Jurnal TICOM: *Technology of Information and Communication*
Volume 10, Nomor 3, Mei 2022 Hal: 187-196 E-ISSN: 2962-7982

Analisis Sentimen Masyarakat terhadap Tim Nasional Indonesia pada Piala AFF 2020 Menggunakan Algoritma *K-Nearest Neighbors*

Aditya Eka Pratama[1], Atik Ariesta[2], Grace Gata[3*]

[1,2,3]Fakultas Teknologi Informasi, Sistem Informasi, Universitas Budi Luhur, Jakarta, Indonesia
Jl. Raya Ciledug, Petukangan Utara, Kebayoran Lama, Jakarta Selatan 12260
E-mail: [1]1712500360@student.budiluhur.ac.id, [2]atik.ariesta@budiluhur.ac.id, [3*]grace.gata@budiluhur.ac.id
(*: corresponding author)

***Abstrak*— Media sosial *twitter* adalah salah satu media komunikasi yang diminati oleh masyarakat di dunia. Hal ini dapat terjadi karena *twitter* memiliki jumlah pengguna aktif sebesar 313 juta per bulan pada tahun 2016 dan sebagian besar pengguna mengakses *twitter* melakui perangkat *mobile* yaitu sebesar 82%. Karena jumlahnya banyak, maka menimbulkan *tweet* yang banyak. *Tweet* tersebut berisi tentang kabar terbaru atau komentar yang sedang menjadi topik didunia. Hal hal yang menjadi topik atau banyak dikomentari pengguna akan menimbulkan *trending* di *twitter.* Pada piala AFF 2020 ini, Indonesia memanggil 30 pemain dengan rata-rata usianya 23,8 tahun lebih muda dari skuad tim lain yang berlaga di ajang AFF 2020. Tentu banyak risiko menurunkan skuad yang didominasi pemain muda. Mulai mental yang belum teruji, emosi gampang lepas, sampai ketergesa-gesaan. Tapi di sisi lain, deretan anak muda itu juga menumbuhkan harapan, semangat, staminia menggelora, dan bekal persiapan untuk menuju ke turnament event yang akan datang. Peneliti menggunakan metode CRISP-DM (*Cross- Industry Standard Process For Data Mining)* dan mengimplementasikan algoritma *K-Nearest Neighbors(*K-NN) untuk klasifikasi data. Data yang digunakan dalam penelitian ini berjumlah 22.182 data yang diambil dari tanggal 30 November 2021 – 06 Januari 2022. Setelah data diproses dengan metode CRISP-DM dan implementasi algoritma *K-Nearest Neighbors(*K-NN), didapatkan akurasi sebesar 67.49%, *precission* sebesar 78.99% dan *recall* 47.69%.**

***Kata Kunci*— Analisis Sentimen, *K-Nearest Neighbors(*K-NN), *Text Mining,* Timnas Indonesia, AFF 2020**

Abstrak*— *Twitter is one of the most popular communication media in the world. Twitter has 313 million active users per month in 2016 and 82% of user's access Twitter via mobile devices. The tweet contains the latest news or comments that are currently the topic of the world called trending topic. In AFF 2020 CUP, Indonesia called 30 players with an average age of 23.8 years. It was the lowest than the other squad teams that competed in the AFF 2020 event. Of course, there are many risks of called squad dominated by young players. Starting from an untested mentality, easily released emotions, to haste. But on the other hand, the row of young people also fostered hope, enthusiasm, tempestuous stamina, and preparation for the upcoming tournament events. The researcher uses the Cross-Industry Standard Process for Data Mining (CRISP-DM) method and implements the K-Nearest Neighbors (K-NN) algorithm for data classification. The data used in this study amounted to 22,182 data taken from November 30, 2021 - January 6, 2022. After the data was processed using the CRISP-DM method and the implementation of the K-Nearest Neighbors (K-NN) algorithm, the accuracy was 67.49%, precision was 78.99% and 47.69% recalls.

Keywords*— *sentiment analysis, K-Nearest Neighbors(K-NN), Text Mining, Indonesian National Team, AFF 2020

I. PENDAHULUAN

Media sosial *Twitter* adalah salah satu media komunikasi yang diminati oleh masyarakat di dunia. Hal ini dapat dilihat dari peningkatan pengguna *twitter* yang tercatat di seluruh dunia. *Twitter* memiliki jumlah pengguna aktif sebesar 313 juta per bulan pada tahun 2016 dan sebagian besar pengguna mengakses *Twitter* melalui perangkat *mobile* yaitu sebesar 82 persen. Jumlah pengguna yang banyak juga menimbulkan *tweet* yang banyak dari pengguna. Pengguna akan memberikan kabar terbaru atau komentar tentang hal yang sedang menjadi topik utama di dunia. Hal yang sedang menjadi topik utama dan banyak dikomentari oleh pengguna akan menimbulkan *trending* topik di twitter.

Pengguna Twitter yang banyak akan menimbulkan peningkatan tweet yang di posting. Setiap tweet pada Twitter memiliki topik yang berbeda. Tweet tersebut dapat memuat opini dan komentar yang berkaitan dengan bidang ekonomi, sosial, hiburan, pendidikan, olahraga, dan lain- lain. Salah satu olahraga yang digemari di Indonesia adalah sepakbola khususnya Timnas Indonesia. Pengguna Twitter akan memberikan komentar dan opini tentang performa Timnas Indonesia melalui media sosial, salah satunya Twitter. Dalam opini dan komentar *supporter* terdapat penyampaian komentar yang beragam apabila Timnas mendapatkan hasil yang baik maka akan ada opini atau komentar bahagia maupun pujian dan jika Timnas mendapatkan hasil yang buruk maka akan ada opini dan komentar kritikan bahkan cacian. Tweet dari pengguna juga dapat menjadi evaluasi bagi pengelolaan Timnas Indonesia kedepannya agar performa sesuai yang diharapkan oleh supporter. Karena dukungan dari supporter

sangat berpengaruh kepada Timnas mulai dari dukungan semangat, moril maupun pemasukan Timnas itu sendiri. Namun, pengguna akan mengalami kesulitan apabila melihat tweet secara langsung tanpa ada label tweet tersebut bernilai positif atau negatif. Sehingga diperlukan klasifikasi untuk memberikan kemudahan pada pengguna [1].

