

# **SKRIPSI**

## **PEMODELAN KATEGORISASI PRODUK *APPAREL* PADA *MARKETPLACE* DENGAN *MACHINE LEARNING***

Oleh:

Mega Inez Syafira

065116051



**PROGRAM STUDI ILMU KOMPUTER  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS PAKUAN  
BOGOR  
2021**

# **SKRIPSI**

## **PEMODELAN KATEGORISASI PRODUK *APPAREL* PADA *MARKETPLACE* DENGAN *MACHINE LEARNING***

Diajukan sebagai salah satu syarat untuk memperoleh  
Gelar Sarjana Komputer Jurusan Ilmu Komputer  
Fakultas Matematika dan Ilmu Pengetahuan Alam

Oleh:

Mega Inez Syafira

065116051



**PROGRAM STUDI ILMU KOMPUTER  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS PAKUAN  
BOGOR  
2021**

## HALAMAN PENGESAHAN

Judul : Pemodelan Kategorisasi Produk *Apparel* Pada *Marketplace* Dengan  
*Machine Learning*  
Nama : Mega Inez Syafira  
NPM : 065116051

Mengesahkan,

Pembimbing Pendamping  
FMIPA UNPAK



Sufiatul Maryana, S.Kom., M.Kom.

Pembimbing Utama  
FMIPA UNPAK



Dr. Prihastuti Harsani, S.Si., M.Si.

Mengetahui,

Ketua Program Studi Ilmu Komputer  
FMIPA UNPAK



Lita Karlitasari, S.Kom., MMSI.

Dekan

FMIPA UNPAK  


Asep Denih, S.Kom., M.Sc., Ph.D.

## PERNYATAAN KEASLIAN KARYA TULIS SKRIPSI

Dengan ini saya menyatakan bahwa :

Sejauh yang saya ketahui, karya tulis ini bukan merupakan karya tulis yang pernah dipublikasikan atau sudah pernah digunakan untuk mendapatkan gelar Sarjana di Universitas lain, kecuali pada bagian-bagian dimana sumber informasinya dicantumkan dengan cara referensi yang semestinya.

Demikian pernyataan ini saya buat dengan sebenar-benarnya. Apabila kelak di kemudian hari terdapat gugatan, penulis bersedia dikenakan sanksi sesuai dengan peraturan yang berlaku.

Bogor, 03 Agustus 2021



Mega Inez Syafira

065116051

## **PERNYATAAN PELIMPAHAN SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA**

---

Saya yang bertandatangan di bawah ini :

Nama : Mega Inez Syafira

NPM : 065116051

Judul Skripsi : *Pemodelan Kategorisasi Produk Apparel Pada Marketplace  
Dengan Machine Learning*

Dengan ini saya menyatakan bahwa Paten dan Hak Cipta dari produk Skripsi dan Tugas Akhir di atas adalah benar karya saya, dengan arahan dari komisi pembimbing dan belum diajukan dalam bentuk apapun kepada perguruan tinggi manapun.

Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir Skripsi ini.

Dengan ini saya melimpahkan Paten, Hak Cipta dari karya tulis saya kepada Universitas Pakuan.

