

Sentimen Analisis pada Data Tweet Pengguna Twitter Terhadap **Produk** **Penjualan Toko Online** Menggunakan Metode **K-Means**

Andris Faesal, Aziz Muslim, Aditya Hastami Ruger, Kusrini
Program Studi Megister Teknik Informatika Universitas Amikom Yogyakarta



APRISCHA
NAUVA M



140810180063



Latar Belakang

Tokopedia merupakan salah satu *market place* yang menjadi 10 top brand e-commerce yang paling banyak dibicarakan di Twitter pada quarter pertama tahun 2019.



Twitter memiliki fitur bernama *Trends for you* yang memungkinkan pengguna untuk mendapatkan informasi yang sesuai trend yang tampil saat itu. Pengguna juga dapat melakukan pencarian dengan kata untuk sebuah postingan. Postingan pada fitur tersebut menampilkan postingan kata yang begitu cepat dan jumlah postingan yang sangat banyak.



Maka jika topik pembicaraan adalah sebuah produk, selanjutnya akan memungkinkan pengguna mencari dimana produk tersebut didapatkan sehingga pelaku bisnis yang menggunakan *market place* seperti Tokopedia akan sangat terbantu dalam proses penjualan dan promosi produknya.



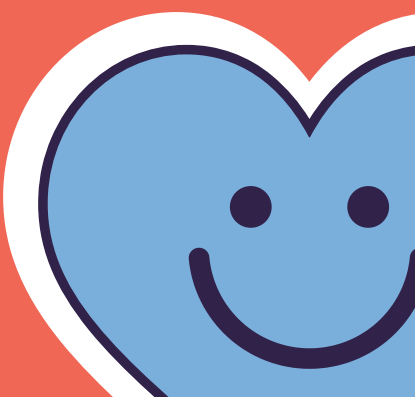
Namun ★

Penulis menemukan masalah yaitu pelaku bisnis tersebut **salah sasaran** saat mempromosikan produk serta market placenya



★ Sehingga

Di penelitian ini, digunakan metode **K-Means** untuk proses *cluster* dimana akan ada **pra-prosesing terlebih dahulu** untuk mengolah data tweet yang bertujuan untuk menghasilkan pengelompokan tweet berdasarkan quantity tweet.





Metode

Penelitian Eksperimen

Penelitian yang dilakukan dengan mengadakan manipulasi terhadap objek penelitian serta adanya variable yang terkontrol.