Pada turnament piala AFF 2020 ini, Indonesia memanggil 30 pemain dengan rata-rata usinya 23,8 tahun lebih muda dari skuad tim lain yang berlaga di ajang AFF 2020 [2]. Tentu banyak risiko menurunkan skuad yang didominasi pemain muda. Mulai mental yang belum teruji, emosi gampang lepas, sampai ketergesa-gesaan. Tapi di sisi lain, deretan anak muda itu juga menumbuhkan harapan, semangat, staminia menggelora, dan bekal persiapan untuk menuju ke turnament event yang akan datang.

Penelitian yang telah dilakukan oleh Mungki Astiningrum, mamluatul Hani'ah, dan Yanuar Rahmat Yoga Pradana Nugroho dengan judul "Analisis Sentimen Tentang Opini Terhadap Performa Timnas Sepak Bola Indonesia Pada *Twitter".* Data yang dikumpulkan pada penelitian ini merupakan data tweet, yang diambil dari media sosial *twitter* dengan menggunakan *keywords* #TimnasDay, #TimnasJuara, #TimnasIndonesia, #InfoTimnas dan #KitaGaruda data yang berhasil dikumpulkan pada penelitian ini sebanyak 530 *tweet* yang diambil pada tanggal 17 Februari 2020. Hasil dari penelitian menunjukan sentimen positif sebanyak 185 dan sentimen negatif sebanyak 239. Algoritma *Naive Bayes* dapat digunakan untuk mengklasifikasikan tweet kedalam positif atau negatif terutama tweet mengenai timnas sepak bola Indonesia. Dari tiga pengujian didapatkan hasil nilai algoritma Naive Bayes pada komposisi data training 371 dan data testing 159 sebesar 78%, komposisi data training 424 dan data testing 106 sebesar 84% dan komposisi data training 477 dan data testing 53 sebesar 87%. Nilai akurasi terendah adalah 78% dan tertinggi adalah 87% [1].

Dalam penelitian ini penulis bertujuan untuk menemukan nilai sentimen positif dan negatif dari opini masyarakat terhadap performa timnas Indonesia di piala AFF 2020, banyak masyarakat yang beropini di media sosial *twitter* jika timnas Indonesia menunjukkan performa permainan yang kurang baik di piala AFF 2020 dan juga ada masyarakat yang merasa puas dan bangga terhdap performa timnas Indonesia. Dengan diadakannya penelitian ini dapat menemukan nilai sentimen positif dan negatif dari opini masyarakat, dan nantinya penelitian ini diharapkan bisa membuat timnas Indonesia membenahi diri dalam menunjukkan performa yang baik pada masa yang akan datang. Berdasarkan latar belakang yang telah dikemukakan sebelumnya, maka penulis mengidentifikasikan masalah, yaitu: Bagaimana cara mendapatkan sentimen positif dan negatif dari *tweet* yang relevan dengan topik "analisis sentimen masyarakat terhadap timnas Indonesia pada piala AFF 2020 menggunakan algoritma *k-nearest neighbors*. Berdasarkan opini dari *tweet* yang didapatkan, bagaimana tingkat kepuasan masyarakat terhadap timnas Indonesia dalam menunjukkan performa permainan yang baik.

Tujuan dari penulisan penelitian ini adalah untuk melakukan analisis terhadap sentimen positif dan negatif dari opini masyarakat indonesia yang dituangkan melalu sosial media *twitter.* Manfaat dari penulisan ini adalah untuk menemukan nilai sentimen positif dan negatif opini masyarakat terhadap timnas Indonesia di piala AFF 2020 dalam menunjukkan performa permainan yang nantinya akan berguna untuk penelitian lanjutan dan sebagai kritik terhadap timnas indonesia untuk membenahi diri dalam menunjukkan performa yang lebih baik lagi pada masa yang akan datang.

II. METODE PENELITIAN

A. *Twitter*

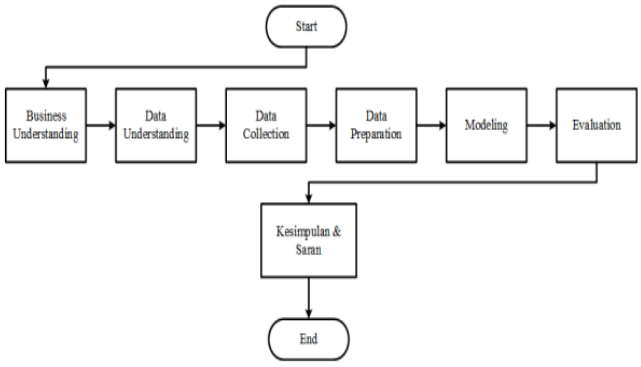
Twitter adalah situs micro blogging yang dioperasikan oleh *Twitter, Inc.* Disebut micro blogging karena situs ini memungkinkan penggunanya mengirim dan membaca pesan seperti blog pada umumnya. Pesan tersebut dinamakan *tweet*, yaitu teks tulisan sebanyak 140 karakter yang ditampilkan pada halaman profil pengguna. *Twitter* juga dapat memberikan dampak yang negatif bagi penggunanya. *Twitter* sebagai salah satu dari sekian banyak ragam media sosial biasanya dimanfaatkan oleh para remaja untuk berinteraksi dengan individu-individu lainnya di dunia maya, seperti teman-teman, kerabat, keluarga, kenalan baru dan lain sebagainya [3].

B. *Data Mining*

Menurut [4] Fina Nasari dalam jurnal *Data Mining* merupakan proses menemukan korelasi baru yang bermanfaat, pola dan trend dengan menambang sejumlah repository data dalam jumlah besar, menggunakan teknologi pengenalan pola seperti statistik dan teknik matematika. *Data mining* merupakan proses analisis data menggunakan perangkat lunak untuk menemukan pola dan aturan (rules) dalam himpunan data [5]. *Data Mining* merupakan suatu metode untuk menemukan pengetahuan dalam suatu tumpukan data yang cukup besar [6].

C. Tahapan Penelitian

Seperti yang terlihat pada gambar 1 menunjukkan tahapan yang digunakan dalam penelitian ini, serta metodologi *CRISP-DM*. Tahapan ini terdiri dari beberapa tahap yang akan dijelaskan sebagai berikut:



Gambar 1. Kerangka Penelitian

Pada tahapan ini terdapat sub-tahapan dengan menerapkan metodologi *CRISP-DM.* Model proses *CRISP-DM* memiliki beberapa fase diantaranya adalah sebagai berikut :

1) Pemahaman Bisnis *(Business Understanding) :* Di dalam tahapan business understanding peneliti mencoba untuk memahami permasalahan yang ingin diangkat yaitu analisis sentimen masyarakat terhadap timnas Indonesia pada AFF 2020 , pertama peneliti akan mencoba mempelajari artikel atau berita yang ada di internet tentang informasi performa timnas, kemudian mencoba melihat trending topik yang berkaitan dengan timnas Indonesia pada media sosial twitter.