Bogor, 03 Agustus 2021



Mega Inez Syafira  
065116051

## RIWAYAT HIDUP



Mega Inez Syafira dilahirkan di Bogor pada tanggal 07 Oktober 1998 dari pasangan Bapak Gatot Chrismiyono dan Ibu Merry Edly Rachmawati K. sebagai anak pertama dari tiga bersaudara. Pada tahun 2010 penulis menyelesaikan pendidikan Sekolah Dasar di SD Negeri Pengadilan 5 Kota Bogor, kemudian pada tahun 2013 penulis menyelesaikan pendidikan Sekolah Menengah Pertama di SMP Negeri 12 Kota Bogor, dan tahun 2016 penulis menyelesaikan pendidikan Sekolah Menengah Atas di SMA Negeri 10 Kota Bogor. Pada tahun 2016, penulis meneruskan pendidikan ke jenjang perguruan tinggi di Universitas Pakuan, Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam. Selama berkuliah di Universitas Pakuan, penulis pernah aktif sebagai anggota di Himpunan Mahasiswa Ilmu Komputer (HIMAKOM) periode 2016 – 2020, dan sebagai asisten praktikum di Laboratorium Komputer (LABKOM) periode 2017 – 2021. Penulis juga ikut berpartisipasi sebagai peserta pada program *Student Mobility* yang diselenggarakan oleh Program Studi Ilmu Komputer Universitas Pakuan bekerjasama dengan Universiti Tun Hussein Onn Malaysia (UTHM) pada tahun 2018, dan berpartisipasi sebagai perwakilan mahasiswa pada acara Pemilihan Mahasiswa Berprestasi tingkat Fakultas Matematika dan Ilmu Pengetahuan Alam dan tingkat Universitas Pakuan pada tahun 2018. Pada tahun 2019, penulis melaksanakan Praktek Lapang (PL) di Bahagian ICT, Fakulti Sains Komputer & Teknologi Maklumat, Universiti Tun Hussein Onn Malaysia (UTHM). Pada tahun 2021, penulis menyelesaikan penelitian dengan judul *Pemodelan Kategorisasi Produk Apparel Pada Marketplace Dengan Machine Learning*.

## RINGKASAN

**Mega Inez Syafira 2021.** *Pemodelan Kategorisasi Produk Apparel Pada Marketplace Dengan Machine Learning* dibawah bimbingan **Dr. Prihastuti Harsani, S.Si., M.Si.** dan **Sufiatul Maryana, S.Kom., M.Kom.**

Kategorisasi produk pada *e-commerce* dianggap sebagai salah satu hal penting dalam kesuksesan penjualan suatu produk. Kategorisasi produk dapat membantu para penjual dalam mendiferensiasi produk yang akan dijual yang mungkin memiliki kemiripan jenis antar satu produk dengan produk yang lain, atau dengan produk yang dijual oleh penjual lainnya. Selain itu, kategorisasi produk juga dapat membantu para penjual untuk memahami jenis produk yang dijual dan selanjutnya dapat memudahkan pembuatan dan pelaksanaan strategi pemasaran produk. Apabila terjadi mis-kategorisasi pada suatu produk seperti salah mengelompokkan produk yang sesuai dengan jenisnya, maka akan menyulitkan pembeli yang sedang mencari produk tersebut karena kategorisasi produk yang tidak sesuai dengan jenis produknya, sehingga mengakibatkan penjualan pada produk tersebut tidak akan maksimal. Agar produk yang ingin dijual dapat dengan mudah diakses oleh calon pembeli, maka penjual harus menyusun produknya ke berbagai kategori yang sesuai dengan spesifikasi produk tersebut.

Tujuan dari penelitian ini adalah untuk membuat pemodelan kategorisasi produk *apparel* pada *marketplace* dengan *machine learning* menggunakan pendekatan *supervised learning* dan *unsupervised learning*. Penelitian ini menggunakan metode *Latent Dirichlet Allocation* (LDA) dan *Naïve Bayes* untuk dilakukan komparasi antar kedua metode dan mengetahui hasil yang paling sesuai untuk dilakukan kategorisasi produk *apparel*. Penelitian ini dibatasi dengan hanya menggunakan 3 (tiga) jenis produk yaitu produk *fashion* dengan kategori *tees*, *jeans*, dan *jacket* yang terdapat pada *website e-commerce Aliexpress.com*.

## KATA PENGANTAR

*Assalamu 'alaikum warahmatullahi wabarakatuh.*

Puji dan syukur penulis ucapkan kepada Allah SWT. atas segala rahmat dan karunia-Nya, dan tidak lupa *shalawat* serta salam tercurahkan kepada Nabi Muhammad SAW., keluarganya beserta para sahabatnya, hingga sampai kepada kita para umatnya.

Penulis telah menyelesaikan Laporan Penelitian dengan judul “**Pemodelan Kategorisasi Produk Apparel Pada Marketplace Dengan Machine Learning**”. Laporan Penelitian ini disusun sebagai salah satu syarat kelulusan di Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Pakuan.

