Penerapan Algoritma *K-Nearest Neighbors* untuk Analisis Sentimen pada Buletin APTIKOM

Yogi Firman Alfiansah
Universitas Buana Perjuangan
Karawang, Indonesia
If17.yogialfiansah@mhs.ubpkarawang.ac.id

Amril Mutori Siregar
Universitas Buana Perjuangan
Karawang, Indonesia
Amril.mutoi99@gmail.com

Anis Fitri Nur Masruriyah
Universitas Buana Perjuangan
Karawang, Indonesia
Anis.masruriyah@ubpkarawang.ac.id

***Abstract*—**
Membaca menjadi salah satu hal mendasar yang cukup penting dalam pembelajaran dan untuk menambah pengetahuan. Berbagai ilmu pengetahuan bisa didapatkan dengan membaca dan membaca juga dapat mengantarkan pada kesuksesan. Permasalahan yang pada saat ini yaitu masih kurangnya minat daya tarik dalam membaca, maka dari itu APTIKOM membuat sebuah media cetak dan daring yang dapat menarik minat baca yaitu Buletin. Namun, belum dapat dipastikan sentimen penulisan dari buletin apakah banyak mengandung kalimat positif atau negatif. Maka dari itu, dibutuhkan sebuah metode khusus untuk mengkategorikan secara otomatis isi dari Buletin tersebut banyak mengandung kalimat positif atau negatif. Data yang diperoleh dari Buletin merupakan sebuah data berbentuk teks atau kalimat yang akan diklasifikasi menggunakan algoritma *K-Nearest Neighbors*. Untuk mendapat hasil analisis sentimen, dokumen Buletin APTIKOM di *filtering* terlebih dahulu melalui tahapan *text preprocessing*. Setelah melalui tahapan *text preprocessing*, data tersebut diolah analisis sentimennya dan mendapatkan sebanyak lebih dari 150 kalimat yang mengandung sentimen positif dan tidak lebih dari 50 kalimat yang mengandung sentimen negatif dan netral. Hasil pengklasifikasian dengan algoritma *K-Nearest Neighbors* yaitu mendapatkan nilai K yang optimal berdasarkan nilai akurasi yaitu K=5 dan di evaluasi dengan *Confusion Matrix* sehingga mendapatkan nilai *Accuracy* 86.2%.

Kata kunci — Analisis Sentimen, Buletin Aptikom, K-Nearest Neighbors, R Studio

I. PENDAHULUAN

Membaca merupakan hal mendasar untuk menggali informasi dan meningkatkan pengetahuan. Kurangnya pengetahuan dalam informasi terbaru sering kali disebabkan karena kurangnya minat membaca [1]. Berdasarkan masalah tersebut Asosiasi Pendidikan Tinggi Informartika (APTIKOM) membuat sebuah media cetak yang bersifat daring untuk menarik minat baca. Media cetak yang diterbitkan oleh APTIKOM ini membahas dan menyebarkan informasi tentang teknologi yang disebut buletin. Buletin ini ditujukan untuk menarik minat daya baca karena buletin bersifat daring sehingga mudah diakses dimana saja dan juga menggunakan bahasa yang lebih sederhana dibandingkan jurnal.

Buletin diterbitkan setiap bulannya dengan judul atau tema yang berbeda beda.Sehingga, edisi buletin pun sudah banyak dan respon dari pembaca pun cukup baik. Dokumen buletin ini lah yang akan di analisis untuk diketahuii sentimennya dengan analisis sentimen. Analisis sentimen merupakan suatu proses yang ada pada *Text Mining* yang biasanya digunakan untuk mengelompokkan suatu sentimen pada teks, pengelompokkan tersebut dilakukan untuk mengetahui suatu sentimen pada dokumen buletin apakah bersifat positif, negatif atau netral. Analisis sentimen juga biasanya digunakan untuk menilai suatu pelayanan, kebijakan, *cyber bullying*, dan masih banyak lagi [2]. Pada penelitian ini sentimen digunakan untuk bahan evaluasi buletin APTIKOM edisi selanjutnya supaya edisi buletin selanjutnya dapat menyampaikan informasi lebih *maximal* dan informasi pun dikemas sesuai judul atau tema yang ada.