2) Pemahaman Data *(Data Understanding):* Di dalam tahapan data understanding yang dilakukan peneliti adalah mencoba untuk menentukan tweets yang relevan dengan topik yang ingin di angkat dengan keywords kitagaruda, PSSI, timnas Indonesia. Kemudian keywords tersebut akan dilakukan pencarian pada media twitter untuk menampilkan tweet sesuai dengan topik permasalahan jika penentuan keywords sudah dilakukan maka selanjutnya peneliti akan menentukan lama periode data tweet yang akan digunakan.

3) Koleksi Data *(Data Collection) :* Di dalam tahapan data collection peneliti akan melakukan pengambilan data tweets dari 30 November 2021 hingga 06 Januari 2022, dengan target data tweet yang dikumpulkan sebanyak 22.182 tweet. Sebelum proses pengambilan tweet dilakukan, peneliti diwajibkan untuk membuat koneksi yang menghubungkan antara rapid miner dengan twitter agar proses pengumpulan data pada penelitian ini dilakukan dengan menggunakan software rapid miner. Data yang diambil hanya data yang sesuai dengan keywords yang sebelumnya telah ditentukan.

4) Persiapan Data *(Data Preparation):* Pada tahapan ini data yang sudah didapatkan melalui tahapan data collection akan ditentukan sentimennya oleh peneliti sendiri menggunakan kamus lexicon. Setelah pelabelan sentimen dilakukan maka data tweet yang sudah memiliki label sentimen akan di lakukan data preprocessing.Tahapan data preprocessing yang dilakukan adalah sebagai berikut:

1. *Remove Duplicate* : Pada tahap ini dilakukan proses menghilangkan duplikasi data yang ada pada data hasil crawling dengan menggunakan *RapidMiner*.
2. *Case Folding* : Pada tahap ini dilakukan proses mengubah semua huruf pada data *tweets* yang sudah didapat menjadi bentuk *lower case* atau huruf kecil.
3. *Cleansing / Filtering* : Pada tahap ini dilakukan proses menghapus karakter atau tanda baca yang tidak diperlukan dalam proses analisis sentimen. Contoh seperti URL, *hastag* (#), *username*, dan *mention*.
4. *Tokenizing* : Pada tahap ini dilakukan proses memecah data tweets yang masih dalam bentuk kalimat, menjadi kata individual
5. *Stopwords* : Pada tahap ini dilakukan proses *stopwords* untuk menghilangkan kata sambungan / *stopwords* yang ada pada data.
6. *Stemming* : Pada tahap ini dilakukan proses *stemming* berfungsi untuk mengembalikan kata menjadi bentuk kata dasar dari sebuah kata.

Data *tweets* yang sudah melalui proses data *preprocessing* nantinya akan berubah bentuknya dari kalimat akan menjadi token / kata. Lalu data akan dipecah menjadi *data training* dan *data testing* yang akan digunakan didalam tahapan *modeling.*

5) Pemodelan *(Modeling):* Di dalam tahapan Modeling peneliti akan melakukan pemodelan terhadap dataset yang sudah dilakukan preprocessing. Kemudian dataset tersebut akan dibagi kedalam data training dan data testing dimana pada penelitian ini menggunakan split data dengan perbandingan data training dan data testing 60:40, 70:30, 80:20 dan menggunakan cross validation dengan pembagian secara acak kedalam 10 bagian (number of folds =10). Pada fase ini peneliti akan melakukan pemodelan dengan menggunakan algoritma k-nearest neighbors untuk mendapatkan nilai sentimen positif atau negatif dari sebuah tweet namun sebelumnya telah dilakukan perbandingan model klasifikasi dengan menggunakan algoritma decision tree, naïve bayes.

6) Evaluasi *(Evaluation):* Pada tahapan ini peneliti akan melakukan evaluasi metode klasifikasi dengan mengukur performa menggunakan confusion matrix terhadap algoritma k-nearest neighbors.

III. HASIL DAN PEMBAHASAN

A. *Data Collection*

Pada tahapan *data* collection peneliti mengumpulkan data yang akan digunakan. Data yang didapatkan cukup banyak yaitu 22.182 data. Peneliti melakukan pengambilan data *tweet* yang berkaitan dengan timnas Indonesia pada tanggal 30 November 2021 hingga 06 Januari 2022 dengan *keywords* yang sedang *trending topic* seperti #kitagaruda, PSSI, dan Timnas Indonesia. Pada proses penelitian ini peneliti mengumpulkan data dari *twitter* dengan menggunakan *tools rapidminer*. Peneliti mengumpulkan data pada tanggal 30 November 2021 – 6 Januari 2022 dengan *keyword* yang sedang *trending topic*. *Keywords* yang digunakan yaitu Kitagaruda, PSSI, dan Timnas Indonesia.

Peneliti menggunakan *hastag* etersebut karena masih kaitannya dengan timnas Indonesia. Berikut ini adalah data *tweet* yang berhasil peneliti kumpulkan dari media sosial *twitter* dengan menggunakan *rapid miner* seperti yang ditunjukan pada (Tabel 1)

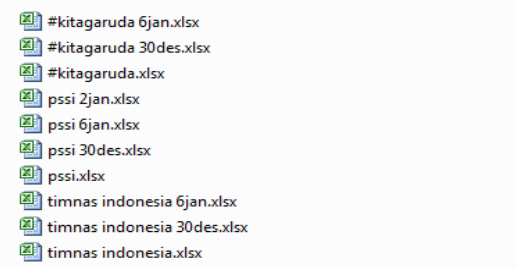
TABEL I
JUMLAH DATA TWEET HASIL CRAWLING

Keywords	**Tanggal**	**Jumlah**
Kitagaruda	30 November 2021 – 6 Januari 2022	556
PSSI	28 Desember 2021 – 6 Januari 2022	8.641
Timnas Indonesia	01 Desember 2021 – 6 Januari 2022	12.985
Jumlah		22.182