Dalam penyusunan Laporan Penelitian ini, penulis banyak mendapat bantuan dari berbagai pihak, oleh sebab itu penulis ingin mengungkapkan rasa terima kasih kepada :

1. Dr. Prihastuti Harsani, S.Si., M.Si. selaku Pembimbing Utama yang selalu berkenan membantu dalam pengarahan dan bimbingan selama penyusunan Laporan Penelitian ini.
2. Sufiatul Maryana, S.Kom., M.Kom. selaku Pembimbing Pendamping yang selalu berkenan membantu dalam pengarahan dan bimbingan selama penyusunan Laporan Penelitian ini.
3. Lita Karlitasari, S.Kom., MMSI. selaku Ketua Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Pakuan.
4. Kedua orangtua dan keluarga tercinta yang selalu memberikan semangat, do'a, dan dukungan baik moril maupun materil.
5. Rekan-rekan mahasiswa yang telah memberikan semangat, motivasi dan inspirasi dalam penyusunan Laporan Penelitian ini.

Penulis menyadari keterbatasan waktu dan kemampuan dalam penulisan Laporan Penelitian ini masih jauh dari sempurna. Oleh karena itu, segala kritik dan saran yang membangun sangatlah penulis harapkan demi penyempurnaan penyusunan Laporan Penelitian ini agar lebih baik lagi. Semoga Laporan Penelitian ini dapat bermanfaat bagi para pembaca.

*Wassalamu 'alaikum warahmatullahi wabarakatuh.*

Bogor, 03 Agustus 2021

Mega Inez Syafira  
065116051



## DAFTAR ISI

<b>HALAMAN PENGESAHAN.....</b>	<b>i</b>
<b>PERNYATAAN KEASLIAN KARYA TULIS SKRIPSI.....</b>	<b>ii</b>
<b>PERNYATAAN PELIMPAHAN SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA .....</b>	<b>iii</b>
<b>RIWAYAT HIDUP .....</b>	<b>iv</b>
<b>RINGKASAN .....</b>	<b>v</b>
<b>KATA PENGANTAR .....</b>	<b>vi</b>
<b>DAFTAR ISI .....</b>	<b>vii</b>
<b>DAFTAR GAMBAR .....</b>	<b>x</b>
<b>DAFTAR TABEL .....</b>	<b>xi</b>
<b>DAFTAR LAMPIRAN .....</b>	<b>xii</b>
<b>BAB I PENDAHULUAN.....</b>	<b>1</b>
1.1 Latar Belakang.....	1
1.2 Tujuan .....	2
1.3 Ruang Lingkup .....	3
1.4 Manfaat .....	3
<b>BAB II TINJAUAN PUSTAKA .....</b>	<b>4</b>
2.1 <i>E-Commerce</i> .....	4
2.2 <i>Machine Learning</i> .....	4
2.3 <i>Supervised Learning dan Unsupervised Learning</i> .....	4
2.4 <i>Topic Modelling</i> .....	4
2.5 Klasifikasi .....	4
2.6 <i>Latent Dirichlet Allocation (LDA)</i> .....	5
2.7 <i>Naïve Bayes Classifier</i> .....	6
2.8 <i>Text Mining</i> .....	6
2.9 <i>Silhouette Analysis</i> .....	6
2.10 Tahap Evaluasi.....	6
2.11 Penelitian Terdahulu.....	7
2.12 Perbandingan Penelitian Terdahulu .....	8
<b>BAB III METODE PENELITIAN .....</b>	<b>9</b>
3.1 Metode Penelitian .....	9
3.1.1 Pengumpulan Data.....	9
3.1.2 Seleksi dan Pembersihan Data.....	9