Sentimen digunakan untuk melakukan proses analisis, memahami dan mengklasifikasi sebuah teks secara otomatis untuk menghasilkan informasi yang akan membagi klasifikasi dokumen tekstual yaitu sentimen positif, negatif, atau netral. Penelitian terkait dilakukan oleh [3], Text Mining dapat melakukan pengembangan profil iklan pekerjaan industri 4.0. *Text Mining* pada penelitan terkait ini diterapkan pada iklan pekerjaan yang tersedia untuk umum karena sering digunakan sebagai salursan guna mengumpulkan informasi yang relevan tentang pengetahuan dan keterampilan yang dibutuhkan oleh industri. Penelitan terkait selanjutnya yaitu tentang mencari opini sentimen positif dan negatif pada suatu produk kosmetik dengan metode *Naive Bayes* sehingga mendapatkan nilai akurasi 90,50% dan AUC 0.715 [4], Penelitian terkait selanjutnya membahas penilaian sebuah opini atau sentimen masyarakat terhadap produk hijab dengan algoritma *Naive Bayes* [5], Penelitian selanjutnya membahas identifikasi pada tema utama yang akan dibahas di masa lalu dan melacak evolusinya dari waktu ke waktu menggunakan *text mining* [6]. Berdasarkan paparan masalah yang sudah dijelaskan, penelitian ini ditujukan untuk mendapatkan sentimen penulisan pada buletin APTIKOM lalu diklasifikasi sehingga mendapatkan nilai *Accuracy*.

II. DATA DAN METODE

A. Bahan dan Peralatan Penelitian

Konsep metode pada penelitian ini diawali dengan mengumpulkan data terlebih dahulu. Data diperoleh dari buletin yang diterbitkan setiap bulannya oleh Aptikom. Dataset atau bahan penelitian ini berupa kumpulan teks pada buletin. Tahapan penelitian membutuhkan *hardware dan software* tentunya untuk membantu proses penelitian, Adapun *hardware* dan *software* yang digunakan adalah sebagai berikut:

1) *Hardware* / Perangkat Keras

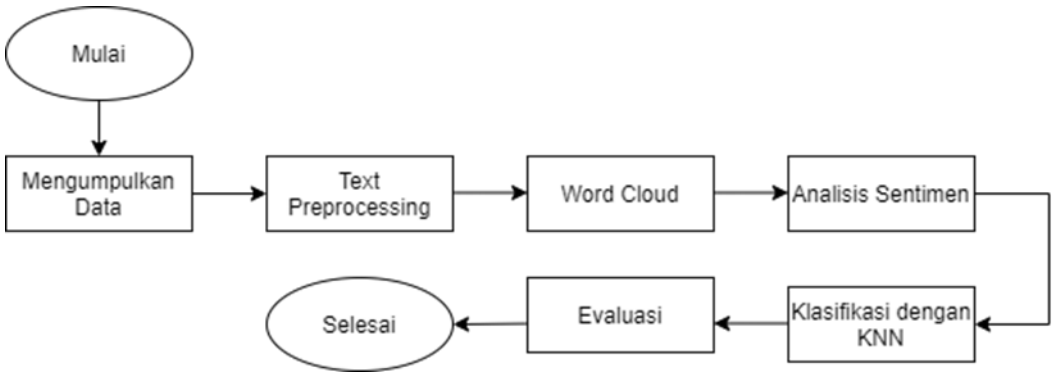
- Laptop Acer E5-475G, processor (Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz (4 CPUs), ~2.4GHz) RAM 8,00GB, Harddisk 1TB, dan sistem operasi Windows 10

