak

Kinerja KPU pada Pemilu 2019 ini ramai menjadi perbincangan masyarakat awam maupun elite politik. Banyak masyarakat atau pihak yang berkomentar mengenai proses perhitungan hasil Pemilu tahun 2019 melalui media *social*. *fanpage* Facebook milik KPU pun ramai diserang oleh komentar positif atau *negative*. Analisis sentiment atas pendapat-pendapat masyarakat tersebut dapat diteliti untuk mengetahui seberapa besar presentase yang bersentimen positif dan bersentimen negatif dari kebijakan ini melalui komentar-komentar yang telah dikirim ke sosial-sosial media. Data yang dianalisis yaitu data yang diambil dari Facebook KPU sebanyak 200 komentar yang dibagi menjadi 150 data latih dan 50 data uji. Penelitian ini menggunakan algoritma K-Means dengan nilai k=2 untuk menentukan sentimen akhir yaitu positif dan negatif, algoritma Levensthein Distance untuk normalisasi kata dan algoritma *Confix Stripping Stemmer* pada proses *stemming*. Hasil yang didapatkan dari sentimen masyarakat terhadap kinerja KPU ini yaitu lebih banyak yang bersentimen negatif daripada yang positif. Hasil dari tingkat akurasi yang didapatkan dari penggunaan algortima K-Means saja yaitu 84% dengan nilai akurasi yang lebih rendah dibandingkan dengan kombinasi algoritma diatas yaitu 86%. Saran untuk penelitian selanjutnya sebaiknya menggunakan data yang lebih banyak lagi, dan menggunakan teknik perhitungan akurasi k-fold cross validation sebagai uji coba selanjutnya.

Kata kunci: *confix stripping stemmer, facebook,* KPU, *k-means, levensthein distance,* pemilu, Sentimen analisis

Abstract

The performance of the KPU in the 2019 elections was a lively conversation between the general public and the political elite. Many people or parties commented on the process of calculating the results of the 2019 elections through social media. KPU's Facebook fanpage is also busy being attacked by positive or negative comments. Sentiment analysis of public opinion can be investigated to find out how much percentage of positive and negative sentiment of this policy through comments that have been sent to social-social media. Data analyzed were 200 data taken from Facebook KPU, divided into 150 training data and 50 test data. This study uses the K-Means algorithm with a value of k = 2 to determine the final sentiments of positive and negative, the Levensthein Distance algorithm for word normalization and the Confix Stripping Stemmer algorithm in the stemming process. The results obtained from the public sentiment on the performance of the KPU are more negative than positive. The results of the accuracy obtained from the use of the K-Means algorithm are 84% with a lower accuracy value compared to the combination of the algorithm above, namely 86%. Suggestions for further research should use even more data, and use the k-fold cross validation accuracy calculation technique as a further trial.

***Keywords**: confix stripping stemmer, facebook,* KPU, *k-means, levensthein distance,* pemilu, Sentimen analysis

Pendahuluan

Era globalisasi saat ini sangat memengaruhi pesatnya kemajuan teknologi informasi seperti dalam bidang ekonomi, kebudayaan, seni, pendidikan dan bahkan dunia politik. Di Indonesia pada tahun 2019 ini dari jumlah populasi 268,2 juta penduduk, pengguna internet dan pengguna social media memiliki pencapaian angka yang sama yaitu 150 juta pengguna atau sekitar 56 persen. Begitu pula pertumbuhan internet Indonesia mencapai 17,3 juta pertahun. Hasil itu naik 13 persen dari tahun

*) Penulis Korespondensi : arini@uinjkt.ac.id

sebelumnya, namun angka tersebut masih tergolong paling rendah dibandingkan dengan negara-negara Asia Tenggara lainnya [1].Teknologi membuat jarak tidak lagi menjadi masalah dalam berkomunikasi, dan sekarang sosial media menjadi kebutuhan wajib bagi sebagian kalangan masyarakat. Media sosial adalah media yang digunakan oleh individu agar menjadi sosial, secara daring dengan cara berbagi isi, berita, foto dan lain-lain dengan orang lain [2]. Tidak hanya pengguna sosial media yang semakin hari semakin meroket, namun juga semakin beragam pula jenis dari sosial media yang ditawarkan [3]. Keberadaan media sosial yang beragam tersebut telah membantu masyarakat untuk mendapatkan informasi terbaru terkait peristiwa atau kejadian di lingkungan sekitar ataupun lingkungan yang lebih luas [4]. *Facebook* memiliki beberapa fitur, yaitu : *wall*, status, *chat*, *friend request*, *inbox*, *notification*, *search bar*, dan *games* [5] (Jayanti, Sentinuwo, Lantang, & Jacobus, 2016). *Facebook* tidak hanya digunakan oleh perseorangan saja, akan tetapi organisasi atau lembaga resmi di pemerintahan pun memaksimalkan *Facebook* sebagai salah satu sarana untuk memberikan informasi kepada masyarakat luas. Salah satu lembaga pemerintah yang memanfaatkan *Facebook* untuk memberikan informasi kepada masyarakat yaitu KPU. Kinerja KPU pada Pemilu 2019 ini ramai menjadi perbincangan masyarakat awam maupun elite politik. Analisis sentimen atau *opinion mining* merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini [6]. Analisis sentimen akan mengelompokkan polaritas dari teks yang ada dalam kalimat atau dokumen untuk mengetahui pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif, negatif atau netral [7]. Analisis sentimen dilakukan untuk menentukan apakah opini atau komentar terhadap suatu permasalahan, memiliki kecenderungan positif atau negatif dan dapat dijadikan sebagai acuan dalam meningkatkan suatu pelayanan, ataupun meningkatkan kualitas produk [8]. Besarnya pengaruh dan manfaat dari analisis sentimen menyebabkan penelitian dan aplikasi berbasis analisis sentimen berkembang pesat [9]. Penggunaan analisis sentimen dapat diterapkan pada opini masyarakat terhadap kinerja KPU pada Pemilu 2019. Hal ini disebabkan oleh beberapa faktor seperti penulisan kata yang disingkat, penggunaan bahasa *modern* atau slang, salah dalam mengetik huruf dan tidak baku dalam penulisan opini [10]. Teknik yang berkembang untuk penggalian dokumen teks saat ini adalah text mining [11]. *Text mining* dapat diolah untuk berbagai macam keperluan diantaranya adalah untuk *summarization*, pencarian dokumen teks dan sentimen analisis [9].