3.1.3	Transformasi Data.....	9
3.1.4	Pemrosesan Data.....	9
3.1.5	Penggalian Data .....	10
3.1.6	Evaluasi Pola.....	10
3.1.7	Representasi Pengetahuan.....	10
3.2	Gambaran Umum Penelitian.....	10
3.3	Alat dan Bahan.....	11
3.3.1	Alat.....	11
3.3.2	Bahan .....	11
<b>BAB IV PERANCANGAN DAN IMPLEMENTASI .....</b>		<b>12</b>
4.1	Tahap Pelaksanaan Penelitian.....	12
4.2	Tahap Pengumpulan Data .....	12
4.3	Tahap Seleksi dan Pembersihan Data .....	12
4.4	Tahap Transformasi Data.....	12
4.5	Tahap Pemrosesan Data ( <i>Preprocessing</i> ).....	13
4.5.1	<i>Case Folding</i> .....	13
4.5.2	<i>Tokenizing</i> .....	13
4.5.3	<i>Stopword Removal</i> .....	13
4.5.4	<i>Stemming</i> .....	14
4.5.5	<i>N-gram</i> .....	14
4.5.6	<i>Term Frequency-Inverse Document Frequency (TF-IDF)</i> .....	14
4.6	Tahap Penggalian Data ( <i>Data Mining</i> ).....	15
4.6.1	<i>Latent Dirichlet Allocation (LDA)</i> .....	15
4.6.2	<i>Naïve Bayes Classifier</i> .....	16
4.7	Evaluasi Pola ( <i>Pattern Evaluation</i> ) .....	18
4.8	Representasi Pengetahuan ( <i>Knowledge Representation</i> ).....	20
4.8.1	Tahap Perancangan.....	20
4.8.2	Tahap Implementasi.....	20
<b>BAB V HASIL DAN PEMBAHASAN .....</b>		<b>23</b>
5.1	Hasil.....	23
5.1.1	Kondisi <i>Dataset</i> .....	23
5.1.2	Hasil <i>Preprocessing</i> dan Transformasi Data .....	23

5.1.3	Hasil <i>Term Frequency-Inverse Document Frequency</i> (TF-IDF).....	24
5.1.4	Hasil <i>Topic Modelling</i> .....	24
5.1.5	Hasil Kategorisasi.....	26
5.1.6	Hasil <i>Silhouette Analysis</i> .....	26
5.1.7	Hasil Evaluasi .....	27
5.2	Pembahasan .....	28
<b>BAB VI KESIMPULAN</b> .....		33
6.1	Kesimpulan .....	33
6.2	Saran .....	34
<b>DAFTAR PUSTAKA</b> .....		35
<b>LAMPIRAN</b> .....		37

## DAFTAR GAMBAR

	Halaman
<b>Gambar 1.</b> Metode KDD .....	9
<b>Gambar 2.</b> Gambaran Umum Penelitian .....	11
<b>Gambar 3.</b> Hasil <i>Scraping</i> Data.....	12
<b>Gambar 4.</b> Contoh Data Spesifikasi Produk <i>Men's Jacket</i> .....	13
<b>Gambar 5.</b> Hasil <i>Case Folding</i> .....	13
<b>Gambar 6.</b> Hasil <i>Stopword Removal</i> .....	14
<b>Gambar 7.</b> Hasil <i>Stemming</i> .....	14
<b>Gambar 8.</b> Contoh Data <i>Training</i> .....	16
<b>Gambar 9.</b> Contoh Data <i>Testing</i> .....	16
<b>Gambar 10.</b> Proses <i>Scraping</i> .....	20
<b>Gambar 11.</b> Proses <i>Preprocessing</i> .....	21
<b>Gambar 12.</b> Proses <i>Data Mining</i> .....	21
<b>Gambar 13.</b> Proses Pemodelan Topik .....	22
<b>Gambar 14.</b> Proses Kategorisasi.....	22
<b>Gambar 15.</b> Sebelum <i>Preprocessing</i> dan Transformasi Data .....	23
<b>Gambar 16.</b> Sesudah <i>Preprocessing</i> dan Transformasi Data .....	24
<b>Gambar 17.</b> Hasil TF-IDF <i>unigram</i> .....	24
<b>Gambar 18.</b> Hasil TF-IDF <i>bigram</i> .....	24
<b>Gambar 19.</b> Hasil Visualisasi LDA (a) 4 topik <i>unigram</i> dan (b) 3 topik <i>bigram</i> ....	25
<b>Gambar 20.</b> Distribusi Kata LDA (a) 4 topik <i>unigram</i> dan (b) 3 topik <i>bigram</i> .....	25
<b>Gambar 21.</b> Hasil Kategorisasi.....	26
<b>Gambar 22.</b> Hasil <i>Silhouette Analysis</i> untuk data <i>unigram</i> .....	27
<b>Gambar 23.</b> Hasil <i>Silhouette Analysis</i> untuk data <i>bigram</i> .....	27
<b>Gambar 24.</b> Hasil <i>Confusion Matrix</i> (a) Kategori dan (b) Sub-kategori <i>Apparel</i> ....	27
<b>Gambar 25.</b> Hasil (a) Sebelum dan (b) Sesudah Transformasi Data .....	28
<b>Gambar 26.</b> Hasil Kategorisasi Kategori.....	30
<b>Gambar 27.</b> Hasil Kategorisasi Sub-kategori .....	30
<b>Gambar 28.</b> Hasil Kategorisasi Kategori Data Baru .....	31
<b>Gambar 29.</b> Hasil Kategorisasi Sub-kategori Data Baru.....	31