2) *Software* / Perangkat Lunak

- R Studio
- Google Chrome
- Microsoft Office 2016

B. Prosedur Penelitian

Serangkaian alur proses pada penelitian yang dilakukan secara sistematis untuk mencapai hasil dan tujuan penelitian dimulai dari pengumpulan data hingga tahapan evaluasi. Berikut merupakan gambaran dari alur proses penelitian :



Gambar 1 Prosedur Penelitian

Penelitian dimulai dengan tahapan mengumpulkan data yang diperoleh dari dokumen buletin Aptikom yang berjudul *Artificial Intellegence*. Kemudian, data diolah dengan tahapan *Text Preprocessing* agar terstruktur dengan baik dan dapat menghilangkan *noise*. Setelah mendapatkan data yang telah diolah dengan *Text Preprocessing* data di proses untuk mendapatkan analisis sentimennya sehingga mendapatkan sentimen positif, negatif, atau netral.

III. HASIL DAN PEMBAHASAN

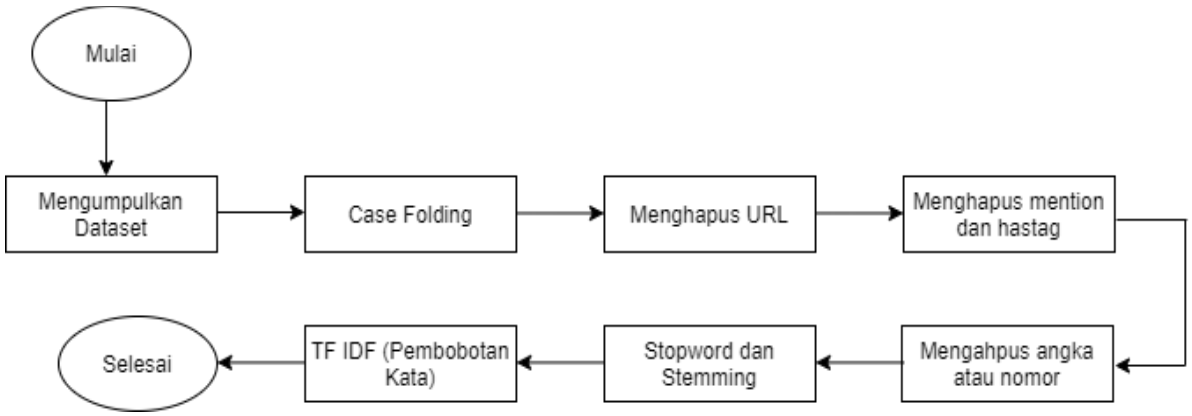
Pada bab ini membahas mengenai tahapan, proses dan hasil pada penelitian dan juga membahas teknik evaluasi dengan *confusion matrix*. Pada pengujian dengan *Confusion Matrix* nantinya akan menghasilkan akurasi dari data yang sudah diproses. Sebelum ke tahap pengujian, tentunya dataset harus melewati beberapa tahapan terlebih dahulu seperti *Text Preprocessing*, Analisis Sentimen, sampai klasifikasi dengan algoritma K-NN.

A. R Studio

R Studio merupakan bahasa pemrograman yang dilakukan pada penelitian ini. R Studio banyak digunakan untuk melakukan statistika komputasi, grafik, dan juga analisis sentimen karena R Studio cukup mudah dalam mengunduh paket-paket atau *library* yang dibutuhkan

B. *Text Preprocessing*

Text Preprocessing merupakan sebuah tahapan untuk dilakukannya seleksi atau tahap *filtering* pada dokumen atau dataset, karena tidak semua teks dapat diproses orientasi sentimennya. *Text Preprocessin* terdiri dari *Case Folding*, menghapus *url*, pemisahan kata penghubung (*tokenization*), *stopword* dan *stemming*, pembobotan kata atau TF IDF. Berikut gambaran alur tahapan *Text Preprocessing* :



