Penerapan PCA (*Principal Component Analysis*) pada Deteksi Outlier untuk Data *Text*

1st Marinda Endi Lestari
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
marindaendi@student.telkomuniversity.ac.id

2nd Ibnu Asror
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
iasror@telkomuniversity.ac.id

3rd Indra Lukmana Sardi
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
indraluk@telkomuniversity.ac.id

Abstrak—Data Mining adalah kegiatan pengumpulan data, pemakaian data historis, untuk menemukan keteraturan pola dalam dataset yang berukuran besar dan mempunyai jumlah yang banyak. Dalam data mining terdapat data yang berbeda dari data pada umumnya yang disebut *outlier*. *Outlier* sendiri berkaitan dengan nilai ekstrem, baik ekstrembesar maupun kecil. Adanya data outlier membuat analisis terhadap serangkaian data menjadi bias, atau tidak mencerminkan fenomena yang sebenarnya. *Outlier detection* digunakan untuk mendeteksi ada atau tidaknya *outlier* dalam sebuah data. *Outlier Detection* dapat digunakan untuk mendeteksi data berupa *categorical*, *numeric*,maupun data teks.*Principal Component Analysis* (PCA) merupakan salah satu metode pendeteksian *outlier* berdasarkan pendekatan *statistical*. Objek yang dianggap sebagai outlier adalah objek yang memiliki probabilitas yang rendah sehubungan dengan model distribusi probabilitas pada data tersebut. Evaluasi yang digunakan untuk mengetahui performansi sistem yaitu : *accuracy, precision,* dan *recall.*

Kata kunci—*outlier*, PCA, outlier detection, data teks

Abstract—Data Mining is an activity of collecting data, using historical data, to find regular patterns in large and large datasets. In data mining there is data that is different from data in general, which is called outliers. Outliers are related to extreme values, both large and small extremes. The existence of outlier data makes the analysis of a series of data biased, or does not reflect the actual phenomenon. Outlier detection is used to detect the presence or absence of outliers in a data. Outlier Detection can be used to detect data in the form of categorical, numeric, and text data. Principal Component Analysis (PCA) is one of the outlier detection methods based on a statisticalapproach. Objects that are considered as outliers are objects that have a low probability with respect to the probability distribution model on the data. The evaluations used to determine the performance of the system are: accuracy, precision, and recall.

Keywords—outlier, PCA, outlier detection, text data

I. PENDAHULUAN

A. Latar Belakang

Artikel berita adalah karangan faktual yang memiliki informasi tentang peristiwa terkini. Artikel berita memiliki panjang tertentu yang dibuat untuk dipublikasikan baik di media online maupun cetak, dan bertujuan untuk menyampaikan gagasan dan fakta yang dapat mendidik dan menghibur Topik umum beritapun beragam seperti laporan berita mengenai pemerintahan, pendidikan, politik, kesehatan, lingkungan, hiburan, ekonomi, bisnis, olahraga, dll. Saat ini lebih banyak orang yang membaca artikel berita melalui situs berita online, baik dalam negri maupun luar negri. Contoh artikel berita di luar negri adalah yang berasal dari situs berita BBC. BBC memiliki banyak sekali artikel berita dalam berbagai kategori, bukan hanya berita negara asalnya tetapi BBC jugamemiliki berita yang mencangkup di seluruh dunia.

Artikel berita diarsipkan dan di kategorikan sesuai dengan isi berita yang tercantum di dalamnya, akan tetapi sering terjadi kesalahan pengelompokkan di mana dokumen yang berisi artikel berita tersebut dapat saja berada di dalam kategori yang salah. Hal inilah yang jika di biarkan akan membuat dokumen dalam kategori yangada menjadi tidak teratur.

Sebuah data yang memiliki sifat dan karakteristik yang berbeda dari data pada umumnya dan kemunculan kejadian yang relatif sedikit dikatakan *outlier*[1]. *Outlier detection* adalah proses menemukan sebuah objek data dengan tingkah laku yang sangat berbeda dari harapan[3]. *Outlier* sendiri dapat didefinisikan sebagai data yang menyimpang terlalu jauh dari data pada umumnya dalam suatu rangkaian data[2]. Istilah *outlier* sendiri sering dikaitkan dengan nilai ekstrem, baik ekstrem besar atau kecil. Adanya *outlier* dalam data dapat menimbulkan masalah seperti *fraud detection*, masalah medis, kesalahan input data, atau kesalahan dalam sebuah dokumen.

Permasalahan dalam analisis *outlier* text menjadi penting dikarenakan banyaknya aplikasi web sentris dan media sosial yang memiliki banyak data text. Beberapa pengaplikasian dari analsis text outlier seperti *website management* yaitu halaman yang tidak biasa dari artikel yang ada di situs web dan dapat ditandai sebagai *outlier*, dan managemen artikel berita untuk menentukan artikel berita yang tidak biasa dari kategori dokumen berita[4].