B. Hasil *Crawling Data*

Dalam penelitian ini, peneliti menggunakan *attribute From-User* dan *Text* untuk isi dari data *tweet* yang diambil

dalam proses *crawling* data yang dilakukan pada tanggal 30 November 2021 – 6 Januari 2022 dengan jumlah data 22182 data. Isi dari kolom *Created-At* yaitu tanggal dimana *tweet* tersebut dibuat atau di postingoleh pengguna *twitter*. Isi dari kolom *From-User* adalah nama pengguna *twitter*. Dan isi dari kolom *Text* yaitu *tweet* yang diungkapkan oleh si pengguna *twitter*. Berikut ini adalah contoh sampel data hasil *crawling* data dengan *rapidminer* (Tabel 2)



Gambar 2. Data *Excel* Hasil *Crawling*

TABEL 2
SAMPLE ISI DATA EXCEL

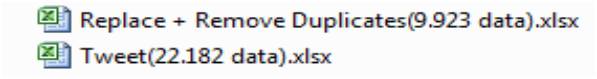
Cread-At	*From-User*	*Text*
04/01/2022 05:23:55 PM	DJ Kasino	Atas permintaan Shin Tae-yong, PSSI berupaya agar Timnas Indonesia gelar pertandingan persahabatan pada akhir Januari untuk memperbaiki peringkat FIFA. #storybola #bolaindonesia #timnas #liga1 #beritabola #quotesanakbola #quotessepakbola #katasepakbola #ligaindonesia #anakbola https://t.co/Z8ifzYVCch
04/01/2022 11:50:14 AM	Delhen	@PSSI pelatih terbaik sepanjang sejara sepak bola indonesia!
30/12/2021 05:57:06 AM	Sjahriel Purwan	Apapun hasilnya, kalian sudah berjuang dan menampilkan yang terbaik. Love you all... #KitaGaruda #TimnasDay
01/01/2022 09:33:11 PM	Galang	memang jelas timnas indonesia kekurangan striker, dulu masih jamannya Bepe bagus secara dia kapten tapi minim goal. Setelahnya kayanya ga ada lagi deh, mau ngomongin gonzales ga juga, beto gonzalves, osvaldo hay apa lagi
01/01/2022 09:29:17 PM	Rizky Adnan	Wasit dari timur tengah ada masalah apa sih sama timnas indonesia? Keputusannya kontroversial mulu

C. *Data Preparation*

Pada tahapan *data preparation* ini, peneliti akan membagi menjadi 3 tahapan sebagai berikut :

1. Pemberian label sentimen, tahapan ini merupakan proses pemberian label sentimen positif atau negatif dari sebuah *tweet* yang dilakukan dengan meggunakan kamus *lexicon* milik [7] dan *tool rstudio.*
2. Tahapan *preprocessing,* tahapan ini merupakan proses yang akan dilakukan oleh peneliti setelah data *tweet* sudah memiliki sentimen positif atau negatif.
3. Menentukan data *training* dan data *testing*, tahapan ini merupakan proses untuk menentukan perbandingan data *training* dan data *testing* terlebih dahulu sebelum masuk ke dalam tahapan *modelling.*

1) *Penentuan Label Sentimen*: Tweet yang sudah didapatkan selanjutnya dilakukan Replace untuk menghilangkan link,hastag, dan mention yang terdapat pada tweet. Pada proses ini terdapat 6680 #hastag yang dihilangkan, 8256 link yang dihilangkan dan 10553 @mention yang dihilangkan dan dilakukan remove duplicate terlebih dahulu agar dataset pada file excel terhindar dari duplikasi data.



Gambar 3. Hasil Proses *Remove Duplicate*

Setelah selesai menjalankan proses *remove duplicate* dengan menggunakan *rapidminer*. Dapat dilihat perbedaan jumlah data seperti yang ditunjukan pada gambar 3. awalnya 22182 data *tweet* berkurang menjadi 9923 data *tweet* karena banyaknya data yang bersifat *retweet*. Kemudian langkah selanjutnya adalah melakukan klasifikasi atau penentuan sentimen berdasarkan masing-masing *tweet*. Karena keterbatasan waktu dan juga belum mendapatkan seorang pakar untuk memvalidasi sentimen dari sebuah *tweet*, maka dari itu selama proses penentuan sentimen berlangsung, peneliti menentukanya sendiri menggunakan kamus *lexicon* milik [7] dan *tools rstudio*. Penelitian ini menggunakan kamus *lexicon* karena mudah digunakan dan menggunakan kamus sebagai sumber bahasa atau leksikal. Pada saat proses pemberian label sentimen terdapat lima *attribute* pada file *excel* yaitu:

1. Kolom *Text* adalah *tweet* yang diungkapkan oleh pengguna *twitter* tersebut.
2. Kolom *Positif* adalah total kata *positif* yang terdapat pada *tweet* tersebut.
3. Kolom *Negatif* adalah total kata *negatif* yang terdapat pada *tweet* tersebut.
4. Kolom *Score* adalah hasil dari pengurangan antara total kata pada kolom *positif* dikurangi total kata pada kolom *negatif.*
5. Kolom *Label* berisikan identifikasi sentimen menurut [8] jika nilai akhir yang dihasilkan dari perhitungan tersebut menghasilkan skor lebih besar atau sama dengan 1, maka skor tersebut diidentifikasikan sentimen *positif.* Sedangkan nilai akhir yang dihasilkan dari perhitungan

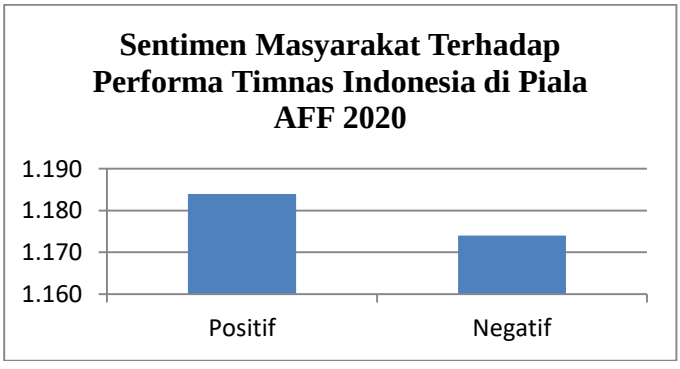
yang menghasilkan skor < 0, maka skor tersebut diidentifikasikan sentimen *negatif*.