Gambar 2 Proses *Text Preprocessing*

1. Mengumpulkan Dataset
   Mengumpulkan dataset dari buletin Aptikom yang berjudul *Aritificial Intellegence*. Dataset yang diambil berbentuk sebuah teks sebanyak 188 kalimat.
2. *Case Folding*
   Case Folding merupakan tahapan awal pada proses *text preprocessing*. Tahapan ini untuk merubah huruf kapital menjadi huruf kecil. Karena, tidak semua dokumen teks dapat konsisten dalam penggunaan huruf. *Case Folding* ini digunakan untuk mengkonversi keseluruhan teks yang ada pada dokumen menjadi bentuk standar.

Sebelum	Sesudah
Dengan perkembangan teknologi yang semakin pesat dan arus pekerjaan yang semakin kencang, tentunya manusia membutuhkan bantuan tambahan dari selain sesama manusia	dengan perkembangan teknologi yang semakin pesat dan arus pekerjaan yang semakin kencang, tentunya manusia membutuhkan bantuan tambahan dari selain sesama manusia

3. Menghapus *URL*
   Proses ini digunakan untuk menghapus link *url* dan juga kata-kata yang tidak relevan karena dapat menganggu tingkat akurasi pada sistem.

Sebelum	Sesudah
SlideShare ,(2020 Juni 27),[Berkas Teks], Diambil dari : https://www.slideshare.net/ikhsanmahruri/sistem-pakar-fuzzy-logic- 85898226	slideshare ,(2020 juni 27),[berkas teks], diambil dari : 85898226

4. Menghapus *mention* dan *hastag*
   Proses ini digunakan untuk menghapus tanda baca seperti *mention* dan *hastag*. Tahapan ini dilakukan karena dapat menganggu tingkat akurasi pada sistem.

Sebelum	Sesudah
Medium.com,(2020 Juni 27),Gambar 8, Diambil dari : https://medium.com/@krismasuccess98/belajar-artificial-intelligencepart-3-8ab18b47384b	medium.com,(2020 juni 27),gambar 8, diambil dari :

5. Menghapus angka atau nomo
   Pada proses ini berguna untuk menghilangkan angka atau nomor. Karena sistem hanya mengolah dataset yang berbentuk teks. Pada proses *filtering* ini benar benar membesihkan semua tanda baca, *mention*, *hastag*, dan juga angka.

Sebelum	Sesudah
Sejarah dari AI dimulai sekitar tahun 1940, namun pada tahun 1956 adalah awal dari AI dan sering disebut dengan The Beginning of AI	sejarah dari ai dimulai sekitar tahun namun pada tahun adalah awal dari ai dan sering disebut dengan the beginning of ai

6. *Stopword* dan *Stemming*
   Stopword berguna untuk menghilangkan kalimat atau teks yang dianggap tidak berguna dan tahapan ini juga dapat menekan waktu pada saat menentukan hasil. Kelemahan *text mining* pada tahapan *stopword* ini yaitu dibutuhkannya kamus yang terus *up to date* agar dapat membaca semua bentuk teks yang berada pada dokumen. Selanjutnya yaitu tahapan *Stemming* yang digunakan untuk menghilangkan kata yang memiliki imbuhan. Semua kata yang memiliki imbuhan akan diubah pada proses *stemming* menjadi kata dasar.