## DAFTAR TABEL

	Halaman
<b>Tabel 1.</b> Perbandingan Penelitian Terdahulu .....	8
<b>Tabel 2.</b> Hasil <i>tokenizing</i> .....	13
<b>Tabel 3.</b> Hasil <i>unigram</i> .....	14
<b>Tabel 4.</b> Hasil <i>bigram</i> .....	14
<b>Tabel 5.</b> Pemodelan Topik <i>unigram</i> .....	15
<b>Tabel 6.</b> Pemodelan Topik <i>bigram</i> .....	15
<b>Tabel 7.</b> Hasil Kategorisasi .....	18
<b>Tabel 8.</b> <i>Confusion Matrix</i> .....	19
<b>Tabel 9.</b> Hasil Kategorisasi Aktual dan Prediksi dengan <i>Naïve Bayes</i> .....	19
<b>Tabel 10.</b> <i>Confusion Matrix</i> untuk Hasil Kategorisasi.....	19
<b>Tabel 11.</b> Kondisi <i>Dataset</i> .....	23
<b>Tabel 12.</b> Interpretasi Topik LDA <i>unigram</i> .....	25
<b>Tabel 13.</b> Interpretasi Topik LDA <i>bigram</i> .....	26
<b>Tabel 14.</b> Nilai <i>accuracy, precision, recall</i> dan <i>F1-score</i> Kategorisasi.....	28

## DAFTAR LAMPIRAN

- Lampiran 1. Surat Keputusan (SK) Tugas Akhir
- Lampiran 2. Hasil Data *Scraping* Awal
- Lampiran 3. Hasil *Preprocessing* Data Manual
- Lampiran 4. Hasil *Preprocessing* Data *Rstudio*
- Lampiran 5. Hasil Pemodelan Topik *Unigram* dan *Bigram*
- Lampiran 6. Hasil Visualisasi Kategorisasi
- Lampiran 7. Hasil Kategorisasi *Actual* dan *Predicted*
- Lampiran 8. Hasil Evaluasi Kategori dan Sub-kategori
- Lampiran 9. Perbandingan Data Latih dan Data Uji
- Lampiran 10. *Stopword List* Bahasa Inggris
- Lampiran 11. Perhitungan Manual TF-IDF
- Lampiran 12. Analisis Kebutuhan Pengerjaan Sistem
- Lampiran 13. Tahap Perancangan Sistem

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

*E-commerce* adalah salah satu bagian dari pemanfaatan teknologi yang bergerak di bidang perdagangan, yang membantu konsumen dalam hal transaksi barang dan jasa melalui media elektronik. Akses yang cepat, murah, mudah, serta dapat melakukan interaksi dan perdagangan dengan orang di seluruh dunia menjadi daya tarik tersendiri. Peluang ini dapat dimanfaatkan oleh siapapun yang ingin bersaing dalam kegiatan jual-beli, khususnya para penjual yang ingin produknya dikenal dan terjual ke penjuru dunia berkat media promosi melalui *e-commerce*. (Aco *et al.*, 2017)