Berikut ini adalah contoh hasil pemberian sentimen menggunakan kamus *lexicon* (Tabel 3)

TABEL 3
CONTOH TWEET DENGAN SENTIMEN

Text	Positif	Negatif	Score	Label
Thailand bagus, ditambah pelatihnya yang melakukan pergantian pemain yang sangat efisien, namun permainan timnas juga lebih baik dibanding leg 1	10	9	1	Positif
Cinta sama timnas Indonesia itu rumit. Kecewa ada tapi senengnya lebih banyak.	5	7	-2	Negatif

Setelah melalui tahapan pemberian label sentimen, maka jumlah data *tweet* berkurang menjadi 4358, hal tersebut dikarenakan banyaknya data yang memiliki sentimen netral sehingga dari 4358 data *tweet* terdapat 2184 *tweet* sentimen *positif* dan 2174 *tweet* sentimen *negatif*. Gambar 4. merupakan kumpulan sentimen masyarakat yang ada didalam *dataset* opini dari performa timnas Indonesia.



Gambar 4. Visualisasi Data Setelah Proses Pemberian Label Sentimen

2) *Tahapan Pre-processing Data*: Membersihkan data dilakukan dengan menggunakan tahapan preprocessing yang terdiri dari transform case, tokenize, filter tokens, filter stopword, dan steam. Dimana dalam penggunakan filter stopword peneliti menggunakan kamus stopwords milik [9] dan penggunaan steam peneliti menggunakan kamus steam milik [10]. Simpan kata-kata yang ada dalam dataset dalam bentuk file notepad atau .txt yang kemudian dimasukan kedalam rapid miner yang selanjutnya membuat kalimat menjadi kata dasar. Berikut ini adalah contoh hasil dari tahapan *preprocessing* (Tabel 4)

TABEL 4
TABEL PERBANDINGAN SEBELUM DAN SESUDAH PREPROCESSING

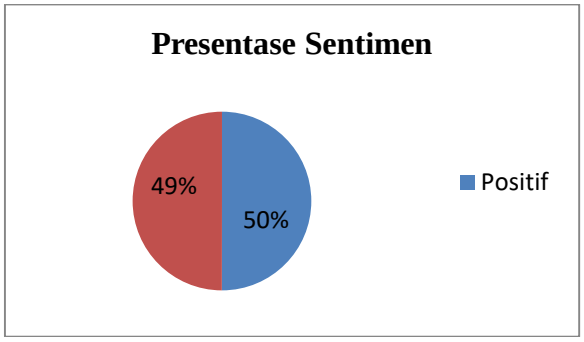
Preprocessing	Sebelum	Sesudah
Transform Case	Terima kasih sudah berjuang sampai detik terakhir, kalian luar biasa Timnas Indonesia. Kami Bangga	terima kasih sudah berjuang sampai detik terakhir, kalian luar biasa Timnas Indonesia. kami bangga
Tokenize	terima kasih sudah berjuang sampai detik terakhir, kalian luar biasa Timnas Indonesia. kami bangga	“terima”,”kasih”,”sudah” ,”berjuang”, ”sampai”,”detik”,”terakhir”,”kalian”, ”luar”,”biasa”,”timnas”,” Indonesia”, ”kami”,”bangga”
Filter Tokens	“terima”,”kasih”,”sudah”, ”berjuang”,”sampai”,”detik”, ”terakhir”,”kalian”,”luar”,”biasa”,”timnas”,”Indonesia”,”kami”, ”bangga”	“terima”,”kasih”,”sudah”, ”berjuang”,”sampai”,”detik”, ”terakhir”,”kalian”,”luar”,”biasa”, ”timnas”,”Indonesia”,”kami”, ”bangga”
Filter Stopwords	“terima”,”kasih”,~~”sudah”~~, ”berjuang”,~~”sampai”~~,” detik”, ~~”terakhir”~~,~~”kalian”~~,~~”luar”~~,~~”biasa”~~,”timnas”, ”Indonesia”,~~”kami”~~, ”bangga”	“terima”,”kasih”,”berjuang”,”detik”, ”timnas”,”Indonesia”,”bangga”
Steam	“terima”,”kasih”,”berjuang”, ”detik”,”timnas”,”Indonesia”,”bangga”	“terima”,”kasih”,”berjuang”,”detik”, ”timnas”,”Indonesia”,”bangga”

Berikut ini adalah penjelasan operator pada tahapan *preprocessing:*

1. *Transform Cases* adalah proses mengubah semua huruf pada data *tweet* menjadi bentuk *lower cases.* Pada proses ini semua huruf dirubah kedalam huruf kecil karena mayoritas text sebagian besar merupakan huruf kecil semua.
2. *Tokenize* adalah proses memecah data *tweet* yang masih dalam bentuk kalimat agar sistem dapat melakukan pengecekan 1 per 1 terhadap tiap-tiap text yang ada pada kalimat. Pada proses ini terdapat 8380 kata yang telah dipecah.
3. Filter Tokenize adalah proses untuk memfilter *token* berdasarkan panjang karakter. Pada proses ini terdapat 803 kata yang terfilter maka kata yang dipecah menjadi 7577 kata.

4. *Filter Stopwords* adalah proses untuk menghilangkan kata sambung, setelah proses ini dijalankan terdapat 576 kata yang dihilangkan sehingga menjadi 7001 kata.
5. *Steamming* adalah proses untuk mengembalikan kata menjadi bentuk kata dasar dari sebuah kata.

Setelah melalui tahapan *data prepraration*, maka jumlah data *tweet* berkurang menjadi 4256 data *tweet*, hal tersebut dikarenakan banyaknya data yang bersifat *duplicate*. Sehingga dari 4245 data *tweet* terdapat 2129 *tweet* sentimen *positif* dan 2127 *tweet* sentimen *negatif*. Jika dipresentasekan terdapat 50% sentimen *positif* dan 49% sentimen *negatif* seperti yang ditunjukan pada (Gambar 5).



Gambar 5. Presentase Sentimen Setelah Proses *Preprocessing*

3) *Penentuan Data Training dan Data Testing*: Data *training* adalah data latih yang digunakan untuk proses latihan bagi model klasifikasi. Sedangkan data *testing* adalah data yang akan digunakan untuk uji *rule* klasifikasi. Pada penelitian ini data akan dibagi menjadi perbandingan 60:40, 70:30, dan 80:20 yang dapat dilihat pada Tabel 5. metode pembagiannya menggunakan metode *stratified sampling*. *Stratified sampling* adalah pemilihan sampel dilakukan secara acak dan terstruktur pada masing-masing kelompok.