Sebelum	Sesudah
Dengan perkembangan teknologi yang semakin pesat dan arus pekerjaan yang semakin kencang, tentunya manusia membutuhkan bantuan tambahan dari selain sesama manusia	kembang teknologi pesat arus kerja kencang manusia butuh bantu manusia

7. TF IDF (Pembobotan Kata)
   TF IDF merupakan tahapan terakhir pada *text preprocessing* yang berguna untuk mengelompokkan kata yang sering muncul pada dokumen. Selain itu TF IDF juga dapat menghitung nilai bobotnya dokumen tersebut. Penelitian ini terdapat dua jenis perhitungan yaitu TF (*Term Frequency*) atau pembobotan lokal dan menggunakan IDF (*Invers Dokumen Frequency*).

$$TF = \begin{cases} 1 + log_{10}(tf_{t,d}), if\ tf_{t,d} > 0 \\ 0, if tf_{t,d} = 0 \end{cases} \quad \textbf{(1)}$$

$$IDF = \log\left(\frac{N}{dft}\right) \quad \textbf{(2)}$$

   Keterangan :
   $tf_{t,d}$: Jumlah kemunculan *term* (t) pada dokumen (d), jika tidak ada *term* atau t=0, maka TF menjadi 0
   N : Jumlah dokumen pada teks
   df_t : Jumlah dokumen yang mengandung *term* (t)

   Perkalian pada perhitungan TF dan IDF dapat menghasilkan bobot kata yang disebut TD IDF.

$$W_{t,d} = TF\ X\ IDF \quad \textbf{(3)}$$

   Berikut merupakan hasil dari proses TF IDF.

Word	N	Total	TF	IDF	TF IDF
Dalam	41	41	1.000000000	3.091042	3.09104245
Teknologi	13	13	0.500000000	3.091042	1.54552123

   Keterangan:
   Word : Kata dasar yang sudah dilakukan *text preprocessing*
   N dan total : Jumlah dari kosa kata yang ada
   TF : *Term Frequency* pada dokumen n
   IDF : *Inverse Document Frequency*
   TF IDF : Jumlah dokumen dalam koleksi dokumen yang mengandung kosa kata

C. Implementasi

1. Hasil Analisis Sentimen
   Tahap implementasi pada penelitian ini yang pertama untuk mendapat hasil sentimennya apakah positif, negatif, atau netral. Dalam tahapan ini dibutuhkannya kamus sentimen dengan kosa kata yang banyak agar hasil sentimen pun lebih akurat. Penelitian ini menggunakan kamus sentimen yang diperoleh dari *library* sentiment dengan jumlah 6518 kata yang masing masing terdapat kata sentimen positif, negatif, dan netral. Berikut merupakan contoh kata yang berlabel sentimen positive.

.

Kata	Sentiment
Asisten	Positive
Akurat	Positive
Biaya	Positive

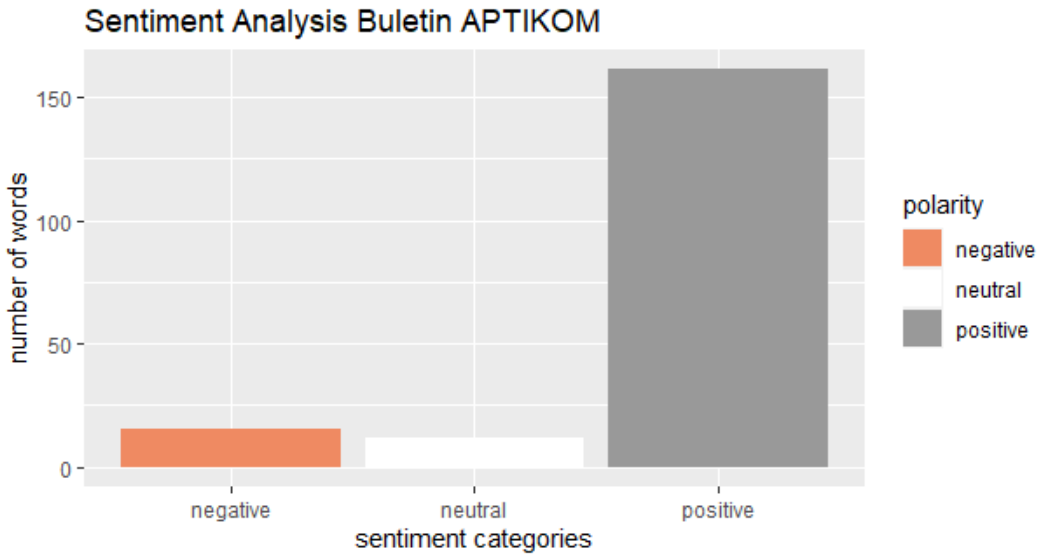
Setelah kata yang berlabel sentimen positive, selanjutnya merupakan contoh kata yang memiliki label sentimen negative.