Alur Penelitian



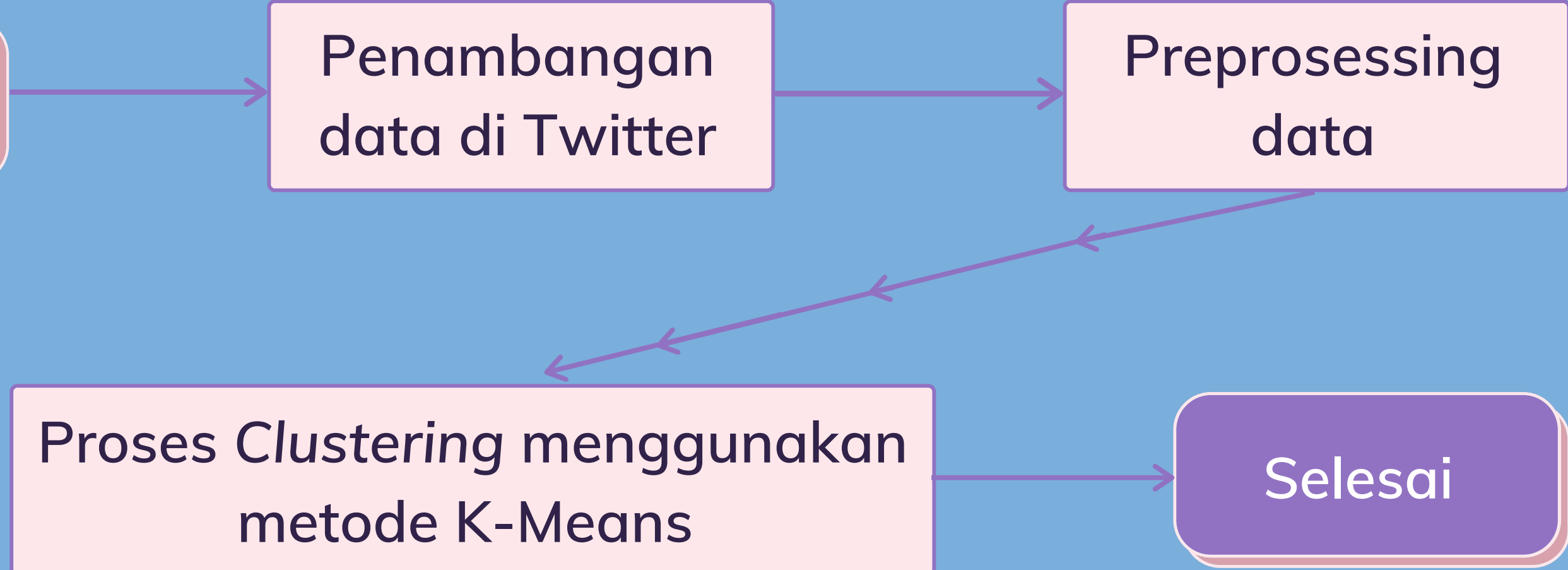
Mulai

Penambahan
data di Twitter

Preprocessing
data

Proses *Clustering* menggunakan
metode K-Means

Selesai





Proses Penambangan Data R Studio - Bahasa Pemrograman R

Untuk mengakses data tweet, diperlukan authentication twitter (API-Twitter) diantaranya `consumer_key`, `consumer_secret`, `access_token`, dan `access_secret`.

```
#crawling data dari twitter (tokopedia tweet)
download.file(url="http://curl.haxx.se/ca/cacert.pem", destfile="cacert.pem")

reqURL <- "https://api.twitter.com/oauth/request_token"
accessURL <- "https://api.twitter.com/oauth/access_token"
authURL <- "https://api.twitter.com/oauth/authorize"

CUSTOMER_KEY <- "pMZc[REDACTED]z" #consumerAPIkey
CUSTOMER_SECRET <- "f[REDACTED]avwCZySuQ15dLRNc68b3spWC9Rhsr" #APISecretkey
ACCESS_TOKEN <- "1104[REDACTED]5yFuq5yANeXkVRnFgit0WlFszB" #Accesstoken
ACCESS_secret<- "eBAI[REDACTED]CPsNrKsmNOTw4ymrINi1L" #Accesstokensecret

setup_twitter_oauth(CUSTOMER_KEY, CUSTOMER_SECRET, ACCESS_TOKEN, ACCESS_secret)
```

Jumlah tweet yang dipanggil sebanyak 2000 tweet menggunakan kata kunci "tokopedia" dan berbahasa indonesia.

```
search.string <- "tokopedia"
no.of.tweets <- 2000
some_tweets1 <- searchTwitter(search.string, n=no.of.tweets, lang = "id",)
some_tweets <- do.call("rbind", lapply(some_tweets1, as.data.frame))
```


2

Preprosesing Data

Menfilter kata tweet yang anti word yaitu menghapus kata-kata yang tidak penting

```
myCorpus <- Corpus(VectorSource(some_tweets$text))
myCorpus <- tm_map(myCorpus, content_transformer(tolower))
removeURL <- function(x) gsub("http[[:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeURL))
removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeNumPunct))
some_txt = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", some_txt)
some_txt = gsub("@\\w+", "", some_txt)
some_txt = gsub("[[:punct:]]", "", some_txt)
some_txt = gsub("[[:digit:]]", "", some_txt)
some_txt = gsub("http\\w+", "", some_txt)
some_txt = gsub("[ \\t]{2,}", "", some_txt)
some_txt = gsub("^\\s+|\\s+$", "", some_txt)
```

PERTAMA. Selanjutnya menghilangkan entitas tweet, username, tanda baca, angka, link, emot, dan objek tidak penting lainnya

```
some_txt<- lapply(tokenize_words(some_txt[]), stemming)
some_txt[]
some_txt <- tokenize_words(some_txt)
some_txt
```