Menurut [11] menyatakan bahwa ada beberapa algoritma atau metode yang di gunakan untuk analisis sentimen, antara lain *Naïve Bayes (NB), Support Vector Machine (SVM)* dan *clustering K-Means*, dan dari hasil pengujian disimpulkan bahwa semakin besar *data set* yang digunakan semakin rendah akurasi *K-Means.*

Menurut [12], *Text clustering* berhubungan dengan proses menemukan sebuah struktur kelompok yang belum terlihat (tak terpandu atau *unsupervised*) dari sekumpulan dokumen. Sedangkan *text classification* dapat dianggap proses untuk membentuk golongan (kelas-kelas) dari dokumen berdasarkan pada kelas kelompok yang sudah diketahui sebelumnya (terpandu atau supervised). Sedangkan algoritma adalah sekumpulan intruksi yang jumlahnya terbatas, yang apabila dilasanakan akan menyelesaikan suatu tugas tertentu [13].

K-Means menguji masing- masing komponen tersebut ke salah satu pusat cluster yang telah didefinisikan tergantung dari jarak minimum antar komponen dengan tiap-tiap cluster. Posisi pusat cluster akan dihitung kembali sampai semua komponen data digolongkan kedalam tiap-tiap cluster dan terakhir akan terbentuk posisi cluster baru [14]

Peneliti [10] menggunakan algoritma sebagai normalisasi kata tidak baku menjadi kata baku dengan menggunakan algoritma *Levensthein Distance*. Pada tahap terakhir untuk pengambilan sentimennya menggunakan algoritma *Naïve Bayes Classifier* menghasilkan nilai *accuracy* yang meningkat yaitu sebesar 98,33%. Pada [14], *dataset* dari 10 pantai yang ada di Indonesia sebanyak 500 tweet. Hasil akurasi dari klasifikasi menggunakan algoritma *Support Vector Machine* sebesar 74,39%. Selanjutnya data opini dari kuesioner ditambahkan untuk mengelompokkan pantai berdasarkan ketersediaan sumber daya, fasilitas, akses, kesiapan masyarakat, potensi pasar dan posisi pariwisata. Dalam proses pengelompokan data ini digunakan metode *K-Means.* Sedangkan pada [15] [Somantri, Wiyono, & Dairoh, 2016] menggunakan algoritma *Support Vector Machine (SVM)* dan *K-Means* dihasilkan tingkat akurasi yang lebih baik dengan tingkat akurasi 86,21%, sedangkan dengan *SVM* saja tanpa *K-Means* yaitu 85,38%. Pada penelitian [10] mencari kata kunci dan hubungan pola antar tiap calon Gubernur DKI Jakarta, metode yang digunakan *Naive Bayes Classifier* pada masing-masing pengukuran performa akurasi, *precision, recall*, dan *F-Measure* sebesar 85.77%; 85.90%; 85.77%; 85.67%. Dengan metode *Support Vector Machine kernel RBF* tiap pengukuran performa akurasi*, precision, recall*, dan *F-Measure* adalah 87.80%; 98.48%; 87.80%; 92.64%. Untuk hasil *SNA* didapatkan hasil yang tinggi untuk *degree centrality* sebeasar 0.865 yang menunjukan pengaruh antar node kata kunci dengan kata kunci

yang lain. Peneliti [16] menggunakan 200 data, menggunakan algoritma *K-Means &* algoritma *levensthein distance* dan nilai k=2. Peneliti [17] menggunakan algoritma *Naïve Bayes Classifier* dan seleksi fitur menggunakan *quadgram* dan *trigram* untuk menganalisa sentimen. Pada [18] menggunakan data mining untuk menganalisa pola, dan [19] analisis sentiment menggunakan algoritma SVM, serta [15] juga menggunakan algoritma SVM untuk optimalisasi dan K-Means untuk klasifikasi.