TABEL 5
PERBANDINGAN DATA TRAINING DAN DATA TESTING

Algoritma	Perbandingan
Decision Tree, Naive Bayes, KNN	60:40
Decision Tree, Naive Bayes, KNN	70:30
Decision Tree, Naive Bayes, KNN	80:20

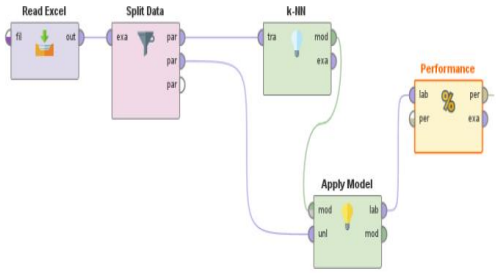
Berikut ini adalah hasil sebaran data *training* dan data *testing* perbandingan 60:40,70:30, dan 80:20 yang dapat dilihat pada (Tabel 6).

TABEL 6
SEBARAN DATA TRAINING DAN DATA TESTING

Perbandingan 60 : 40			
Sentimen	Training	Testing	Total
Positif	852	1277	2129
Negatif	851	1276	2127
Total	1703	2553	4256
Perbandingan 70:30			
Sentimen	Training	Testing	Total
Positif	639	1490	2129
Negatif	638	1489	2127
Total	1277	2979	4256
Perbandingan 80:20			
Sentimen	Training	Testing	Total
Positif	426	1703	2129
Negatif	425	1702	2127
Total	851	3405	4256

D. ***Modeling***

1) *Modelling* Menggunakan Split Data
   a) *Read excel* adalah operator untuk memilih file *excel* yang akan digunakan
   b) *Split data* adalah operator yang digunakan untuk membagi *data training* dan *data testing*.
   c) *K-NN* adalah operator yang menghasilkan model k-nn yang digunakan untuk klasifikasi
   d) *Apply model* digunakan untuk menerapkan model yang telah dilatih sebelumnya menggunakan data *training* pada data *testing*.
   e) *Performance* digunakan untuk mengevaluasi kinerja model yang memberikan daftar nilai kinerja secara otomatis, misalnya *accuracy, precision, dan recall.*

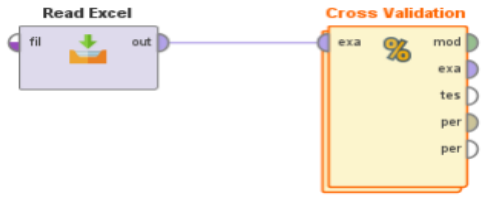


Gambar 6. Penerapan Algoritma *K-NN* dengan Model *Split Data*

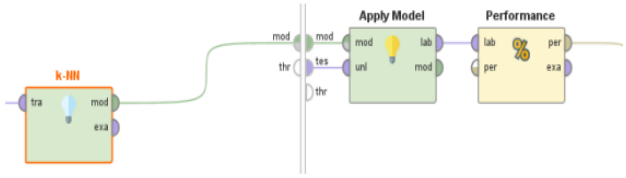
2) *Modelling* Menggunakan *Cross Validation*

Berikut ini adalah penjelasan pada penerapan algoritma - *k-nearest neighbors* menggunakan *cross validation(*Gambar 6) dan (Gambar 7):

a) *Read excel* adalah operator untuk memilih file *excel* yang akan digunakan.
b) *Cross Validation adalah operator yang digunakan untuk membagi data training dan data testing.*
c) *K-NN adalah operator yang menghasilkan model k-nn yang digunakan untuk klasifikasi*
d) *Apply Model digunakan untuk menerapkan* model yang telah dilatih sebelumnya menggunakan data *training* pada data *testing.*
e) *Performance* digunakan untuk mengevaluasi kinerja model yang memberikan daftar nilai kinerja secara otomatis, misalnya *accuracy, precision,* dan *recall*



Gambar 7. Penerapan Algoritma *K-NN* Dengan Model *Cross Validation*



Gambar 8. Proses *Cross Validation*

E. Proses Hasil Pengujian Model

1) Hasil Proses *Modelling* menggunakan *split data*

Setelah melakukan tahapan *modeliing* dengan menggunakan perbandingan data *training* dan data *testing* terhadap *dataset* yang sudah dilakukan *preprocessing*. Maka nilai *accuracy* yang dihasilkan dapat dilihat pada (Tabel 7)

TABEL 7
SEBARAN DATA TRAINING DAN DATA DATA TESTING

Perbandingan	Algoritma		
	Decision Tree	***Naïve Bayes***	***K-Nearest Neighbors***
60:40	60.99%	59.66%	**67.49%**
70:30	61.26%	59.89%	65.69%
80:20	61.53%	61.82%	64.35%

Berdasarkan Tabel 7 maka dapat diambil kesimpulan bahwa hasil *accuracy* tertinggi sebesar 67.49%, yang dihasilkan dengan pemodelan terhadap *dataset* yang sudah dilakukan *preprocessing* dengan menggunakan algoritma *k-nearest neighbor* dan menggunakan perbandingan 60:40 untuk data *training* dan data *testing*.

Pengujian pada algoritma *k-nearest neighbor* dilakukan sebanyak 5 kali dengan pembagian data 60:40. Pengujian dilakukan berdasarkan nilai k, yaitu *k*=1, *k*=3, *k*=5, *k*=7, dan *k*=9 pada Tabel 8. penelitian ini menggunakan angka K ganjil dikarenakan algoritma *K-Nearest neighbors* bekerja dengan cara menentukan kelas berdasarkan kelompok mayoritas hasil dari pemilihan tetangga terdekat sebanyak k tetangga menjadi penentu kelas dari data uji.

TABEL 8
PERBANDINGAN ACCURACY BERDASARKAN NILAI K

Nilai *K*	*Accuracy K-Nearest Neighbors*
K=1	66.35%
K=3	67.10%
K=5	**67.49%**
K=7	65.92%
K=9	65.73%

Nilai *k* = 5 menghasilkan *accuracy* lebih besar dari pada *k*=1, *k*=3, *k*=7, dan *k*=9 karena jumlah sentimen yang kemunculannya paling banyak pada *k* =5 adalah sentimen *positif* dan nilai perhitungan terbesar.

2) Hasil Proses *Modelling* Menggunakan *Cross Validation*

Setelah melakukan tahapan *modeling* terhadap *dataset* yang sudah dilakukan *preprocessing*. Maka nilai *accuracy* yang dihasilkan dapat dilihat pada Tabel 9.