PCA (*Principal Component Analysis*) mendeteksi outlier metode PCA bertujuan untuk memperoleh komponen utama yang tidak dapat dipengahuri oleh keberadaan *outlier*. Tahapan yang digunakan untuk mendapatkan data yang sesuai adalah sebagai berikut : *1)case folding, 2)tokenizing, 3)filtering, 4)stemming*.

B. Topik dan Batasannya

Artikel berita adalah karangan faktual yang

berisi informasi mengenai peristiwa terkini dengan panjang tertentu yang dibuat untuk dipublikasikan di media online maupun cetak. Artikel berita online disimpan dalam sebuah file dokumen dalam kategori berbeda-beda sesuai dengan isi berita dengan jumlah yang cukup banyak dandikelompokkan sesuai dengan kategorinya. Tetapi pengelompokan tersebut bisa saja terjadi kesalahanseperti meng-inputkan dokumen ke dalam kategori yang salah.

Topik yang dibahas pada penelitian ini adalah bagaimana mendeteksi *outlier* dengan PCA (*Principal Component Analysis*). Dalam penelitian ini, data yang diambil adalah data dari situs web berita BBC. Pada penelitian ini digunakan data berkategori *Tech* yang terdapat di empat folder yang berbeda dengan jumlah data yang berbeda di tiap folder. Data akan melalui tahap *preprocessing* agar data menjadi terstruktur. Tahapan yang selanjutnya adalah pembobotan dari data yang telah terstruktur dengan *term frequency-invers document frequency* (tf-idf), lalu dilakukan proses deteksi *outlier* menggunakan PCA. Dari hasil deteksi dapat ditentukan dokumen mana saja yang termasuk *outlier* dan mana yang bukan *outlier.*

C. Tujuan

Membangun sistem yang dapat mendeteksi outlier pada kumpulan dokumen text berita, dengan mengimplementasikan metode *Principal Component Analysis* (PCA) untuk mencari dokumen yang tidak termasuk ke dalam kategori *Tech*. Melakukan pengujian data dengan menganalisa hasil pendeteksian *outlier* menggunakanmetode PCA dan menghitung analisa keakuratan deteksi.

D. Organisasi Tulisan

Pada bab dua studi terkait, yang berisi mengenai teori yang mendukung pada penelitian ini. Pada bab tiga menjelaskan rancangan sistem yang dibangun. Pada bab keempat akan menjelaskan pengujian dan analisis dari penelitian yang dilakukan. Pada bab lima akan menjelaskan mengenai kesimpulan dari penelitian dan saran untuk penelitian selanjutnya.

II. KAJIAN TEORI

A. BBC Dataset

British Broadcasting Corporation (BBC) dibentuk tahun 1927 yang merupakan stasiun televisi dan radio yang berada di Britania Raya. Selain stasiun televisi dan radio BBC juga menyediakan berita di Internet. Dataset yang dipakai merupakan dataset yang didapat dari situs web berita BBC. Dataset ini berupa dokumen berita yang terkait dengan cerita di bidang *busiiness* dari tahun 2004-2005[12]. Di dalam beberapa folder dokumen berisi berita berkategori *Tech* berisi dokumen berekstensi .txt dengan jumlah dokumen berita berbeda di setiap foldernya. Dalam masing-masing folder tersebut berisi dokumen artikel berita yang memiliki panjang berita yang berbeda- beda.

B. Outlier

Outlier merupakan perilaku yang menyimpang dari data pada umumnya dari suatu rangkaian data. Istilah *outlier* sendiri sering dikaitkan dengan nilai ekstrem, baik ekstrem besar maupun ekstrem kecil[8]. Biasanya data yang tidak biasa akan diterjemahkan ke dalam beberapa jenis masalah seperti, masalah medis, cacat struktural, kesalahan input data atau kesalahan dalam dokumen *text*. *Outlier* memiliki informasi yang penting tentang karakteristik yang tidak biasa dari sistem dan dampak apa saja yang mempengaruhi data pada saat menjalankan sistem.

C. Preprocessing

Preprocessing adalah salah satu teknik *text mining* untuk mencari data dan mengekstrak informasi yang berguna dari data tekstual[5]. Tahapa dari *text preprocessing* yaitu :

1. *Case Folding* adalah proses mengubah semua huruf kapital menjadi huruf kecil (*lowercase*). Hanya huruf ‘a’ sampai ‘z’ yang akan di terima, dan juga tahap penghapusan tanda baca atau simbol pada kalimat (*remove punctuation*).
2. *Tokenization* adalah proses untuk memecah kumpulan kalimat menjadi kata, simbol, frasa, atau dapat disebut dengan token[6].
3. *Filtering* (*Stopword Removal*) adalah proses penyaringan kata kunci sehingga kata yang tidak bermakna akan dihapus. Hal ini berfungsi untuk menguraikan *text* data dan dapat meningkatkan kinerja sistem[5,6].
4. *Stemming* digunakan untuk menemukan batang atau akar dari kata. Fungsi lainnya yaitu untuk menghilangkan sufiks, menguraikan jumlah kata, mendapatkan data akar kata yang akurat, menghemat *memory space* dan waktu[5].