Dari beberapa studi literatur yang telah digunakan, berikut adalah fokus penelitian yang peneliti lakukan yang berbeda dengan peneliti-peneliti sebelumnya yaitu :

1. Komentar *public* yang akan dianalisis sentimennya yaitu mengenai kinerja KPU pada Pemilu 2019 diambil dari *Fanpage Facebook* Komisi Pemilihan Umum RI, dengan data latih 150 dan data uji 50
2. Dalam *stemming* kata atau penguraian kata menjadi kata baku, penelitian ini menggunakan algoritma *Confix Stripping Stemmer*.
3. Dalam fitur pembobotan kata, penelitian ini menggunakan algoritma TF-IDF.
4. Text mining dengan menggunakan proses *pre-processing*
5. Dalam perhitungan tingkat akurasi, penelitian ini menggunakan *Confusion Matrix*.
6. Klasterisasi analisis sentiment positif dan negatif menggunakan algoritma *K-Means* dan dibantu oleh algoritma *Levensthein Distance* sebagai normalisasi kata.
7. Sistem yang dibuat berbasis *website* dengan menggunakan Bahasa Pemrograman PHP dan MySQL sebagai *database*.

Untuk algoritma *Confix Stripping Stemmer*, peneliti mengacu pada [20] dan [21], yang digunakan untuk melakukan proses *stemming* terhadap kata-kata berimbuhan. Algoritma *Confix-stripping stemmer* mempunyai aturan imbuhan sendiri dengan model sebagai berikut :

$$\Big[\big[[AW+]AW+\big]AW+\Big]\,Kata-Dasar\,[[+AK][+KK][+P]$$

Keterangan :

AW : Awalan
AK : Akhiran
KK : Kata ganti kepunyaan
P : Partikel

Sedangkan untuk TF-IDF yang merupakan metode untuk menghitung bobot dari kata yang digunakan peneliti akan menghitung nilai *Term Frequency (TF)* dan *Inverse Document Frequency (IDF)* pada setiap *token* (kata) disetiap dokumen dalam korpus dengan mengacu pada [7].

- *Term frequency* (TF) adalah jumlah kemunculan kata pada suatu dokumen. Semakin banyak suatu kata muncul pada dokumen, maka semakin besar kata tersebut berpengaruh pada dokumen tersebut. Sebaliknya, semakin sedikit suatu kata muncul pada dokumen, maka semakin kecil kata tersebut berpengaruh pada dokumen tersebut.
- *Inverse document frequency* (IDF) adalah pembobotan kata yang didasarkan pada banyaknya dokumen yang mengandung kata tertentu. Semakin banyak dokumen yang mengandung suatu kata tertentu, semakin kecil pengaruh kata tersebut pada dokumen. Sebaliknya, semakin sedikit dokumen yang mengandung suatu kata tertentu, semakin besar pengaruh kata tersebut pada dokumen. Pembobotan menggunakan TF-IDF dijelaskan pada Persamaan [7].

Pre-processing dilakukan untuk menghindari data yang kurang sempurna, gangguan pada data, dan data-data yang tidak konsisten. Tahap *preprocessing* memiliki beberapa proses, yaitu *case folding*, *tokenizing*, *stopword removing* dan *stemming* [7].

a. *Case folding* yaitu pengubahan bentuk huruf menjadi huruf kecil. Hanya huruf a sampai z yang diterima. Karakter selain huruf dihilangkan dan dianggap *delimiter*.
b. *Tokenizing* adalah proses memecah teks menjadi kata tunggal dan penghapusan tanda baca serta angka, sesuai dengan kamus data yang telah ditentuka. Strategi umum yang digunakan pada tahap *tokenizing* adalah memotong kata pada *white space* atau spasi dan membuang karakter tanda baca. Tahap *tokenizing* membagi urutan karakter menjadi kalimat dan kalimat menjadi *token*.
c. *Stopword* adalah proses menghilangkan kata tidak penting dalam *text*. Hal ini dilakukan untuk memperbesar akurasi dari pembobotan *term*. Untuk mengoptimalkan perhitungan frekuensi kemunculan kata pada proses pembobotan maka diperlukan kamus sinonim untuk mengecek kata yang memiliki makna yang sama. Jika kata tersebut ditemukan didalam kamus sinonim maka kata tersebut diubah ke bentuk sinonimnya. Untuk proses ini, diperlukan suatu kamus kata-kata yang bisa

dihilangkan. Dalam Bahasa Indonesia, misalnya kata: dan, atau, mungkin, ini, itu, dll adalah kata-kata yang dapat dihilangkan.

d. *Stemming* adalah pengubahan kata ke bentuk kata dasar atau penghapusan imbuhan. *Stemming* disini menggunakan kamus daftar kata berimbuhan yang mempunyai kata dasarnya dengan cara membandingkan kata-kata yang ada di dalam komentar dengan daftar kamus stem. Sebagai contoh, kata bersama, kebersamaan, menyamai, akan di *stem* ke *root word* nya yaitu "sama". Algoritma *stemming* untuk bahasa yang satu berbeda dengan algoritma stemming untuk bahasa lainnya. Sebagai contoh Bahasa Inggris memiliki morfologi yang berbeda dengan Bahasa Indonesia sehingga algoritma *stemming* untuk kedua bahasa tersebut juga berbeda. Pada teks berbahasa Inggris, proses yang diperlukan hanya proses menghilangkan *sufiks*. Sedangkan pada teks berbahasa Indonesia lebih rumit/kompleks karena terdapat variasi imbuhan yang harus dibuang untuk mendapatkan *root word* dari sebuah kata.