Kata	Sentiment
Manipulation	Negative
Liar	Negative
Pola	Negative

Setelah memiliki kamu sentimen, proses selanjutnya yaitu memproses dokumen yang sudah diolah tadi untuk dicari sentimennya. Hasil pada Analisis Sentimen adalah berbentuk teks dan dalam bentuk kalimat. Berikut merupakan contoh hasil analisis sentimen.

Text	Sentimen
sejarah a kitar awal a beginning of a	Positive
beda aimachine learning deep learning	Neutral
evaluasi rule dalam isi hubung variabelvariabel masuk keluar proses hasil nya bentuk fuzzy	Negative

Hasil sentimen diatas dapat divisualisasikan dalam bentuk grafik. Data yang akan ditampilkan juga merupakan data secara keseluruhan yang sudah di proses. Berikut hasil visualisasi pada analisis sentimen.



Gambar 3 Visualisasi *Sentimen Analysis*

Hasil pada analisis sentimen dapat dilihat pada gambar grafik diatas ini bahwa pada dokumen buletin Aptikom yang berjudul *Artificial Intellegence* mendapatkan sebanyak 161 kalimat sentimen positive, 15 kalimat sentimen negative, dan 12 kalimat yang memiliki sentimen neutral.

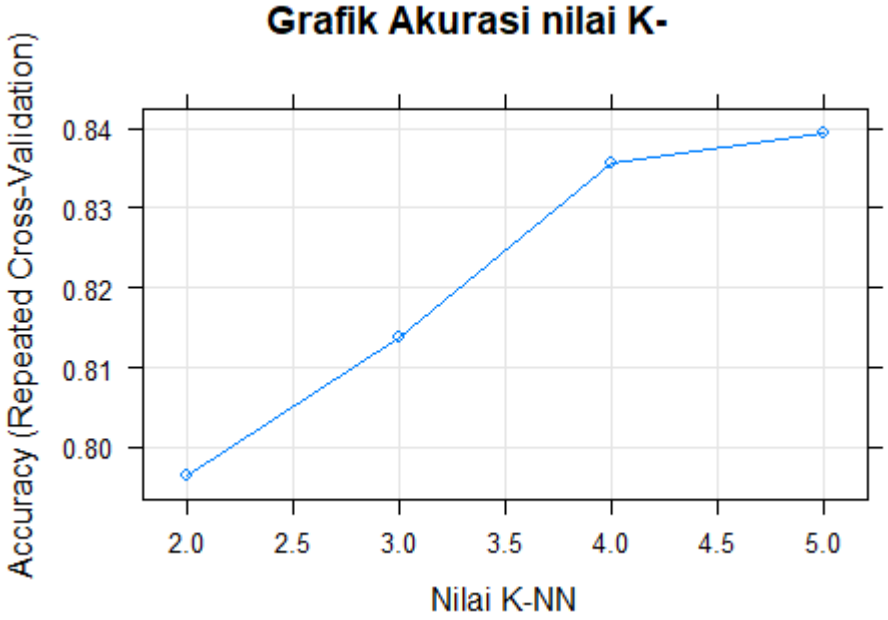
2. Algoritma K-Nearest Neighbors

Pada tahapan ini yaitu sudah memasuki tahapan klasifikasi dengan Algoritma *K-Nearest Neighbors* atau K-NN. Tentunya data yang digunakan merupakan data yang sudah di *filtering* dengan *Text Preprocessing* dan sudah mendapatkan hasil *sentiment analysis*. Berikut hasil pengklasifikasian dengan algoritma *K-Nearest Neighbors*.

```
k-Nearest Neighbors

159 samples
  2 predictor
  3 classes: 'negative', 'neutral', 'positive'

Pre-processing: centered (2), scaled (2)
Resampling: Cross-Validated (15 fold, repeated 0 times)