KEDUA. Tokenizing yaitu pemotongan teks input menjadi kata, istilah, symbol, tanda baca, atau elemen lain yang memiliki arti

```
library(tm)
stopwords <- read.csv("/Users/aprischanauva/Documents/manjah/singkatan-lib.csv", header = FALSE)
stopwords <- as.character(stopwords$V1)
stopwords <- c(stopwords, stopwords())

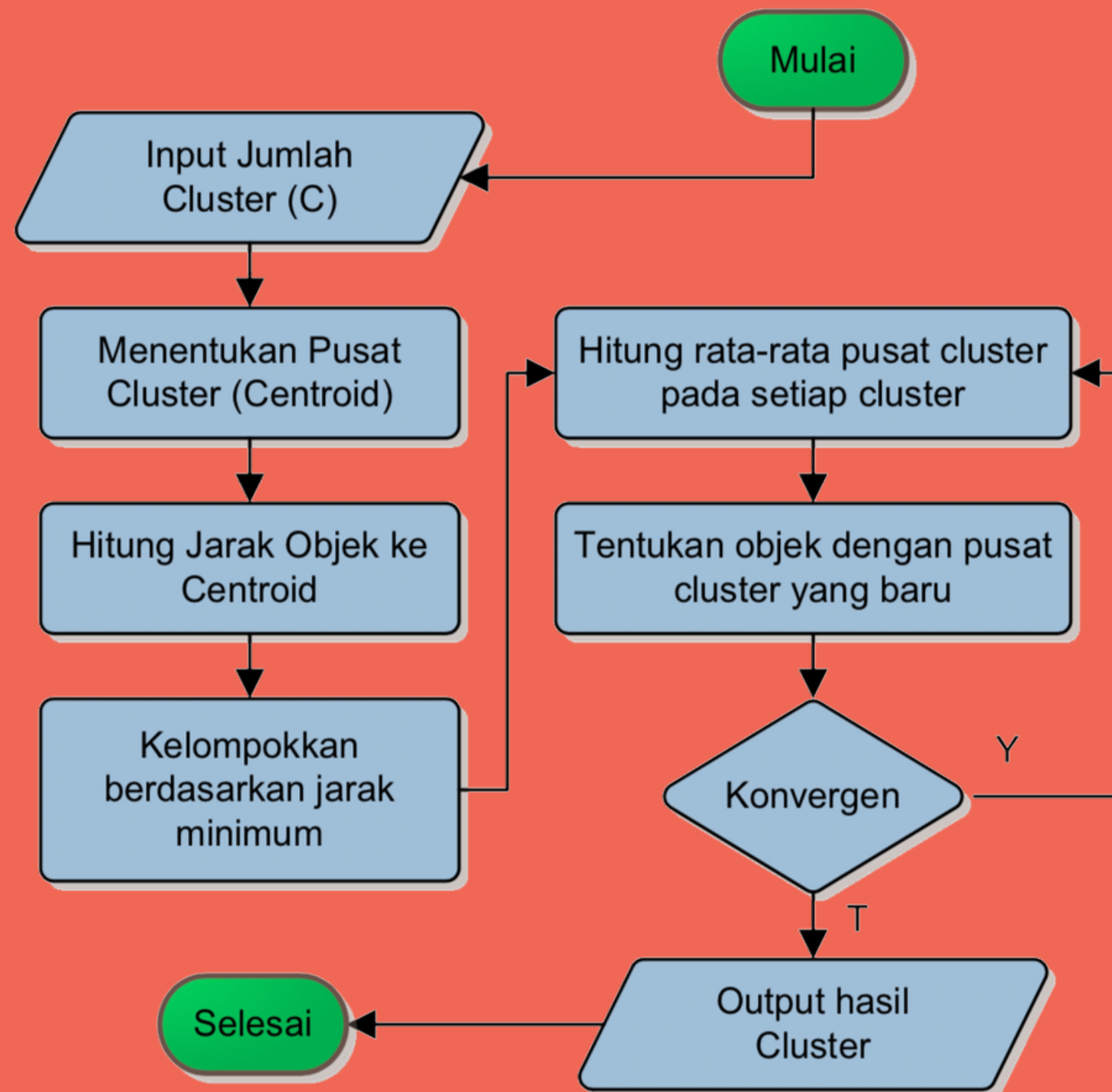
some_txt <- VectorSource(some_txt)
some_txt <- VCorpus(some_txt)
some_txt <- tm_map(some_txt, content_transformer(tolower))
some_txt <- tm_map(some_txt, removeWords, stopwords)
some_txt <- tm_map(some_txt, stripWhitespace)
```

KETIGA. Proses Stop Words merupakan kata yang diabaikan dalam pemrosesan.