Untuk mengukur kinerja/akurasi pada penelitian ini, yaitu dengan menggunakan *confusion matrix*. *Confusion matrix* adalah suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep data mining. Pada dasarnya *confusion matrix* membandingkan hasil klasifikasi yang dilakukan oleh suatu sistem dengan hasil klasifikasi yang sebenarnya [22].

Untuk proses algoritma *K-Means*, peneliti mengacu pada [23], dengan langkah-langkah :

1. Menentukan *k* sebagai jumlah *cluster* yang ingin dibentuk.
2. Membangkitkan nilai *random* untuk pusat *cluster* awal (*centroid*) sebanyak *k*.
3. Menghitung jarak setiap data *input* terhadap masing-masing *centroid* menggunakan rumus jarak *Euclidean* (*Euclidean Distance*) hingga ditemukan jarak yang paling dekat dari setiap data dengan centroid. Berikut adalah persamaan *Euclidian Distance* :

$$d(xi, \mu j) = \sqrt{\sum (xi - \mu j)^2} \quad (1)$$

Keterangan :

xi : data kriteria

μj : *centroid* pada *cluster* ke-*j*

4. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan centroid (jarak terkecil).
5. Memperbaharui nilai *centroid*. Nilai *centroid* baru diperoleh dari rata-rata *cluster* yang bersangkutan dengan menggunakan rumus:

$$\mu j\ (t+1) = \frac{1}{Nsj} \sum_{j \in sj} xj \quad (2)$$

Keterangan:

$\mu j\ (t+1)$: *centroid* baru pada iterasi ke (t+1)

Nsj : banyak data pada *cluster* Sj.

6. Melakukan perulangan dari langkah 2 hingga 5, sampai anggota tiap *cluster* tidak ada yang berubah. Jika langkah 6 telah terpenuhi, maka nilai pusat *cluster* (μj) pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data

Levenshtein distance atau edit distance merupakan algoritma yang digunakan untuk menghitung jumlah operasi yang paling sedikit antara satu kata dengan kata lain. Hasil perhitungan edit distance didapatkan dari matriks yang digunakan untuk menghitung jumlah perbedaan antara dua string [24].

Pada algoritma *levensthein distance* terdapat 3 macam operasi utama yang dilakukan, yaitu [10] :

1. Operasi Penambahan Karakter, Operasi penambahan karakter yaitu operasi yang digunakan untuk menambahkan karakter ke dalam *string*. Contoh pada penulisan *string* "kern" maka diubah menjadi *string* "keren" dengan menambahkan karakter '**e**'.
2. Operasi Pengubahan Karakter, Operasi pengubahan karakter yaitu operasi yang digunakan untuk mengubah karakter dengan cara menukar sebuah karakter dengan karakter lain. Contoh pada penulisan *string* "hidsp" diubah menjadi *string* "hidup" dengan mengubah karakter '**S**' menjadi karakter '**U**'.
3. Operasi Penghapusan Karakter, Operasi penghapusan karakter yaitu operasi yang digunakan untuk menghapus suatu karakter pada *string*. Contoh pada penulisan *string* "hebatt" diubah menjadi *string* "hebat" dengan menghilangkan salah satu karakter '**T**'.

Menurut penelitian [24], algoritma *Levenshtein distance* berjalan mulai dari pojok kiri atas sebuah *array* dua dimensi (matriks) yang telah diisi sejumlah karakter string awal dan *string* target. Entri pada matriks tersebut merepresentasikan nilai terkecil dari transformasi string awal menjadi *string* target. Entri yang terdapat pada ujung kanan bawah matriks adalah nilai *distance* yang menggambarkan jumlah

perbedaan dua *string*. Berikut ini adalah langkah-langkah algoritma *levenshtein distance* dalam mendapatkan nilai *distance*:

1. Inisialisai (nilai awal)
   a) Hitung panjang S (string awal) dan T (string target), misalkan m dan n.
   b) Buat matriks berukuran 0...m baris dan 0...n kolom.
   c) Inisialisasi baris pertama dengan 0...n.
   d) Inisialisasi kolom pertama dengan 0...m.
2. Proses perhitungan matriks
   a) Periksa S[i] untuk $1 < i < n$
   b) Periksa T[j] untuk $1 < j < m$
   c) c. Jika S[i] = T[j], maka entrinya adalah nilai yang terletak pada tepat didiagonal atas sebelah kiri, yaitu d[i,j] = d[i-1,j-1]
   d) Jika S[i] ≠ T[j], maka entrinya adalah d[i,j] minimum dari:
      - Nilai yang terletak tepat diatasnya, ditambah satu, yaitu d[i,j-1]+1.
      - Nilai yang terletak tepat dikirinya, ditambah satu, yaitu d[i-1,j]+1.
      - terletak pada tepat didiagonal atas sebelah kirinya, ditambah satu, yaitu d[i-1,j-1]+1.
3. Hasil entri matriks pada baris ke-i dan kolom ke j, yaitu d[i,j].
4. Langkah 2 diulang hingga entri d[m,n] ditemukan.

Metode Penelitian

Pada penelitian ini peneliti mengumpulkan data dan informasi menggunakan studi lapangan dan studi pustaka. Observasi atau pengamatan secara langsung dan mengambil data dilakukan dari bulan Agustus 2019 – Oktober 2019 melalui *Facebook* secara manual yaitu komentar masyarakat mengenai kinerja KPU pada Pemilu 2019 dari *Facebook Page* KPU RI (Republik Indonesia). Didapat data sebanyak 200 komentar. Setelah mendapatkan data, peneliti melakukan pelabelan sentimen secara manual menggunakan 150 data secara *random* sebagai data latih.