Summary of sample sizes: 148, 148, 148, 148, 149, 148, ...
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  2  0.7964646  0.18158310
  3  0.8138047  0.06860254
  4  0.8356902  0.11514418
  5  0.8393939  0.03036759

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

Gambar 4 Hasil Algoritma K-NN

Berdasarkan pada gambar diatas yang menunjukkan hasil pengklasifikasian dengan algoritma K-NN terlihat terdapat 159 data, 2 kolom dan terdapat 3 kelas yaitu *positive*, *negative*, *neutral*. Dapat diketahui juga nilai dari *Cross Validated* atau K-Fold yaitu 15. Untuk hasil pada data K2 sampe K5 berbeda beda mulai dari 0.79 sampai 0.83 dan mendapatkan nilai K yang optimal yaitu K=5.
Supaya dapat dilihat dengan mudah, berikut merupakan grafik dari hasil pengklasifikasian dengan algoritma K-NN diatas.



Gambar 5 Visualisai grafik algoritma K-NN

D. Evaluasi

Pengujian pada penelitian ini menggunakan metode *Confusion Matrix* untuk mengevaluasi kinerja pada algoritma yang digunakan. *Confusion Matrix* cukup sering digunakan untuk melakukan pengujian pada algoritma K-NN. Hasil evaluasi pada penelitian ini menghasilkan nilai akurasi sebesar 86.2%. Hasil *Confusion Matrix* dapat dilihat pada gambar dibawah ini.

```
Confusion Matrix and Statistics

          Reference
Prediction negative neutral positive
  negative        0       0        0
  neutral         0       0        1
  positive        1       2       25

Overall Statistics

               Accuracy : 0.8621
                 95% CI : (0.6834, 0.9611)
    No Information Rate : 0.8966
    P-Value [Acc > NIR] : 0.8249