D. TF-IDF

TF-IDF (*Term Frequency – Invers Document Frequency*) adalah metode untuk melakukan pembobotsn antar kata (*term*) terhadap dokumen. TF merupakan frekuensi kemunculan sebuah term dalam dokumen. Semakin besar jumlah kemunculan suatu *term* dalam dokumen mka semakin banyak kemunculan suatu kata. Berikut adalah persamaan dari TF :

$$\mathrm{TF}(t,d) = \sum_{x \in d} \mathrm{fr}(\mathrm{x},\mathrm{t}) \quad (2.1)$$

Dimana fr(x,t) dapat didefinisikan sebgai berikut :

$$\mathrm{fr}(x,t) = \begin{cases} 1, & if\ x = t \\ 0, & otherwise \end{cases} \quad (2.2)$$

Sedangkan IDF dapat didefinisikan dengan persamaan berikut :

$$\mathrm{IDF}(t) = \log \frac{|D|}{1+|\{d:t \in d\}|} \quad (2.3)$$

Dimana $|\{d : t \in d\}|$ adalah banyaknya kemunculan *term* pada dokumen *d* dan |D| merupakan banyaknya dokumen yang diolah. Persamaan TF-IDF adalah sebagai berikut :

$$\mathrm{TFIDF}(t) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t) \quad (2.4)$$

Nilai TF-IDF bertabah secara proposional sesuai dengan jumlah kemunculan dari kata pada setiap dokumen. Semakin besar nilai TF maka semakin kecil nilai IDF, hal ini menunjukkan bahwa semakin penting kalimat kata maka nilai akan semakin kecil[7][11].

E. PCA (Principal Component Analysis)

Principal Component Analysis (PCA) adalah metode statistik yang dapat memecah matriks data menjadi matriks vektor yang disebut *principal component*. *Principal Component Anlaysis* pada dasarnya adalah teknik untuk penguran dimensi[9], namun dapat digunakan juga untuk beberapa tujuan berbeda, salah satunya dapat memeriksa sekumpulan data untuk menemukan *outlier.*

PCA (Principal Component Analysis) yang dirancang dapat digunakan untuk mendeteksi outlier. PCA (Principal Component Analysis) adalah teori dan teknik yang bertujuan untuk mendeteksi outlier, dan terlebih dulumenyesuaikan sebagian besar data yang kemudian menandai titik-titik tersebut yang menyimpang[11].

Dalam tugas akhir ini matriks dapat didekomposisi menjadi vektor ortogonal yang disebut dengan vector eigen terkait dengan nilai eigen. Vektor eigen dengan nilai eigen yang tinggi dapat menagkap sebagian besar varians dalam data. Maka hyperplane berdimensi rendah yang dibangun oleh k eigenvector dapat menangkap sebagian besar varians dalam data. Tetapi, outlier berbeda dari titik data nomal yang lebih jelas di hyperplane dibangun oleh vektor eigen dan nilai eigen yang kecil. Maka dari itu, skor outlier dapat diperoleh sebagai jumlah proyeksi jarak sample pada semua vektor eigen

F. Confussion Matrix

Confusion matrix memiliki informasi tentang actual dan prediction clasification yang dilakukan oleh sebuah sistem[10]. Dalam tugas akhir ini untuk mengukur performansi pada suatu data yaitu menggunakan perhitingan *accuracy, precision,* dan *recall*. Performansi pada suatu sistem biasanya dievaluasi menggunakan datadalam matrix. Berikut merupakan tabel *confussion matrix.*

TABEL 1
TABEL CONFUSSION MATRIX

	Predicted Negative	**Predicted Positive**
Actual Negative	a	b
Actual Positive	c	d

Keterangan pembentukan matrix :

a. *Accuration* (AC) merupakan bagian dari total nilai prediksi yang benar. Nilai tersebut dapat dirumuskansebagai berikut :

$$AC = \frac{a+d}{a+b+c+d} \quad (2.6)$$

b. *Recall* atau *True Positive* (TP) merupakan hasil analisa yang positif yang sesuai dengan prediksi. Secaramakna *recall* merupakan kualitas seberapa lengkap hasil relevan yang di tampilkan oleh sistem. Nilai tersebut dapat dirumuskan sebagai berikut :

$$TP = \frac{d}{c+d} \quad (2.7)$$

c. False Positive (FP) merupakan hasil analisa negatif dari data prediksi. Nilai dapat dirumuskan sebagaiberikut :

$$FP = \frac{b}{a+b} \quad (2.8)$$

d. True Negatif (TN) hasil analisa positif dari data aktual. Nilai dapat dirumuskan sebagai berikut :

$$TN = \frac{a}{a+b} \quad (2.9)$$

e. False Negatif (FN) merupakan hasil analisa negatif dari data aktual. Nilai tersebut dapat dirumuskansebagai berikut :