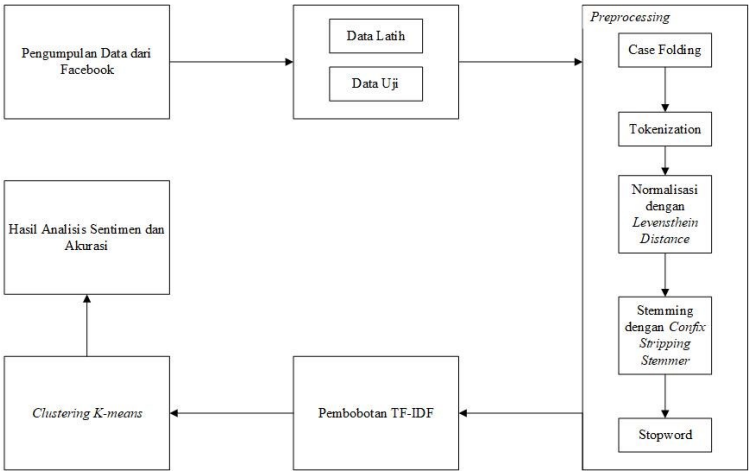
Sedangkan untuk proses mendapatkan hasil sentimen masyarakat dari objek yang diteliti yaitu kinerja KPU di pemilu 2019, dengan menggunakan metode simulasi yaitu dengan tahapan :

1. Formulasi Masalah (*Problem Formulation*)
2. Model Pengkonsepan (*Conceptual Model*)
3. Data Masukan/Keluaran (*Input/Output Data*)
4. Pemodelan (*Modelling*)
5. Simulasi (*Simulation*)
6. Verifikasi dan Validasi (*Verification and Validation*)
7. Eksperimentasi (*Experimentation*)
8. Analisis Keluaran (*Output Analysis*).

Formulasi Masalah

Data yang digunakan pada penelitian ini adalah komentar berbahasa Indonesia tentang kinerja KPU pada Pemilu 2019 yang diambil dari *Facebook page,* dalam proses penentuan sentimen masyarakat mengkombinasikan antara algoritma *k-mean,* algoritma *levensthein distance* dan *confix stripping stemmer.*

Model Pengkonsepan



Gamb ar 1. Alur Model Pengkonsepan

Konsep pertama membuat konsep pada proses *text mining* yang ingin digunakan. Kedua, mengidentifikasikan *input*, yaitu komentar-komentar masyarakat terkait kinerja KPU pada Pemilu 2019 dari *facebook page* KPU RI, kemudian komentar yang telah dikumpulkan kemudian diolah dan diproses dengan secara manual untuk pelabelan terhadap data latih. Ketiga, membuat konsep untuk tahap uji pada skenario 1 yaitu dengan melihat hasil sentimen dan tingkat akurasi menggunakan algoritma *K-Means* saja. Keempat, tahap uji pada skenario 2 yaitu melihat hasil sentimen dan tingkat akurasi menggunakan kombinasi algoritma *K-Means* dan dibantu algoritma *Levensthein Distance* sebagai normalisasi kata serta *confix stripping stemmer* sebagai *stemming* kata pada tahap *pre-processing*.

Data Masukan/Keluaran

Pada penelitian ini data yang diperoleh sebagai data *input* yaitu komentar masyarakat terhadap kinerja KPU pada Pemilu 2019 yaitu sebanyak 200 data yang diambil dari komentar posting *Facebook* page KPU RI. Hasil atau *output* yang didapatkan dari penelitian ini yaitu sentimen akhir dari suatu komentar dan hasil akurasi pada skenario pertama dan kedua.

Pemodelan (Modelling)

Pada tahapan ini peneliti menentukan model skenario yang akan digunakan. Pada tahap ini penelitian ini melakukan pemodelan dalam membuat rancangan sistem yang akan dibuat secara manual. Pemodelan atau skenario yang dibuat yaitu skenario kombinasi antara algoritma *K-Means* dan algoritma *Levensthein Distance* dan algoritma *Confix Stripping Stemmer* serta skenario tanpa menggunakan algoritma *Levensthein Distance* dan *Confix Stripping Stemmer* (hanya menggunakan algoritma *K-Means* dan saja).

Simulasi

Sistem akan dijalankan untuk mensimulasikan kinerja masing-masing algoritma sesuai dengan konsep dan skenario yang telah ditentukan sebelumnya. Simulasi yang akan dilakukan adalah dengan melakukan input *dataset* latih dan uji, melakukan pelabelan terhadap data latih secara manual untuk dikelompokkan sentimennya, melakukan pelatihan terhadap data latih dan melakukan *clustering* data uji. Hasil simulasi berupa perbandingan akurasi dari algoritma yang dijadikan penelitian, kemudian akan dicatat dan kemudian dilakukan tahap verifikasi.