Kategorisasi produk pada *e-commerce* dianggap sebagai salah satu hal penting dalam kesuksesan penjualan suatu produk. Kategorisasi produk dapat membantu para penjual dalam mendiferensiasi produk yang akan dijual yang mungkin memiliki kemiripan jenis antar satu produk dengan produk yang lain, atau dengan produk yang dijual oleh penjual lainnya. Selain itu, kategorisasi produk juga dapat membantu para penjual untuk memahami jenis produk yang dijual dan selanjutnya dapat memudahkan pembuatan dan pelaksanaan strategi pemasaran produk. Apabila terjadi miskategorisasi pada suatu produk seperti salah mengelompokkan produk yang sesuai dengan jenisnya, maka akan menyulitkan pembeli yang sedang mencari produk tersebut karena kategorisasi produk yang tidak sesuai dengan jenis produknya, sehingga mengakibatkan penjualan pada produk tersebut tidak akan maksimal.

Agar produk yang ingin dijual dapat dengan mudah diakses oleh calon pembeli, maka penjual harus menyusun produknya ke berbagai kategori yang sesuai dengan spesifikasi produk tersebut. Kategorisasi dilakukan dengan cara memasukkan produk ke dalam daftar kategori. Kategorisasi seperti “*Women’s Fashion > Blouses & Shirts*”, adalah salah satu contoh pengelompokan yang berisikan produk khusus yang sesuai dengan nama kategori tersebut. Kategorisasi ini banyak diterapkan oleh *e-commerce* besar seperti *eBay*, *Aliexpress*, dan *Amazon*. *Aliexpress* adalah salah satu *e-commerce* raksasa di dunia yang dimiliki oleh *Alibaba Group* dan diluncurkan pada tahun 2010. *Aliexpress* lebih mengedepankan portal transaksi jual beli *Business-to-Business*, dengan memiliki 30 (tiga puluh) kategori dan lebih dari 150 (seratus lima puluh) sub-kategori.

Banyaknya jenis barang jadi yang dapat diperjualbelikan—salah satunya produk *apparel*—menjadi salah satu penyebab sulitnya menetapkan kategorisasi. Setiap produk *apparel* memiliki spesifikasi bahan dan kegunaan yang berbeda, tidak terlepas dari tren mode yang mengikuti perkembangan zaman. Tak hanya mengikuti tren mode yang sedang ramai diperbincangkan, tidak sedikit pula produk *apparel* yang dikhususkan untuk orang dengan profesi tertentu dipasarkan di *e-commerce*. Perbedaan spesifikasi, *material* dan fungsi yang terdapat pada setiap produk *apparel* mengakibatkan perlunya pendataan dan pengelompokan produk yang akan dijual sehingga dapat tersusun dengan baik. Pada saat penginputan spesifikasi produk, penjual juga harus melakukan proses pemilihan kategori produk secara manual dengan banyaknya kategori yang disediakan oleh *e-commerce*. Hal ini dapat menimbulkan kesalahan pemilihan kategori apabila penjual ingin menambahkan produk dalam

jumlah banyak. Selain itu, proses ini akan memakan waktu yang cukup lama, sehingga menjadi tidak efisien dan kategori data produk yang dihasilkan menjadi kurang akurat. Berdasarkan permasalahan tersebut, dibutuhkan adanya suatu alternatif kategorisasi pada produk *e-commerce* yang sesuai dengan nama dan spesifikasi produk yang akan ditambahkan, dengan menerapkan metode *machine learning* seperti *clustering* atau klasifikasi dalam proses kategorisasi yang dilakukan.