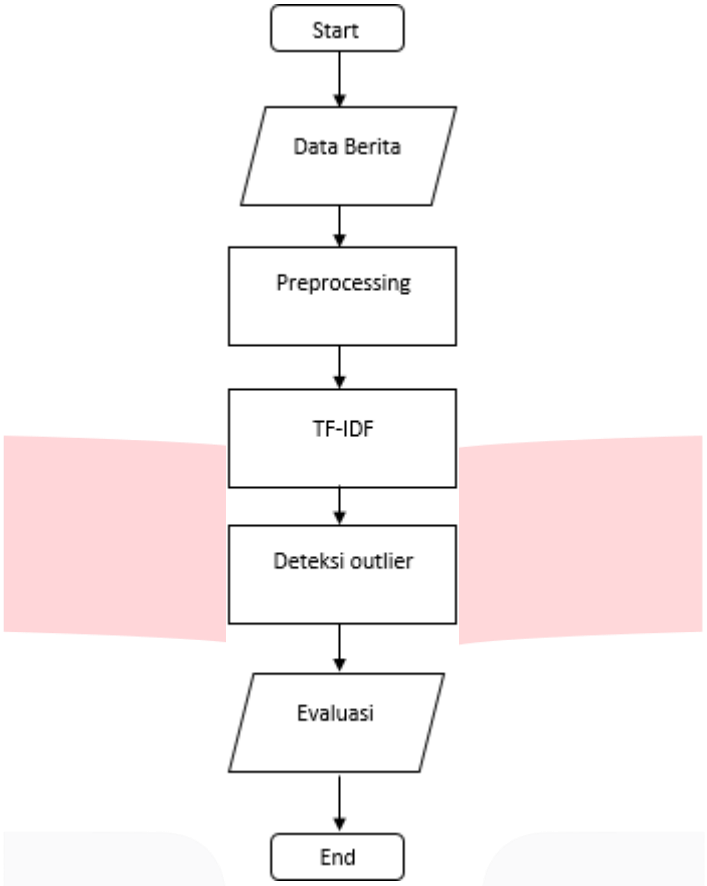
$$FN = \frac{c}{c+d} \quad (2.10)$$

f. *Precision* (P) merupakan hasil prediksi positif yang benar. Secara makna *precision* merupakan pengukuran kualitas seberapa berguna sistem. Nilai tersebut dapat dirumuskan sebagai berikut :

$$P = \frac{d}{b+d} \quad (2.11)$$

III. METODE

Gambaran umum sistem merupakan gambaran bagaimana sistem akan dibangun dan dilaksanakan pada penelitian ini. Sistem pada penelitian ini akan dibangun dengan tujuan mendapatkan hasil dari pendeteksian *outlier* pada dataset BBC yang nantinya akan di analisa lebih lanjut. Gambar 3.1 merupakan gambaran umum yang akandibangun.



GAMBAR 1
GAMBARAN SYSTEM SECARA UMUM

Pada Gambar 1 diatas, data .txt adalah kumpulan berita berkategori *tech*, terlebih dahulu melalui tahapan *preprocessing* agar data dapat diolah ke tahapan yang selanjutnya. Lalu setelah itu data masuk ke dalamtahap TF-IDF untuk pemberian bobot nilai pada pada setiap *term* yang selanjutnya masuk ke dalam perhitungan.Tahapan selanjutnya data akan di proses menggunakan *Principal Component Analysis* (PCA) untuk mendeteksi outlier. Tahapan akhir adalah evaluasi datauntuk mengetahui akurasi dari hasil run code.

A. BBC Dataset

Data yang dipakai merupakan data text dari web berita BBC. Dataset berkategori *tech* yang berisi dokumen berita. Kategori *tech* memiliki jumlah dokumen berbeda di setiap foldernya, dan data awalnya yaitu data *text*.

Berikut adalah contoh data artikel berita :

Dollar gains on Greenspan speech

The dollar has hit its highest level against the euro in almost three months after the Federal Reserve head said the US trade deficit is set to stabilise.

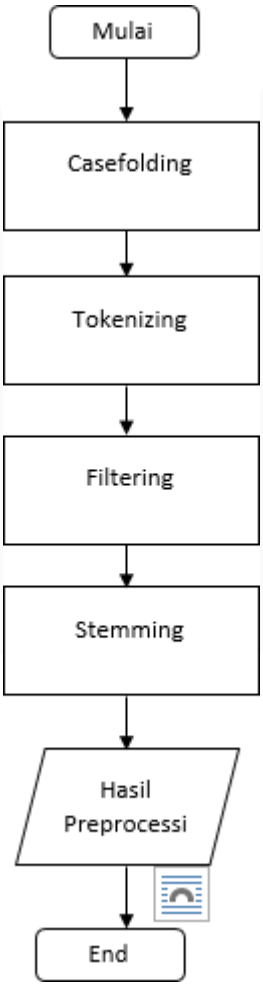
And Alan Greenspan highlighted the US government's willingness to curb spending and rising household savings as factors which may help to reduce it. In late trading in New York, the dollar reached $1.2871 against the euro, from $1.2974 on Thursday. Market concerns about the deficit has hit the greenback in recent months. On Friday, Federal Reserve chairman Mr Greenspan's speech in London ahead of the meeting of G7 finance ministers sent the dollar higher after it had earlier tumbled on the back of worse-than-expected US jobs data. "I think the chairman's taking a much more sanguine view on the current account deficit than he's taken for some time," said Robert Sinche, head of currency strategy at Bank of America in New York. "He's taking a longer-term view, laying out a set of conditions under which the current account deficit can improve this year and next."