Verifikasi dan Validasi

Verifikasi dilakukan untuk memastikan bahwa setiap tahapan pada sebelumnya saling memiliki hubungan, dalam hal ini setiap tahapan diulas kembali untuk memastikan tiap tahap tersebut saling terkait. Pada tahap verifikasi dilakukan untuk memastikan adanya kesalahan atau tidak yang terjadi dalam beberapa tahapan atau proses simulasi. Sedangkan tahapan validasi dilakukan untuk memastikan kesesuaian proses simulasi yang dibuat berdasarkaan model pengkonsepan dengan formulasi masalah yang dibuat. Pada intinya, verifikasi dan validasi bertujuan untuk menyakinkan hasil dari aplikasi sentimen ini sesuai dengan yang dikonsepkan sebelumnya.

Eksperimentasi

Eksperimen yang dilakukan yaitu dengan membandingkan hasil skenario yaitu hasil sentimen data uji menggunakan algoritma *K-Means* dan kombinasi algoritma *K-Means*, algoritma *Levensthein Distance*, serta algoritma *Confix Stripping Stemmer*. Parameter yang digunakan untuk nilai k pada algoritma K-Means yaitu k = 2. Eksperimen disini bertujuan untuk mengevaluasi hasil simulasi aplikasi. Dengan 150 data latih dan 50 data uji

Analisis Keluaran (Output Analysis)

Peneliti melakukan analisa terhadap *output* berdasarkan skenario yang sudah dilakukan yaitu menggunakan algoritma *K-Means* dan kombinasi algoritma *K-Means,* algoritma *Levensthein Distance*, serta algoritma *Confix Stripping Stemmer* beserta hasil tingkat akurasi dari setiap skenario tersebut.

Hasil dan Pembahasan

Pada tahap analisis keluaran, dilakukan analisis terhadap hasil sentimen masyarakat terhadap kinerja KPU pada Pemilu 2019 menggunakan algoritma *K-Means* dan kombinasi algoritma *K-Means,* algoritma *Levensthein Distance*, serta algoritma *Confix Stripping Stemmer* beserta hasil tingkat akurasi dari setiap scenario tersebut.

Skenario 1 : Hasil Sentimen Algoritma K-Means Dan Kombinasi Algoritma K-Means, Algoritma Levensthein Distance, Serta Algoritma Confix Stripping Stemmer

Pengujian dilakukan menggunakan algoritma *K-Means* dan kombinasi algoritma *K-Means,* algoritma *Levensthein Distance*, serta algoritma *Confix Stripping Stemmer* dibandingkan dengan hasil sentimen yang dilakukan secara manual sebanyak 50 data uji.

Tabel 1. Hasil Sentimen dari Skenario Pada Data Uji

Data ke-n	Algoritma *K-Means*	Kombinasi Algoritma *K-Means* dan Algoritma *Levensthein* dan Algoritma *Confix*	Sentimen sebenarnya
1	Positif	Positif	Positif
2	Negatif	Negatif	Positif
3	Positif	Positif	Positif
4	Positif	Positif	Positif
5	Negatif	Negatif	Negatif
6	Positif	Positif	Positif
7	Negatif	Negatif	Positif
8	Negatif	Negatif	Negatif
9	Negatif	Negatif	Negatif
10	Negatif	Negatif	Negatif
11	Negatif	Negatif	Negatif
12	Negatif	Negatif	Negatif
13	Negatif	Negatif	Negatif
14	Negatif	Negatif	Negatif
15	Negatif	Negatif	Negatif
16	Negatif	Negatif	Negatif
17	Positif	Positif	Positif
18	Negatif	Negatif	Positif
19	Negatif	Negatif	Negatif
20	Positif	Positif	Positif
21	Positif	Positif	Positif
22	Positif	Positif	Positif
23	Negatif	Negatif	Negatif
24	Negatif	Negatif	Negatif
25	Negatif	Negatif	Positif
26	Positif	Positif	Positif
27	Negatif	Negatif	Positif
28	Positif	Positif	Positif
29	Negatif	Negatif	Negatif
30	Negatif	Negatif	Negatif
31	Negatif	Negatif	Negatif
32	Negatif	Negatif	Negatif
33	Negatif	Negatif	Negatif
34	Negatif	Negatif	Negatif
35	Negatif	Negatif	Negatif
36	Positif	Positif	Positif
37	Negatif	Negatif	Negatif

38	Negatif	Negatif	Negatif
39	Positif	Positif	Positif
40	Positif	Negatif	Negatif
41	Negatif	Negatif	Negatif
42	Positif	Positif	Positif
43	Negatif	Negatif	Negatif
44	Positif	Positif	Positif
45	Positif	Positif	Negatif
46	Positif	Positif	Positif
47	Positif	Positif	Positif
48	Negatif	Negatif	Negatif
49	Positif	Positif	Positif
50	Positif	Positif	Positif

Dari Tabel 1 dihasilkan untuk skenario 1 menghasilkan 42 data uji yang sesuai dengan manual dan 8 data uji yang tidak sesuai. Sedangkan pada skenario 2 terdapat 43 data uji yang sesuai dengan manual dan 7 data uji yang tidak tidak sesuai.