3

K-Means

Proses Clustering



Di dalam menentukan titik centroid kita menggunakan perhitungan jarak Euclidian Distance:

$$d_{ij} = \sqrt{[(x_i - x_j)^2 + (y_i - y_j)^2]}$$

x_i = koordinat x untuk fasilitas i

y_i = koordinat y untuk fasilitas i

d_{ij} = jarak antar fasilitas i dan j

Menghitung nilai centroid cluster dari i dan seterusnya menggunakan rumus:

$$V_{ij} = \frac{\sum_{k=1}^{N_i} X_{kj}}{N_i}$$

N_i = jumlah data yang merupakan anggota dari cluster ke-i



Hasil dan Pembahasan





Penambahan Data Tweet

Dari pemanggilan data twitter sebanyak 2000 data, (n = 2000), data tweet yang berkaitan dengan kata kunci “tokopedia” berhasil dikumpulkan sebanyak **1.975 tweet** yang merupakan konsumen atau akun twitter yang berasal dari Indonesia. Berikut contoh hasil data tweet teratas.

	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID	statusSource
1	RT @mecitysx: Bismillahirrahmanirrahim Kalo ini lewa...	FALSE	0	NA	2020-12-30 19:04:29	FALSE	NA	1344358713994268672	NA	<a href="http://tv
2	RT @ryanandr_n_: Pantangan terbesar lelaki adalah wa...	FALSE	0	NA	2020-12-30 18:58:21	FALSE	NA	1344357174131052544	NA	<a href="http://tv
3	jual #Chloskin di #tokopedia https://t.co/1e91k1ehEZ	FALSE	0	NA	2020-12-30 18:58:17	FALSE	NA	1344357156770856962	NA	<a href="https://i
4	Nungguin tokopedia ipo 🙌	FALSE	0	NA	2020-12-30 18:56:59	FALSE	NA	1344356830030413824	NA	<a href="http://tv
5	https://t.co/znEemSacZk apakah kamu sedang menc...	FALSE	0	NA	2020-12-30 18:54:04	FALSE	NA	1344356093003141120	NA	<a href="http://tv
6	Hp baru https://t.co/2wq32RjevJ	FALSE	0	NA	2020-12-30 18:53:56	FALSE	NA	1344356061952688128	NA	<a href="http://tv
7	RT @AyahKhaira81: @Hobii_luvv @tokopedia Aamiin,...	FALSE	0	NA	2020-12-30 18:49:01	FALSE	NA	1344354822124802048	NA	<a href="http://tv
8	menikah sukses uwu??? that uwu... https://t.co/xOy1...	FALSE	0	NA	2020-12-30 18:47:41	FALSE	NA	1344354488740507648	NA	<a href="http://tv
9	@tokopedia @TokopediaCare Sukses	FALSE	0	tokopedia	2020-12-30 18:44:00	FALSE	1344179666353967104	1344353561950388224	72293042	<a href="http://tv
10	RT @tokopedia: Untuk yang belum menang! Jangan s...	FALSE	0	NA	2020-12-30 18:44:00	FALSE	NA	1344353560784293893	NA	<a href="http://tv
11	RT @Bukanorangbija2: Triglav spinX Bahan eva pylon...	FALSE	0	NA	2020-12-30 18:42:59	FALSE	NA	1344353304009003009	NA	<a href="http://tv
12	jual Stiker Kulkas 2 Master Chef Aneka Warna ~ https:...	FALSE	0	NA	2020-12-30 18:42:55	FALSE	NA	1344353288670519296	NA	<a href="http://tv
13	@shufi_ https://t.co/rsNn463WAZ Ini enak. https://t....	FALSE	0	wotkins	2020-12-30 18:41:32	TRUE	1344352132065943554	1344352941906354176	38210335	<a href="https://r
14	https://t.co/OVpyxRWKm8 dj sakit pinggang full albu...	FALSE	0	NA	2020-12-30 18:38:21	FALSE	NA	1344352139804479488	NA	<a href="http://tv
15	RT @mecitysx: Bismillahirrahmanirrahim Kalo ini lewa...	FALSE	0	NA	2020-12-30 18:37:04	FALSE	NA	1344351818118172672	NA	<a href="http://tv
16	jual Stiker Pintu Kamar Mandi Keroppi ~ https://t.co/...	FALSE	0	NA	2020-12-30 18:35:55	FALSE	NA	1344351526647595008	NA	<a href="http://tv
17	RT @Choko_Publisher: [PRE-ORDER TUAN MUDA KIM ...	FALSE	0	NA	2020-12-30 18:35:47	FALSE	NA	1344351493034414083	NA	<a href="https://r
18	Baikan https://t.co/wBAGSrGcHL	FALSE	0	NA	2020-12-30 18:34:39	FALSE	NA	1344351206911578112	NA	<a href="http://tv
19	Dirinya https://t.co/wWgIVTL7Xi	FALSE	0	NA	2020-12-30 18:33:47	FALSE	NA	1344350987977256961	NA	<a href="http://tv
20	@emerson_yuntho @tokopedia @TokopediaCare Pern...	FALSE	0	emerson_yuntho	2020-12-30 18:33:41	FALSE	1344202490749730816	1344350963490988032	155928815	<a href="https://r
21	@andishaizadi tokopedia	FALSE	0	andishaizadi	2020-12-30 18:27:53	FALSE	1344344413992935425	1344349503940227072	1191006456209801217	<a href="http://tv