Worries about the deficit concerns about China do, however, remain. China's currency remains pegged to the dollar and the US currency's sharp falls in recent months have therefore made Chinese export prices highly competitive. But calls for a shift in Beijing's policy have fallen on deaf ears, despite recent comments in a major Chinese newspaper that the "time is ripe" for a loosening of the peg. The G7 meeting is thought unlikely to produce any meaningful movement in Chinese policy. In the meantime, the US Federal Reserve's decision on 2 February to boost interest rates by a quarter of a point - the sixth such move in as many months - has opened up a differential with European rates. The half-point window, some believe, could be enough to keep US assets looking more attractive, and could help prop up the dollar. The recent falls have partly been the result of big budget deficits, as well as the US's yawning current account gap, both of which need to be funded by the buying of US bonds and assets by foreign firms and governments. The White House will announce its budget on Monday, and many commentators believe the deficit will remain at close to half a trillion dollars.

GAMBAR 2
CONTOH ARTIKEL BERITA

B. Preprocessing

Tujuan dari tahapan *preprocessing* adalah untuk penyeragaman data, agar data dapat diolah ke tahapselanjutnya. Tahapan *preprocessing* dapat dilihat pada Gambar :

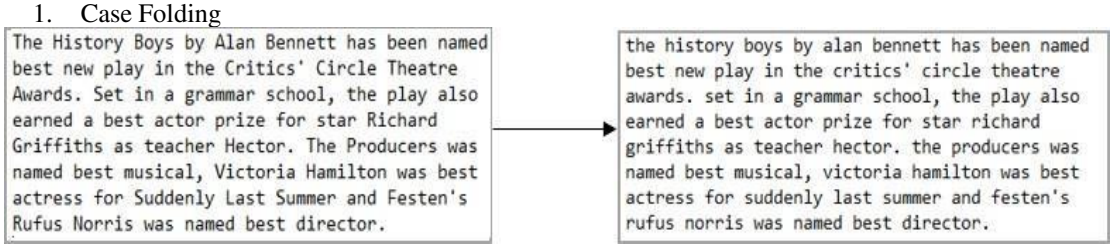


GAMBAR 3
TAHAPAN *PREPROCESSING*

Adapun contoh penerapan tahap *preprocessing* dapat di lihat pada Gambar :

1. Case Folding

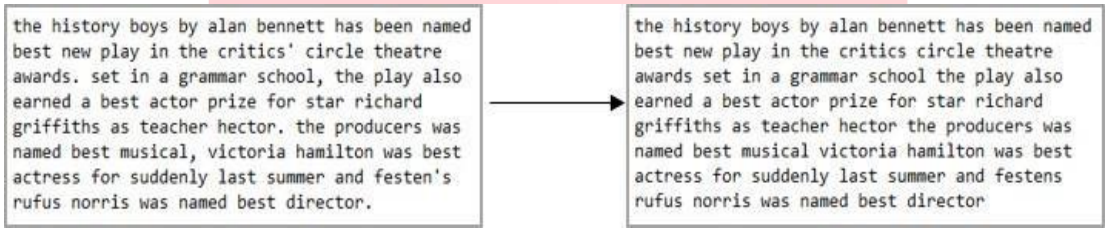
Pada proses *case folding* huruf yang ada dalam dokumen diubah menjadi huruf kecil (*lowercase*).



GAMBAR 4
CONTOH *CASE FOLDING*

2. Remove Punctuation

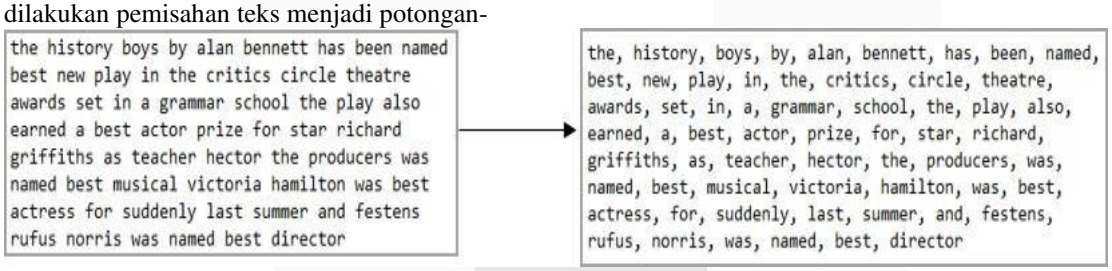
Pada proses ini dokumen yang dimiliki akan dilakukan penghapusan tanda baca atau simbol padakalimat.