TABEL 9
HASIL ACCURACY CROSS VALIDATION

Algoritma	*Accuracy*
Decision Tree	62.38%
Naïve Bayes	59.91%
K-Nearest Neighbors	**66.94%**

Berdasarkan Tabel 7 dan Tabel 9 yang berisi hasil pengukuran performa model klasifikasi *k-nearest neighbors* dengan 2 cara yaitu pembagian *dataset* dengan *split data* dan *cross validation*. Maka dapat disimpulkan bahwa pemodelan dengan *split data* 60:40 memiliki akurasi sebesar 67.49% sedangkan dengan pemodelan *cross validation* memiliki akurasi sebesar 66.94%, sehingga akurasi tertinggi yang dihasilkan pada penelitian ini adalah dengan pemodelan dengan *cross validation*.

3) Hasil Perhitungan Algoritma *K-Nearest Neighbors*

Perhitungan manual dengan algoritma *k-nearest neighbors* ini menggunakan sample dari *dataset* yang diambil secara acak sebanyak 5 data *tweet* yang dapat dilihat pada Tabel 10 yang dimana sample *tweet* D1 Sampai D4 sebagai data *training* dan D5 sebagai data *testing*.

TABEL 10
CONTOH TWEET

Sample	*Tweet*	Sentimen
D1	Terima kasih perjuangannya timnas indonesia! Kalian luar biasa!	Positif
D2	Terima kasih timnas Indonesia	Positif
D3	Cinta sama timnas indonesia itu rumit. Kecewa ada tapi senengnya lebih banyak.	Negatif
D4	Wasit dari timur tengah ada masalah apa sih sama timnas indonesia?keputusannya kontoversial mulu	Negatif
D5	Terima kasih Timnas Indonesia utk perjuangannya dan pencapaiannya	?

Setelah menentukan *sample tweet* yang ingin digunakan, maka selanjutnya adalah menentukan parameter *k*, kemudian melakukan *preprocessing* dan menentukan bobot *term* dari 5 *tweet* menggunakan *term weighting(TF-IDF.)*

Jika sudah menentukan bobot *term* dari 5 *tweet* menggunakan *term weighting (TF-IDF*, maka selanjutnya adalah menghitung perkalian skalar antara data *testing* (D5) pada Wdt dengan Data *training* (D1,D2,D3,D4) pada Wdt, lalu jumlahkan hasil perkalian skalar pada masing-masing data *training*.

TABEL 11
HASIL PERKALIAN SKALAR DATA TRAINING DAN DATA TESTING

X1	**X2**	**X3**	**X4**
0.0492	0.0492	0	0
0.0492	0.0492	0	0
0,1584	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0.2568	0.0984	0	0

Kemudian selanjutnya adalah menghitung kemiripan panjang *vector* setiap *tweet* seperti yang ditunjukan pada Tabel 12 dengan cara, kuadratkan bobot setiap dokumen pada Wdt, lalu jumlahkan nilai kuadrat tersebut dan jumlah nilai kuadrat diakarkan.

Penjelasan dari Tabel 12 adalah sebagai berikut:
Y1 = D1 pada Wdt tabel 10 dikuadratkan
Y2 = D2 pada Wdt tabel 10 dikuadratkan
Y3 = D3 pada Wdt tabel 10 dikuadratkan
Y4 = D4 pada Wdt tabel 10 dikuadratkan
Y5 = D5 pada Wdt tabel 10 dikuadratkan

Baris yang berwarna biru pada Tabel 12 adalah hasil penjumlahan dari nilai kuadrat, sedangkan untuk baris yang berwarna hijau adalah hasil setelah penjumlahan nilai kuadrat yang telah diakarkan.

TABEL 12
PANJANG VEKTOR SETIAP TWEET

Y1	Y2	Y3	Y4	Y5
0.0492	0.0492	0	0	0.0492
0.0492	0.0492	0	0	0.0492
0.1584	0	0	0	0.1584
0	0	0	0	0
0	0	0	0	0
0.4886	0	0	0	0
0.4886	0	0	0	0
0.4886	0	0	0	0
0	0	0.4886	0	0
0	0	0.4886	0	0
0	0	0.4886	0	0
0	0	0.4886	0	0
0	0	0	0.4886	0
0	0	0	0.4886	0
0	0	0	0.4886	0
0	0	0	0.4886	0
0	0	0	0.4886	0
0	0	0	0.4886	0
0	0	0	0	0.4886
0	0	0	0	0.4886
1.7226	**0.0984**	**1.9544**	**2.9316**	**1.2340**
1.312	**0.313**	**1.398**	**1.712**	**1.111**

Jika sudah menghitung kemiripan panjang *vector* dari setiap *tweet*, maka selanjutnya adalah menerapkan persamaan *cosine similarity* sebagai berikut ini:

- $Cos(D5,D1)= \frac{0.2568}{1.111 \times 1.312} = 0.1761$
- $Cos(D5,D2) = \frac{0.0984}{1.111 \times 0.313} = 0.2829$
- $Cos(D5,D3) = \frac{0}{1.111 \times 1.398} = 0$
- $Cos(D5,D4) = \frac{0}{1.111 \times 1.712} = 0$

Setelah menerapkan persamaan *cosine similarity*, maka selanjutnya adalah mengurutkan berdasarkan jarak dari nilai terbesar ke terkecil yang ditunjukan pada Tabel 13.

TABEL 13
JARAK BERDASARKAN NILAI TERBESAR KE TERKECIL

D2	**D1**	**D3**	**D4**
0.2829	0.1761	0	0

Jadi dipilih nilai *k=2* maka akan dipilih 1 nilai *similarity* yang tertinggi, kemudian ambil sebanyak nilai *k (k=2)* yang paling tinggi tingkat kemiripannya dengan D5 seperti pada tabel

TABEL 14
TWEET YANG PALING TINGGI KEMIRIPANNYA

D2	**D1**
Positive	*Positive*

Lalu pilih sentimen yang paling banyak kemunculannya, berdasarkan *k=2*, masing-masing D1 dan D2 sentimen *positive*.

Dari 2 nilai tertinggi tersebut, kelas sentimen yang paling banyak muncul adalah positif sehingga *tweet* D5 masuk ke dalam sentimen positif.

F. *Evaluation*

Penelitian ini akan melakukan pengujian dengan menggunakan *confusion matrix* yang diperoleh dari tahapan *modelling* dengan menggunakan algoritma *k-nearest neighbors* mendapatkan akurasi terbesar dengan menggunakan *split data* yang didapatkan ketika nilai *k* berada pada angka 5. Sehingga *confusion matrix* yang disajikan pada Tabel 15 adalah hasil dari evaluasi pengukuran klasifikasi algoritma *k-nearest neighbors* dengan *dataset* yang berjumlah 4256 terdapat 1703 data *training* dan 2553 data *testing*.