Analisis Hasil Akurasi Algoritma K-Means

Tabel 2. Hasil pengujian Skenario 1

	Actual Positive (+)	**Actual Negative (-)**
Prediction Positive (+)	True positive = 18	False Positive = 3
Prediction Negative (-)	False negative = 5	True negative = 24

Dari Tabel 2, sentimen dari masyarakat terhadap kinerja KPU pada Pemilu 2019 yaitu lebih banyak yang bersentimen negatif dibanding yang bersentimen positif. Nilai akurasi dari skenario 1 yaitu perbandingan antara hasil sentimen dari algoritma *K-Means* tanpa menggunakan algoritma normalisasi kata dan algoritma *stemming* serta hasil sentimen secara manual mendapatkan tingkat akurasi 84%, perhitungan hasil bisa dilihat dibawah ini.

$$\frac{Jumlah\ data\ yang\ diperiksa\ benar}{Total\ data\ yang\ diperiksa} = \frac{(18+24)}{(18+3+5+24)} \times 100\% = 84\%$$

Analisis Hasil Akurasi Kombinasi Algoritma K-Means Dan Algoritma Levensthein Distance Dan Algoritma Confix Stripping Stemmer Dan Sentimen Manual

Tabel 3. Hasil pengujian Skenario 2

	Actual Positive (+)	**Actual Negative (-)**
Prediction Positive (+)	True positive = 18	False Positive = 2
Prediction Negative (-)	False negative = 5	True negative = 25

Dari Tabel 3. sentimen dari masyarakat terhadap kinerja KPU pada Pemilu 2019 yaitu lebih banyak yang bersentimen negatif dibanding yang bersentimen positif. Nilai akurasi dari skenario 2 yaitu perbandingan antara hasil sentimen dari kombinasi algoritma *K-Means,* algoritma normalisasi kata yaitu algoritma *Levensthein distance* dan algoritma stemming yaitu algoritma *Confix Stripping Stemmer* serta hasil sentimen secara manual mendapatkan tingkat akurasi 86%, lebih tinggi dibandingkan skenario 1 karena menggunakan algoritma normaliasi kata dan algoritma *stemming*. Perhitungan hasil bisa dilihat dibawah ini.

$$\frac{Jumlah\ data\ yang\ diperiksa\ benar}{Total\ data\ yang\ diperiksa} = \frac{(18+25)}{(18+2+5+25)} \times 100\% = 86\%$$

Berdasarkan hasil akurasi tabel 2 dan 3, didapatkan hasil bahwa pada kombinasi algoritma *K-Means,* algoritma *Levensthein Distance* serta algoritma *Confix Stripping Stemmer* memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan algoritma *K-Means* saja. Pada skenario 1 (menggunakan algoritma *K-Means* saja) mendapatkan nilai akurasi sebesar 84%. Pada skenario 2 (menggunakan algoritma *K-Means,* algoritma *Levensthein Distance* dan algoritma *Confix Stripping Stemmer* mendapatkan nilai akurasi sebesar 86%. Hal ini disebabkan karena di dalam data uji ditemukan suatu kata yang *typo* dan di skenario 1 ini tidak adanya proses normalisasi kata dan *stemming*, setelah dilakukan pencocokan data pada data latih yang mengandung kata tersebut (*term frekuensi*) = 0. Berbeda halnya pada skenario 2, kata yang *typo* ini akan dilakukan normalisasi kata menggunakan algoritma *Levensthein Distance* dan pengubahan kata imbuhan menjadi kata dasar atau *stemming* menggunakan algoritma *Confix Stripping Stemmer*. Hasilnya untuk kata yang telah diperbaiki oleh algoritma *Levensthein Distance* dan *Confix Stripping Stemmer* ini setelah dilakukan pencocokan data pada data latih yang mengandung kata tersebut (*term frekuensi*) > 0. Hasil pencocokan itu berpengaruh pada nilai pembobotan kata dan inilah yang menyebabkan hasil sentimen dari proses *K-Means* mengalami perubahan dan peningkatan akurasi dibandingkan dengan skenario 1.

Kesimpulan

Hasil penelitian ini yaitu tingkat akurasi yang didapatkan dari 2 skenario. Skenario 1 hanya menggunakan algoritma *K-Means* tanpa bantuan algoritma normalisasi kata dan algoritma stemming didapatkan nilai akurasinya yaitu 84% sedangkan pada skenario 2 menggunakan kombinasi algoritma *K-Means,* algoritma *Levensthein Distance* sebagai algoritma normalisasi kata serta algoritma *Confix Stripping Stemmer* sebagai algoritma stemming mengalami peningkatan akurasi yaitu 86%. Pada penelitian analisis teks diharuskan menggunakan fitur tambahan berupa normalisasi kata dan stemming yang baik. Masyarakat pada umumnya, mereka menggambarkan ekspresi ataupun pendapat mereka menggunakan kata – kata singkatan yang tidak ada di dalam KBBI (jika berbahasa Indonesia). Dengan menggunakan fitur normalisasi kata dan stemming, peneliti bisa mengekstrak informasi dengan lebih jelas dan akurat dari data–data yang sudah diambil melalui target sumber.

Daftar Pustaka

[1] We Are Social, “Digital in 2019,” 2019, Tersedia : https://wearesocial.com/global-digital-report-2019 [diakses 2 Januari 2020]

[2] I. Buyung and A. Raharja, “Pengaruh Pensaklaran Video Otomatis (Video Automatic Switch Effect),” *Teknol. Inf.*, vol. VIII, no. 23, pp. 57–74, 2013.

[3] N. D. Mentari, M. A. Fauzi, and L. Muflikhah, “Analisis Sentimen Kurikulum 2013 Pada Sosial Media Twitter Menggunakan Metode K-Nearest Neighbor dan Feature Selection Query Expansion Ranking,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 8, pp. 2739–2743, 2018.