GAMBAR 5
CONTOH PUNCTUATION REMOVAL

3. Tokenization

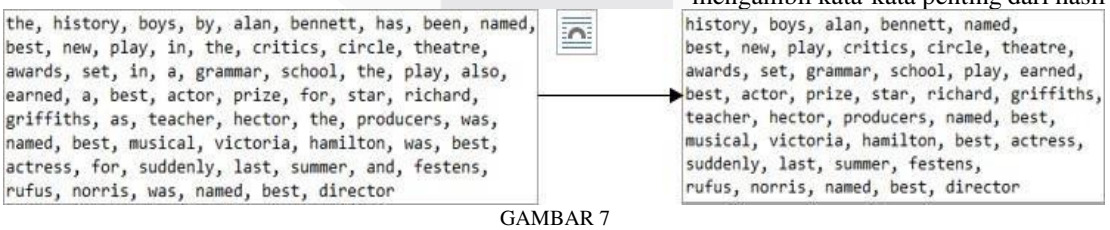
Pada tahap *tokenization* atau tokenisasi dilakukan pemisahan teks menjadi potongan-potonganyangdisebut sebagai token.



GAMBAR 6
CONTOH TOKENISASI PADA KALIMAT

4. Filtering (Stopword Removal)

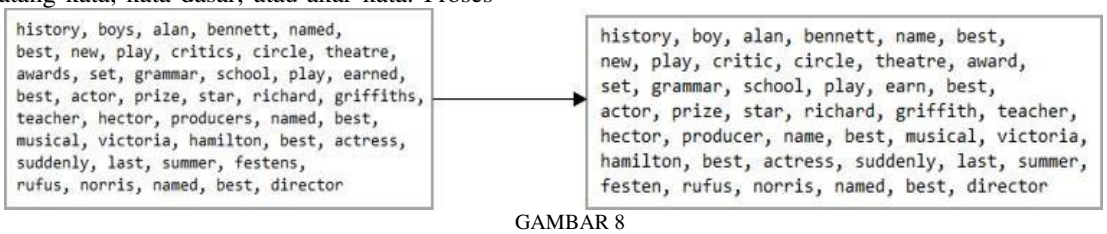
Filtering (*stopword removal*) adalah tahap mengambil kata-kata penting dari hasil token.



GAMBAR 7
CONTOH PROSES FILTERING (STOPWORD) REMOVAL PADA KALIMAT

5. Stemming

Stemming adalah proses mereduksi kata menjadi batang kata, kata dasar, atau akar kata. Proses *stemming* pada teks berbahasa Inggris hanya proses menghilangkan sufiks.



GAMBAR 8
CONTOH PROSES STEMMING PADA KALIMAT

C. TF-IDF

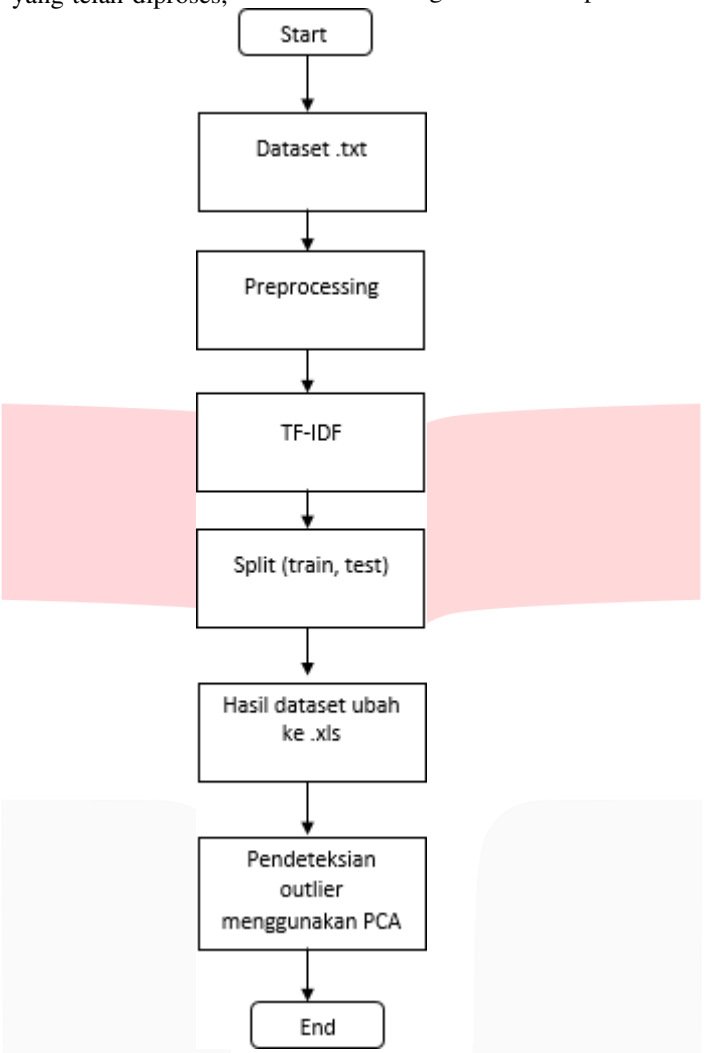
Pada tahap ini dokumen yang telah melalui *preprocesing* akan dihitung bobot (nilai) dengan menggunakan metode TF-IDF. Matrix yang terbentuk akan disesuaikan dengan jumlah kemunculan tiap kata dari tiap dokumen file sehingga matrix yang dihasilkan akan berbeda-beda. Berikut adalah contoh pembentukanmatrix TF-IDF berdasarkan hasil dari *preprocessing* :