Contoh hasil data tweet



2 Proses Preprocessing

PERTAMA. Data tweet yang terkumpul akan di hilangkan entitas tweetnya, nama pengguna, tanda baca, angka, link URL, emot, dan objek yang tidak penting lainnya. Berikut hasil tweet setelah proses preprocessing pertama.

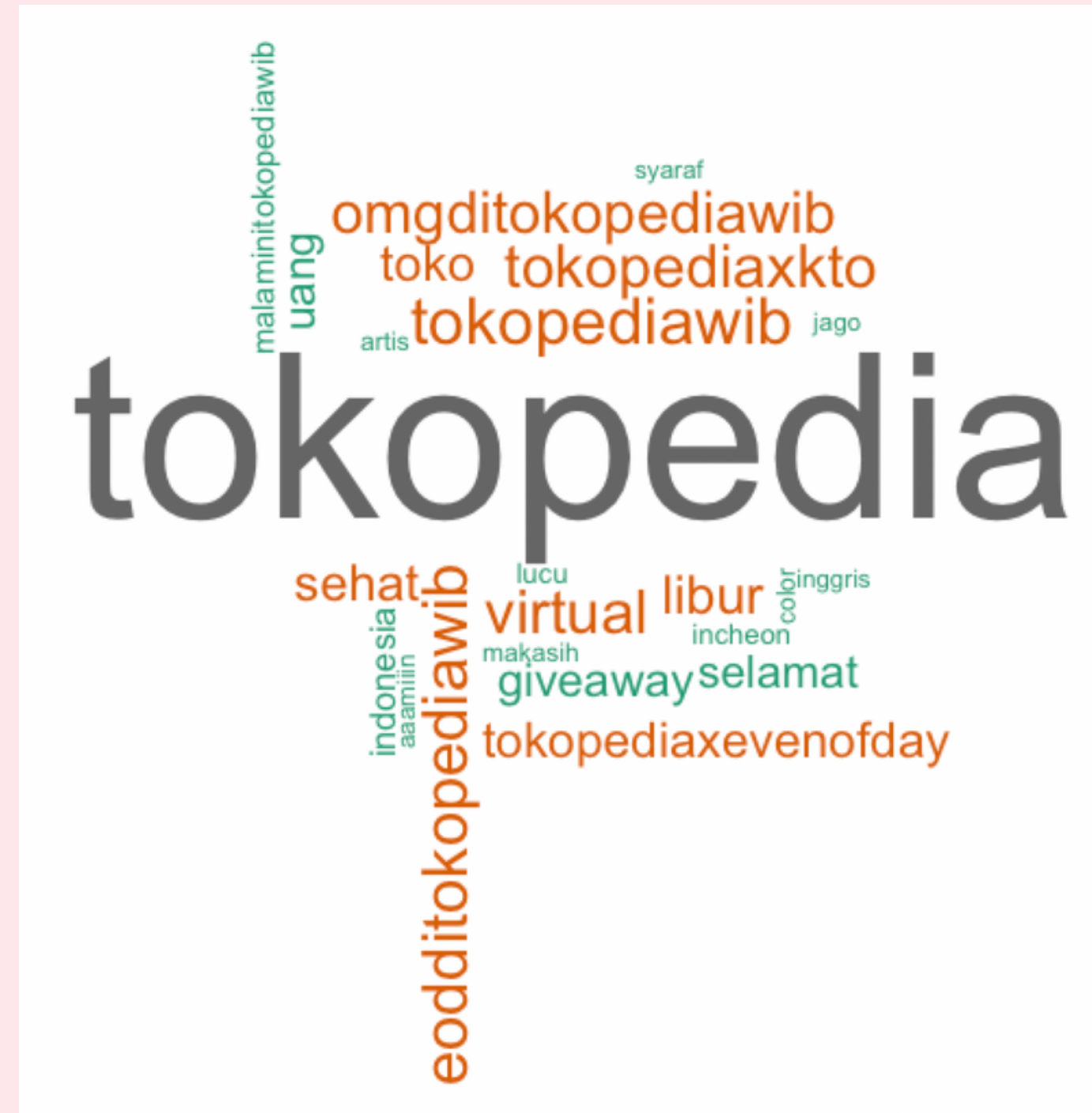
tokopedia sukses menikah aamiin
ner nget belasan paket dr toped
yang retweet semoga jadi pahala
preorder tuan muda kim a taekook fanfiction by jadwal po desember sdjanuari via tokopedi
yang mau order pomade bisa langsung gtgtgt tokopediadeklimisterpomadeshop tokopedia
menikah pacar
sukses
tokopedia ini emg ya sering ngasik cashback trus beli kuota harganya segini biasanya an seringseringlah begi
ada kabar ttg taeyong
pacar sukses tokopedia pacar udah jd suami semoga jadi sukses nyari diskon di tokopedia
sukses tokopedia pacar wkwkwk
sandal slop pria yang nyaman di pakai untuk aktifitas harian kamu triglav spinx bahan eva pylonruber size ringa
makan martabaktelor bebek
menikahsuksesuwu
kursi pijat ini cocok untuk para tenaga kesehatan di rumah sakit xuanwu beijing profesor qi
peneliti memeriksaorang yang melaporkan sakit punggung kronis individu dibagi menjadi
semangat pagi menjemput rezeki izinkan kang sandal menawarkan kembali sandal gunung triglav premium size bahan e
sandal slop pria yang nyaman di pakai untuk aktifitas harian kamu triglav spinx bahan eva pylonruber size ringa