                  Kappa : -0.045

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: negative Class: neutral Class: positive
Sensitivity                  0.00000        0.00000          0.9615
Specificity                  1.00000        0.96296          0.0000
Pos Pred Value                   NaN        0.00000          0.8929
Neg Pred Value               0.96552        0.92857          0.0000
Prevalence                   0.03448        0.06897          0.8966
Detection Rate               0.00000        0.00000          0.8621
Detection Prevalence         0.00000        0.03448          0.9655
Balanced Accuracy            0.50000        0.48148          0.4808
```

Gambar 6 *Confusion Matrix*

Untuk memastikan hasil yang ada pada sistem seperti diatas, nilai *Accuracy*, *Precision*, dan *Recall* dapat dihitung dengan menggunakan rumus seperti dibawah ini.

- *Accuracy* berguna untukm mengetahui seberapa akurat model dalam mengklasifikasikan dengan benar. *Accuracy* dapat dihitung dengan

$$Accuracy = (TP)/(Jumlah\ Data)$$

$$= TP = (0 + 0 + 25)/(29) \quad \textbf{(4)}$$

$$= 25/29 = 0.862$$

$$= 0.862 \times 100 = 86.2\%$$

- *Precision* berguna untuk menggambarkan akurasi diantara data yang diminta dengan hasil prediksi yang diberikan oleh model.

$$Precision = (TP)/(TP + FP) \quad \textbf{(5)}$$

$$= 0 + 0 + 8.33/\ 3 = \ 2{,}77$$

- *Recall* atau *Sensitivity* memiliki fungsi untuk menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi.

$$Recall\ = TP/(TP + FN) \quad \textbf{(6)}$$

$$= 0 + 0 + 25/3 = 8.33$$

Keterangan pada rumus diatas yaitu:

1. True Positive (TP) merupakan sebuah jumlah record pada data positif yang diklasifikasikan sebagai nilai positif.
2. False Positive (FN) merupakan sebuah jumlah record pada data yang bergolong negatif yang diklasifikasikan sebagai nilai positif

3. False Negative (FN) ialah sebuah jumlah record data positif yang tergolong hasil klasifikasi sebagai nilai positif
4. True Negative (TN) merupakan jumlah record data yang bergolong negatif yang memiliki hasil klasifikasi sebagai nilai negatif.

IV. KESIMPULAN DAN SARAN

Berdasarkan hasil tahapan penelitian yang telah dilakukan dapat dihasilkan dan melakukan pengujian dengan *confusion matrix,* maka dapat disimpulkan bahwa pengujian dengan *confsuion matrix* mendapatkan nilai *accuracy* sebesar 86.2% . Data yang digunakan sampai tahapan pengujian tentunya merupakan dataset yang telah diolah melalui *text preprocessing* dan siap untuk di analisis sentimennya. Hasil pada tahapan *text preprocessing* yaitu menghasilkan dataset yang sudah di filter dan menampilkan kata yang sering muncul pada buletin Aptikom, yaitu terdapat kata "dalam" berjumlah 41 kata, "ambil" 30 kata, dan "data" 27 kata. Peneltian ini juga menghasilkan dominan sentimen positif sebanyak 161 kalimat, 15 kalimat mengandung sentimen negatif, dan 12 sentimen netral. Penelitian ini menggunakan algoritma *K-Nearest Neighbors* untuk proses pengklasifikasian.

Saran untuk penelitian selanjutnya yaitu diharapkan menggunakan algoritma yang berbeda untuk melakukan perbandingan sehingga mendapatkan hasil yang lebih akurat dan juga dengan dataset yang lebih banyak. Untuk proses analisis sentimen dan *stopword* pada tahapan *text preprocessing* diharapkan menggunakan kamus yang terus *up to date* agar kosa kata yang digunakan semakin luas sehingga data yang diproses dapat lebih akurat.

PENGAKUAN

Naskah ilmiah ini merupakan sebagian penelitian pada Tugas Akhir milik Yogi Firman Alfiansah dengan judul Analisis Sentimen menggunakan Algoritma *K-Nearest Neighbors* pada buletin Aptikom, yang dibimbing langsung oleh Bapak Amril Mutori Siregar dan Ibu Anis Fitri Nur Masruriyah

DAFTAR PUSTAKA

[1] D. A. Francis, N. Carauna, J. L. Hudson and G. M. McArthur, "The association between poor reading and internalising problems: A systematic review and meta-analysis," *Clinical Psychology Review,* pp. 45-60, 2019.

[2] R. Kurniawan and A. Apriliani, "Analisis sentimen masyarakat terhadap virus corona berdasarkan opini dari Twitter berbasis Web Scrapper," *Jurnal Instek,* vol. 5, pp. 67-75, 2020.

[3] M. P. Bach, T. Bertoncel, M. Meško and Ž. Krstić, "Text mining of industry 4.0 job advertisements," *International Journal of Information Management,* vol. 50, pp. 416-431, 2020.

[4] E. Indrayuni, "Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes," *JURNAL KHATULISTIWA INFORMATIKA,* vol. VII, pp. 29-36, 2019.

[5] A. Heryanto and R. Pramudita, "Opini Media Sosial Facebook Terhadap Produk Hijab Menggunakan Metode Text Mining," *INFORMATION SYSTEM FOR EDUCATORS AND PROFESSIONALS,* vol. 4, pp. 168 - 177, 2020.

[6] F. Galati and B. Bigliardi, "Industry 4.0: Emerging themes and future research avenues using a text mining approach," *Computers in Industry,* vol. 109, pp. 100-113, 2019.