TABEL 2
CONTOH PEMBENTUKAN MATRIKS TF-IDF

		TF					TF-IDF				
No	Kata	1	2	3	4	5	1	2	3	4	5
1	light	2	0	0	0	0	7,964	0	0	0	0
2	buy	0	0	1	0	0	0	0	3,982	0	0
3	paint	0	0	0	0	0	0	0	0	0	0
4	importantli	0	0	0	0	0	0	0	0	0	0
5	foreman	0	0	0	0	0	0	0	0	0	0
6	profit	0	0	0	0	0	0	0	0	0	0
7	see	0	0	0	0	0	0	0	0	0	0
8	work	1	1	1	1	0	1,869	1,869	1,869	1,869	0
9	hold	0	1	0	0	1	0	4,3005	0	0	4,3005
10	world	0	1	0	1	0	0	2,1032	0	2,1032	0
11	classic	0	0	1	0	0	0	0	3,6073	0	0
12	create	0	0	0	1	1	0	0	0	3,4355	3,4355
13	children	1	0	0	0	2	2,4679	0	0	0	4,9357
14	success	0	0	1	2	0	0	0	2,4481	4,8961	0
15	plan	0	1	1	0	0	0	3,161	3,161	0	0
16	assist	1	0	0	0	0	4,7705	0	0	0	0
17	katherin	0	0	0	0	0	0	0	0	0	0
18	helena	0	1	0	0	0	0	5,6868	0	0	0
19	intend	0	0	0	0	1	0	0	0	0	4,5881

D. Outlier Detection

Data matrix TF-IDF yang telah diproses, akan diolah langsungdalam sistem. Berikut adalah gambarandari proses :



GAMBAR 9
PROSES *OUTLIER DETECTION*

Pada mulanya dataset berita dalam bentuk format .txt yang selanjutnya dilakukan tahapan preprocessingdan tf-idf, setelah melalui tahapan tersebut data text di split/di bagi dua untuk keperluan data training dan data testing. Setelah di split dataset diubah ke dalam bentuk .xls agar dapat di terima sistem dan di proses, yang terakhir yaitu pendeteksian outlier menggunakan PCA.

IV. HASIL DAN PEMBAHASAN

A. Dataset

Pengujian yang dilakukan pada tugas akhir ini adalah untuk mengetahui tingkat keakuratan sistem deteksi *outlier* terhadap dokumen data text. Pengujian ini mengambil 5 folder dataset berita yang berkategori *tech*, masing- masing folder memiliki jumlah data yang berbeda. Masing-masing dataset nantinya akan di split datanya untuk keperluan data training dan data testing. Data training berguna untuk sistem agar belajar terlebih dahulu untuk membedakan data mana yang termasuk ke dalam kategori *tech* dan mana yang bukan.

B. Skenario Pengujian

Pada skenario pengujian ini data yang telah di proses sebelumnya melalui tahapan preprocessing dan tf-idfakan di split untuk keperluan data training dan data testing. Data yang semula dalam bentuk .txt diubah ke dalam bentuk .xls agar sistem dapat membaca data yang siap melalui tahapan pendeteksian *outlier* untuk mengetahui data mana yang tidak termasuk ke dalam kategori *tech,* data yang bukan kategori *tech* akan disebut sebagai *outlier*. Setelah itu akan dilanjutkan dengan perhitungan *confussion matrix* dengan menghitung nilai *accuracy, precision,*dan *recall.*

1. Skenario Pengujian 1

Pada skenario pengujian 1, akan menggunakan data yang terdapat di folder 1 yang berisi 20 data. Data tersebut akan dibagi menjadi 10 training dan 10

testing. Pengujian akan dilakukan dengan menggunakan data yang jumlahnya sedikit untuk melihat hasil dari pengujian tersebut, yang nantinya data akan bertambah banyak sesuai dengan skenario pengujian. Hasil dari perhitungan *accuracy, precision*, dan *recall* ada pada *Gambar 6, Gambar 7,Gambar 8* dan *tabel 1, tabel 6,* dan hasil run code ada pada *Gambar 1*