KEDUA. Proses Tokenizing. KETIGA. Proses Stop Words atau menghilangkan entries data yang tidak diperlukan seperti "aku", "kamu", "kita", dll. Hasil entries setelah ketiga proses tersebut sebagai berikut:

word	freq
tokopedia	286
tokopediawib	58
eodditokopediawib	55
virtual	52
tokopediavaxto	51
omgditokopediawib	49
libur	42
sehat	38
toko	37
tokopediaxevenofday	36
giveaway	33
selamat	31
uang	30
indonesia	17
malamininitokopediawib	12
incheon	7
lucu	7
artis	4
makasih	4
...	...



2 Proses Preprocessing



Word Cloud dari Proses
Preprocessing



2 Proses Preprocessing

beberapa contoh entries data yang tidak digunakan:

"libur", "hmchh")
"kongrets", "lupa", "undang", "aqz", "ke..."
"keyword", "back", "strecher", "massage")
"insyaallah", "kerja", "tokopedia")
"hmmmm", "tokopedia", "aminkeunnn")
"hm", "lets")
"hm")
"hm", "hm")

("tbtb", "rezeki", "nomplok", "wkwk")
("tanggap", "kendala", "status", "transaksi", "nomor", ...
("tanda", "jodoh", "tokopedia")
("tahunbanyak", "berkah", "giat", "lancarin", "diridhoi"...
("taehyung", "jodoh", "ku")
("syukur", "syukur", "krnsdh", "dgn", "suka", "duka")
("syukur", "kasih", "nafas", "sehat", "kyknya", "ngeluh"...
("syukur", "capai")
("suskes", "tokopediaaaaa")
("suksesmenikahpacar", "wkwkwkw")



3 Proses Clustering dengan K-Means

```
dtm <- TermDocumentMatrix(some_txt)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
```