TABEL 15
CONFUSION MATRIX

	True Positive	***True Negative***	***Class Precision***
Pred.Positive	609	162	78.99%
Pred.Negative	668	1114	62.51%
Class Recall	47.69%	87.30%	

Perhitungan dari Tabel 15 adalah sebagai berikut :

- *Precision*= $\frac{TP}{TP+FP} = \frac{609}{609+162} = 0.7899 = 78.99\%$
- *Recall*= $\frac{TP}{TP+FN} = \frac{609}{609+668} = 0.4769 = 47.69\%$
- *Accuracy*= $\frac{TP+TN}{TP+TN+FP+FN} = \frac{609+1114}{609+1114+162+668} = 0.6749 = 67.49\%$

Berdasarkan perhitungan pada Tabel dapt dilihat bahwa hasil dari pengolahan data sentimen demgan menggunakan KNN menghasilkan *precision* sebesar 78.99%, *recall* 47.69%, dan *accuracy* sebesar 67.49%

Adapun uraian hasil proses pembentukan *confusion matrix* sebagai berikut:

1. Hasil aktual positif yang diprediksi positif oleh KNN (TP) sebanyak 609 data.
2. Hasil aktual positif yang diprediksi negatif oleh KNN (FN) sebanyak 668 data
3. Hasil aktual negatif yang diprediksi positif oleh KNN (FP) sebanyak 162 data
4. Hasil aktual negatif yang diprediksi negatif oleh KNN (TN) sebanyak 1114 data.

IV. PENUTUP

Berdasarkan hasil yang didapat setelah melakukan penelitian ini, maka dapat diambil kesimpulan sebagai berikut: Berdasarkan penerapan model dengan menggunakan algoritma *K-Nearest Neighbors(*K-NN) terhadap *dataset* yang sudah didapatkan dengan pembagian data untuk data training dan data testing sebesar 60:40, didapatkan nilai akurasi sebesar 67.49%, *precission* sebesar 78.99%, dan *recall* 47.69%. Berdasarkan hasil penerapan model dengan menggunakan algoritma *K-Nearest Neighbor(K-NN)* dengan data yang berhasil didapatkan dari proses *crawling data* dari *twitter* berjumlah 22.182 data yang didapatkan dari masyarakat Indonesia pengguna *twitter* yang ber opini terhadap performa timnas Indonesia. Didapatkan total data hasil pengujian sebesar 4256 data dan terbagi menjadi dua jenis sentimen yaitu sentimen *positive* dan *negative*, dengan pembagian data *positive* sebanyak 2129 dengan *presentase* 50% dan data *negative* sebanyak 2127 dengan *presentase* 49%. Dari hasil ini, didapatkan jumlah sentimen *positive* lebih besar dibanding dengan sentimen *negative.* Berdasarkan pengujian menggunakan algoritma *K-Nearest Neighbor (*K-NN) yang menunjukkan tanggapan *positive* dari masyarakat Indonesia pengguna *twitter* yang melihat performa timnas Indonesia di piala AFF 2020, peneliti menyimpulkan bahwa masyarakat Indonesia merasa puas dan bangga terhadap performa timnas Indonesia di piala AFF 2020.

Penelitian selanjutnya mungkin bisa menggunakan seorang ahli atau pakar untuk membantu menentukan sentimen *tweet* yang didapat agar sentimen yang telah di dapatkan menjadi lebih akurat.

Pengembangan teknik *preprocessing* bisa lebih bersih dan akurat, dikarenakan banyaknya *slangwords* yang digunakan didalam *tweet* masyarakat Indonesia, dan juga banyak *tweet* yang dibuat dengan bahasa daerah ataupun bahasa inggris.

REFERENSI

[1] M. Astiningrum, M. Haniah, and Y. R. Y. Pradana, "Analisis Sentimen Tentang Opini Terhadap Performa Timnas Sepak Bola Indonesia Pada Twitter," *Semin. Inform. Apl. Polinema*, pp. 35— 39, 2020.

[2] I. Safutra, "Piala AFF 2020, Harapan Baru di Generasi 23,8 Tahun." 2021, [Online]. Available: https://www.jawapos.com/sepak-bola/sepak-bola-indonesia/09/12/2021/piala-aff-2020-harapan-baru-di-generasi-238-tahun/. diakses tanggal 29 Juli 2022.

[3] C. B. Saputra, A. Muzakir, and D. Udariansyah, "Analisis Sentimen Masyarakat Terhadap #2019Gantipresiden Berdasarkan Opini Dari Twitter Menggunakan Metode Naive Bayes Classifier," *Bina Darma Conf. Comput. Sci.*, pp. 403–413, 2019.

[4] K. Fatmawati and A. P. Windarto, "Data Mining: Penerapan Rapidminer Dengan K-Means Cluster Pada Daerah Terjangkit Demam Berdarah Dengue (Dbd) Berdasarkan Provinsi," *Comput. Eng. Sci. Syst. J.*, vol. 3, no. 2, p. 173, 2018.

[5] M. Arifin, "Implementasi Data Mining Pada Prediksi Pemesanan Menggunakan Algoritma Apriori (Studi Kasus : Kimia Farma)," *J. Pelita Inform.*, vol. 8, no. 3, pp. 353–356, 2020.

[6] S. Rodiyansyah, "Algoritma Apriori untuk Analisis Keranjang Belanja pada Data Transaksi Penjualan," *Infotech J.*, vol. 1, no. 2, pp. 36–39, 2015.

[7] Fajri, "Kamus Lexicon Positif dan Negatif." 2018, [Online]. Available:

https://github.com/fajri91/InSet. diakses tanggal 29 Juli 2022.

[8] A. Pandu *et al.*, "Analisis Sentimen Twitter Pasca Pengumuman Hasil Pilpres 2019 Menggunakan Metode Lexicon Analysis," *J. Tek. Inform.*, vol. 15, no. 1, pp. 33–44, 2020.

[9] S. K. Trajd, "Indonesia Stopwords, github.com." 2021, [Online]. Available: https://github.com/SokKanaTorajd/indonesia-stopwords. Diakses tanggal 27 Juli 2022.

[10] A. Librian, "Kata Dasar, github.com." 2015, [Online]. Available: https://github.com/sastrawi/sastrawi/tree/master/data. Diakses tanggal 29 Juli 2022.