[4] M. Ariful Furqon, D. Hermansyah, R. Sari, A. Sukma, Y. Akbar, and N. A. Rakhmawati, “Analisis sosial media pemerintah daerah di indonesia berdasarkan respons warganet,” pp. 2–4, 2018.

[5] G. A. Buntoro, “Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter,” *Integer J. Maret*, vol. 1, no. 1, pp. 32–41, 2017.

[6] A. R. T. Lestari, R. S. Perdana, and M. A. Fauzi, “Analisis Sentimen Tentang Opini Pilkada Dki 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1718–1724, 2017.

[7] Oktinas and Willa, “Analisis Sentimen Pada Acara Televisi Menggunakan Improved K-Nearest Neighbor,” pp. 5–30, 2017.

[8] W. E. Nurjanah, R. S. Perdana, and M. A. Fauzi, “Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 1, no. 12, pp. 1750–1757, 2017.

[9] F. N. Hasan and M. Wahyudi, “Analisis Sentimen Artikel Berita Tokoh Sepak Bola Dunia Menggunakan Algoritma Support Vector Machine Dan Nakive Bayes Berbasis Particle Swarm Optimization,” *Director*, vol. 15, no. 2, pp. 2017–2019, 2018.

[10] P. Antinasari, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1733–1741, 2017.

[11] Budi, S. 'Text Mining Untuk Analisis Sentimen Review Film Menggunakan Algoritma K-Means, " *Jurnal teknologi Informasi*, Vol. *16, No.* 1, p.1–8, 2017

[12] Nuansa, E. P, "Analisis Sentimen Pengguna Twitter Terhadap Pemilihan Gubernur DKI Jakarta Dengan Metode Naïve Bayesian Classification Dan Support Vector Machine," 2017.

[13] Sofiyanto, A., & Sw, "Perancangan Aplikasi Pertukaran Mata uang Asing Berbasis Android, " *Teknokris*, Vol.11, No.11, p. 50–57, 2017.

[14] Syaifudin, Y. W., & Irawan, R. A. "Implementasi Analisis Clustering Dan Sentimen Data Twitter Pada Opini Wisata pantai Menggunakan K-Means, " *JIP*, Vol. *4*, No. 3, p. 189–194, May 2018.

[15] Somantri, O., Wiyono, S., & Dairoh, "Optimalisasi Support Vektor Machine (SVM) Untuk Klasifikasi Tema Tugas Akhir Berbasis K-Means*, " TELEMATIKA*, Vol. 13, No. 02, Juli 2016, Pp. 59 – 68.

[16] Haris, M., "Analisis Sentimen Komentar Masyarakat Terhadap Kebijakan Pemerintah Tentang Sistem Zonasi Sekolah Menggunakan Algoritma K-Means dan Algoritma Levenshtein Distance, " S.Kom. Skripsi, Teknik Informatika, FST, UIN Syarif Hidayatullah Jakarta, 2019

[17] Kahfi, M., "Analisis *Sentimen Komentar Kebijakan Full Day School (FDS) dari Facebook Page Kemendikbud RI Menggunakan Algoritma Naïve Bayes Classifier, "* S.Kom. Skripsi, Teknik Informatika, FST, UIN Syarif Hidayatullah Jakarta, 2017

[18] Jayanti, L., Sentinuwo, S. R., Lantang, O. A., & Jacobus, A., " Analisa Pola Penyalahgunaan Facebook Sebagai Alat Kejahatan Trafficking Menggunakan Data Mining ", *Jurnal Teknik Informatika*, Vol. *8, No* 1, 2016.

[19] Maulana, M. A., Setyanto, A., & Kurniawan, M. P., "Analisis Sentimen Media Sosial Universitas AMIKOM Yogyakarta Sebagai Sarana Penyebaran Informasi Menggunakan Algoritma Klasifikasi SVM," *Semnasteknomedia Online*, Vol. 6, No. 1, p.7–12, 2018.

[20] Ariadi, D., & Fithriasari, K., "Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer", *JURNAL SAINS DAN SENI ITS Vol. 4, No.2*, p. 248–253, 2015.

[21] Adriani, M., Asian, J., Nazief, B., Williams, H. E., & Tahaghoghi, S. M. M. (2007). Stemming Indonesian: A confix-stripping approach. *Conferences in Research and Practice in Information Technology Series*, *38* (September 2018), 307–314.

[22] Anam, C., & Santoso, H. B., "Perbandingan Kinerja Algoritma C4 . 5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa, " *Energy Jurnal Ilmiah Ilmu-ilmu Teknik*, Vol. *8, No.* 1, p. 13–19, Mei 2018.

[23] Rohmawati W, N., Defiyanti, S., & Jajuli, M., "Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa", JITTER Jurnal Ilmiah Teknologi Informasi Terapan, Vol. I, No. 2, p. 62–68, April 2015,

[24]. Gunawan, F., Fauzi, M., & Adikara, P., "Analisis Sentimen Pada Ulasan Aplikasi Mobile Menggunakan Naive Bayes dan Normalisasi Kata Berbasis Levenshtein Distance (Studi Kasus Aplikasi BCA Mobile)," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, Vol. 1, No.* 10, 1082-1088, 2017.