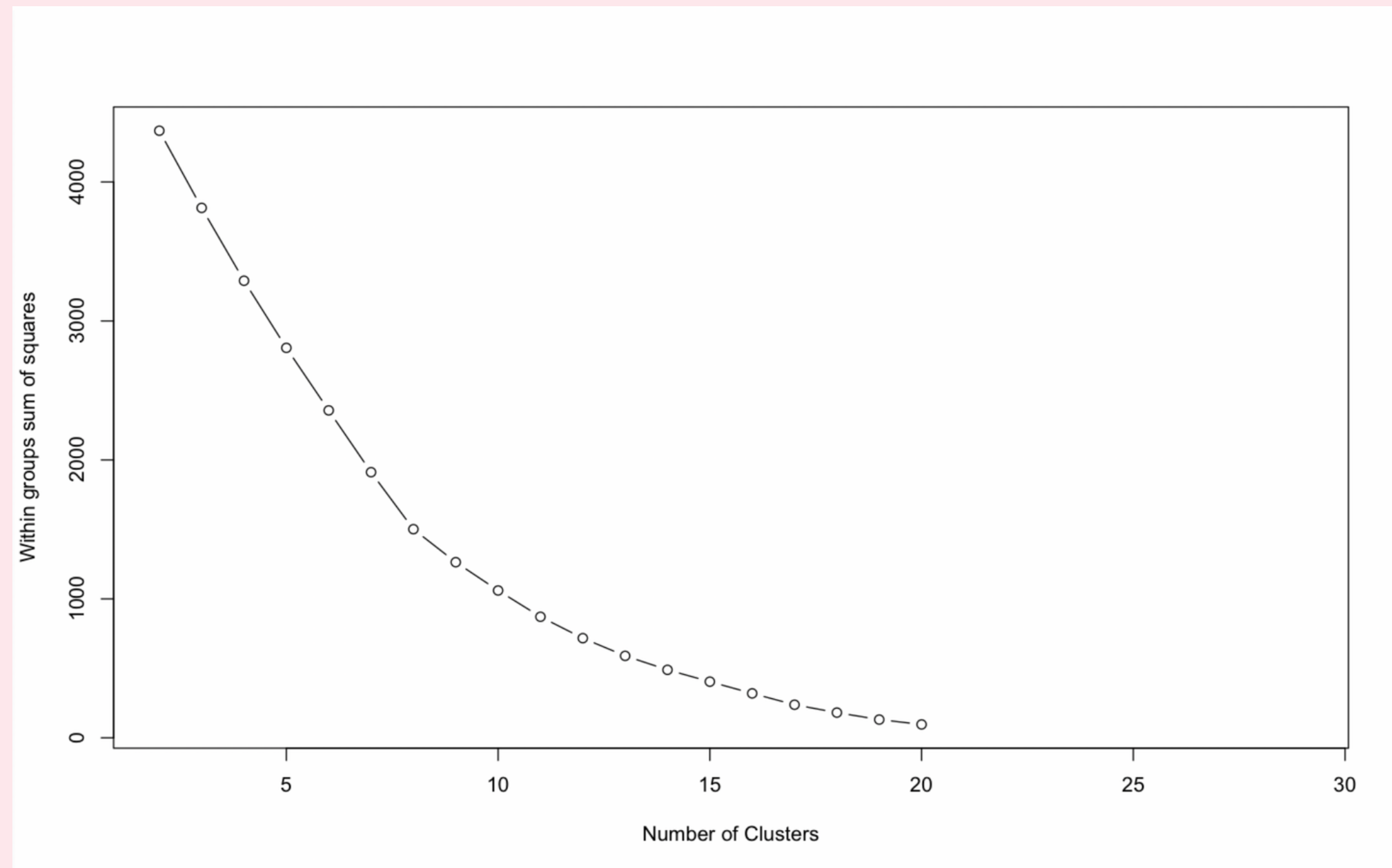
```
tdm2 <- dtm[-c(2,4,19,28,29,30,31,39,40,41,49,73,91,100,101,110,164,194,201,220,250),
tdm2
```

```
tdm <- tm::DocumentTermMatrix(some_txt)
tdm.tfidf <- tm::weightTfIdf(tdm2)
tdm.tfidf <- tm::removeSparseTerms(tdm.tfidf, 0.999)
tfidf.matrix <- as.matrix(tdm.tfidf)
dist.matrix = proxy::dist(tfidf.matrix, method = "cosine")
```

	1	2	3	4	5	6	7	8	9	10	11	12	13
akal	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
alami	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
artis	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
eodditokopediawib	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
giveaway	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
giveawaynya	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
incheon	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
indonesia	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
inggris	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
internet	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
kumpul	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
libur	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
lucu	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
makasih	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
malaminitokopediawib	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
menikahliburansukses	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
omgditokopediawib	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0
sehat	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	5.126926	0	0.000000	0
selamat	0	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	0.000000	0	0.000000	0

3 Proses Clustering dengan K-Means

```
wss <- 2:20  
for (i in 2:20) wss[i] <- sum(kmeans(tfidf.matrix,centers=i,nstart=25)$withinss)  
dev.new(width = 10000000, height = 10000000, unit = "px")  
plot(2:29, wss[2:29], type="b", xlab="Number of Clusters",ylab="Within groups sum of squares")
```