2. Skenario Pengujian 2

Pada skenario pengujian 2, akan menggunakan data yang terdapat di folder 2 yang berisi 40 data. Data tersebut akan dibagi menjadi 20 training dan 20 testing. Hasil dari perhitungan *accuracy, precision*, dan *recall* ada pada *Gambar 6, Gambar 7, Gambar 8* dan *tabel 2, tabel 6,* dan hasil run code ada pada *Gambar 2*

3. Skenario Pengujian 3

Pada skenario pengujian 3, akan menggunakan data yang terdapat di folder 3 yang berisi 50 data. Data tersebut akan dibagi menjadi 25 training dan 25 testing. Hasil dari perhitungan *accuracy, precision*, dan *recall* adapada *Gambar 6, Gambar 7, Gambar 8* dan *tabel 3, tabel 6,* dan hasil run code ada pada *Gambar 3*

4. Skenario Pengujian 4

Pada skenario pengujian 4, akan menggunakan data yang terdapat di folder 4 yang berisi 60 data. Data tersebut akan dibagi menjadi 30 training dan 30 testing. Hasil dari perhitungan *accuracy, precision*, dan *recall* adapada *Gambar 6, Gambar 7, Gambar 8* dan *tabel 4, tabel 6,* dan hasil run code ada pada *Gambar 4*

5. Skenario Pengujian 5

Pada skenario pengujian 5, akan menggunakan data yang terdapat di folder 5 yang berisi 70 data. Data tersebut akan dibagi menjadi 35 training dan 35 testing. Hasil dari perhitungan *accuracy, precision*, dan *recall* adapada *Gambar 6, Gambar 7, Gambar 8* dan *tabel 5, tabel 6, ,* dan hasil run code ada pada *Gambar 5*

V. KESIMPULAN

A. Kesimpulan

Kesimpulan yang dapat diambil pada penelitian Tugas Akhir ini adalah sebagai berikut:

1. Implementasi PCA pada dokumen artikel berita yang mendeteksi adanya penyimpangan / outlierhanya mendeteksi sebegian kecil dari o*utlier* yang ada.
2. Hasil dari sistem deteksi *outlier* menggunakan PCA menunjukkan bahwa PCA bisa mendeteksi outlier data text tetapi dengan akurasi yang kurang baik..

   Karena PCA disini mendeteksi outlier berdasarkan pada nilai eigen terbesar, jika menggunakan data text harus juga memperhitungkan setiap kata yang terdapat pada dokumen, apakah dokumen tersebutsama dengan dokumen lain yang berisi informasi mengenai *tech* atau tidak.

B. Saran

Saran yang diperlukan untuk perbaikan dan pengembangan untuk penelitian terkait selanjutnya adalahsebagai berikut:

1. Pengambilan data uji bisa di dapatkan di situs berita lainnya, baik itu situs berita luar negeri ataupunberita dalam negeri.
2. Pengambilan data uji bisa dalam bentuk lain tidak harus data text
3. Peneliti dapat menggunakan metode atau data yang berbeda, dan
4. Pastikan data yang dimiliki sama dengan data pada contoh code agar bisa mendapatkan hasil yangmaksimal

REFERENSI

[1] J. Han and M. Kamber, *Data Mining : Concepts and Technique*. 2006

[2] Han. Jiawei, Kamber. Micheline, 2006. *Data Mining : Concepts and Tehniques*, MorganKaufmann.

[3] Dr. S Vijayarani and Ilamathi J, '*Preprocessing Techniques for Text Mining - An Overview -Semantic Scholar'*, 2015.

[4] V. Gurusamy and S. Kannan, *'Preprocessing Techniques for Text Mining'*, 2014.

[5] B. A. Kuncoro and B. H. Iswanto, 'TF-IDF method in ranking keywords of Instagram users' image captions', in *2015 International Conference on Information Technology Systems and Innovation (ICITSI),* 2015, pp. 1–5.

[6] 'Confusion Matrix'. [Online]. Available: http://researchhubs.com/post/ai/fundamentals/confusion-matrix.html. [Accessed: 01-Jan-2020].

[7] J. Han and M. Kamber, *Data Mining : Concepts and Technique*. 2006.

[8] R. Kannan, H.Woo, C. Aggarwal, and H. Park. *Outlier detection for text data : An extended version*. CoRR, abs/1701.01325, 2017.

[9] E. J. Candés, X. Li, Y. Ma, J. *Wright. Robust principal component analysis?* Journal of theACM v.58 n.11 May 2011

[10] Wright, J., Ganesh, A., Rao, S., Ma, Y.: *Robust principal component analysis: Exact recoveryof corrupted low-rank matrices via convex optimization.* submitted to Journal of the ACM (2009)

[11] Z. Lin, M. Chen, Y. Ma. *The Augmented Lagrange Multiplier Method for Exact Recovery ofCorrupted Low-Rank Matrices*,

arXiv:1009.5055
[12] 'Dataset'. Available : http://mlg.ucd.ie/datasets/bbc.html.