3 Proses Clustering dengan K-Means

```
k2 <- kmeans(tfidf.matrix, 3)
print(k2)
```

K-means clustering with 3 clusters of sizes 3, 1, 22

Clustering vector:

akal	alami	artis	eodditokopediawib	giveaway	giveawaynya	incheon	indonesia
3	3	3	3	3	3	3	3
inggris	internet	kumpul	libur	lucu	makasih malaminitokopediawib	menikahliburansukses	
3	3	3	3	3	3	3	3
omgditokopediawib	sehat	selamat	toko	tokopedia	tokopediawib tokopediaxevenofday	tokopediaxevenofday	tokopediaxevenofday
3	1	2	3	3	3	3	1
uang	virtual						
3	1						

Within cluster sum of squares by cluster:

```
[1] 617.8373 0.0000 3308.2839
(between_SS / total_SS = 20.4 %)
```

3 Proses Clustering dengan K-Means

Jumlah cluster yang ditentukan adalah 3 cluster. Berikut hasil dari proses clustering:

tweet sedang

sehat
virtual
tokopedia xkto

tweet jarang muncul

selamat

tweet terbanyak/sering muncul

akal
inggris
oomg ditokopedia wib
uang
alami
internet
artis
kumpul
eod ditokopedia wib
libur
toko
giveaway
lucu
tokopedia
tookopedia wib
incheon
malam ini tokopedia wib
tokopedia x even of day
indonesia
menikah libur sukses

Hasil Analisa



Clustering pada data tweet untuk kata kunci tokopedia dengan mengelompokkannya menjadi 3 cluster yaitu **sering**, **sedang**, dan **jarang**. Hal ini memberi arti bahwa kata-kata yang sering muncul merupakan kata yang di sarankan untuk digunakan dalam tweet sebagai media promosi.

Berdasarkan wordcloud dan cluster sering muncul didapatkan jika kata **tokopedia**, **libur**, dan **tokopedia wib** adalah kata yang paling sering dibicarakan di Twitter dan dapat dijadikan sebagai bahan promosi kedepannya.

★ Kesimpulan ★



Berdasarkan hasil analisa dari 1.975 tweet yang berhubungan tentang Tokopedia dengan menggunakan metode K-Means untuk 3 kelompok kata berdasarkan frekuensi kemunculannya yaitu sering digunakan, sedang, dan jarang digunakan.



Keyword "Tokopedia", "Libur", dan "Tokopedia WIB" bisa digunakan menjadi media promosi untuk kedepannya. Antusias dari masyarakat mengenai Campaign Tokopedia WIB masih sangat besar, ini dapat dijadikan sebuah potensi untuk Tokopedia melakukan promosi. Selain itu kata "Libur" dapat dijadikan salah satu konten untuk meningkatkan engagement Tokopedia.



THANKYOU 😊



Daftar Pustaka

C. Gallagher, E. Furey, and K. Curran, “The application of sentiment analysis and text analytics to customer experience reviews to understand what customers are really saying,” *Int. J. Data Warehous. Min.*, vol. 15, no. 4, pp. 21–47, 2019, doi:10.4018/IJDWM.2019100102.

S. A. F. Alvi Pranandha Syah, Adiwijaya, “Analisis Sentimen Pada Data Ulasan Produk Toko Online Dengan Metode Maximum Entropy Sentiment Analysis on Online Store Product Reviews With Maximum,” *e-Proceeding Eng.*, vol. 4, no. 3, pp. 4632–4640, 2017.

Y. W. Syaifudin and R. A. Irawan, “Implementasi Analisis Clustering Dan Sentimen Data Twitter Pada Opini Wisata Pantai Menggunakan Metode K-Means,” *J. Inform. Polinema*, vol. 4, no. 3, p. 189, 2018, doi: 10.33795/jip.v4i3.205.

M. H. Siregar, “Data Mining Klasterisasi Penjualan Alat-Alat Bangunan Menggunakan Metode K-Means (Studi Kasus Di Toko Adi Bangunan),” *J. Teknol. Dan Open Source*, vol. 1, no. 2, pp. 83–91, 2018, doi: 10.36378/jtos.v1i2.24.

J. W. Creswell, *Education Research. Planning, coduction and evaluating quantitative and qualitative research*, 4th ed. Boston: Pearson, 2012.