Penelitian terdahulu telah dilakukan oleh Daniel *et al.* (2014). Dalam penelitian tersebut dilakukan perancangan *expert system* untuk klasifikasi barang impor pada kantor Bea Cukai Tanjung Perak, Surabaya. Penelitian ini menggunakan pendekatan *forward chaining*, perancangan *prototype*, dan pembuatan *user interface* dirancang menggunakan bahasa pemrograman PHP, pengelolaan database dengan MySQL, dan menggunakan sistem operasi Linux. *Prototype* tersebut diujikan kepada 16 orang PFPD sebagai user. Hasil survei dari penelitian ini didapatkan bahwa responden PFPD menyatakan bahwa rancangan *expert system* tersebut dapat memenuhi kebutuhan penetapan tarif, membantu klasifikasi barang secara tepat dan cepat, dapat merasa dibantu oleh ahli klasifikasi, serta mudah dioperasikan.

Penelitian selanjutnya telah dilakukan oleh Cevahir *et al.* (2016). Dalam penelitian tersebut diusulkan alat klasifikasi otomatis untuk memprediksi kategori yang cocok dengan judul dan deskripsi produk yang diberikan dengan studi kasus pada *e-commerce* Rakuten Ichiba. Penelitian ini menggunakan metode jaringan saraf tiruan *Deep Belief Network* (DBN) dan *Deep Autoencoder* (DAE), kemudian diimplementasikan pada *Graphics Processing Unit* (GPU) dan *library* CUDA dengan C++. Pelatihan model menggunakan 150 juta produk dengan *taxonomy tree* sebanyak 5 (lima) *level*, dan memiliki 28.338 kategori. Hasil pengujian menunjukkan bahwa 81% prediksi kategori yang ditentukan sesuai dengan pemilihan kategori yang dilakukan oleh penjual.

Penelitian selanjutnya telah dilakukan oleh Gottipati *et al.* (2016). Dalam penelitian tersebut dilakukan kategorisasi produk yang sesuai dengan judul dan deskripsi produk pada *e-commerce website* Amazon. Penelitian ini menggunakan pemodelan topik *unsupervised* yaitu *Latent Dirichlet Allocation* (LDA) dan menggunakan 3 (tiga) *classifier* yaitu *Naïve Bayes*, *Support Vector Machine* (SVM), dan *K-Nearest Neighbors* (K-NN). Penelitian ini juga menggunakan ekstraksi fitur untuk pengolahan kata *unigram* dan *bigram* dengan *split* sebesar 50%. Proses pengolahan data diimplementasikan menggunakan *library Apache Lucene*, dan penghitungan *chi-square* menggunakan aplikasi WEKA. Hasil pengujian menunjukkan bahwa pada model *chi-square* memiliki nilai akurasi terbaik pada *unigram* dengan menggunakan *classifier Naïve Bayes*, diikuti oleh SVM dan K-NN.

Berdasarkan ulasan di atas, penelitian ini melakukan pemodelan kategorisasi produk *apparel* berdasarkan pendekatan *supervised learning* dan *unsupervised learning*. Metode yang akan dilakukan pada penelitian ini yaitu *Latent Dirichlet Allocation* (LDA) dan *Naïve Bayes*.

## 1.2 Tujuan

Tujuan dari penelitian ini adalah untuk membuat pemodelan kategorisasi produk *apparel* pada *marketplace* dengan *machine learning* menggunakan pendekatan *supervised learning* dan *unsupervised learning*.



### 1.3 Ruang Lingkup

Penelitian ini dilakukan untuk membuat model pemodelan topik dan kategorisasi otomatis pada produk *apparel*, dengan jenis produk yaitu produk *fashion* yang terdapat pada *website e-commerce* Aliexpress.com. Data yang akan digunakan diambil sebanyak 960 data pada kategori *Men's Jacket*, 900 data pada kategori *Women's Jacket*, 960 data pada kategori *Men's Jeans*, 960 data pada kategori *Women's Jeans*, 960 data pada kategori *Men's Tees*, dan 960 data pada kategori *Women's Tees*, sehingga total data yang akan digunakan sebanyak 5.700 data. Penelitian ini menggunakan seleksi fitur dengan TF-IDF untuk *unigram* dan *bigram*, pemodelan topik dengan *Latent Dirichlet Allocation* (LDA) untuk *unigram* dan *bigram*, dan kategorisasi dengan *Naïve-Bayes classifier*. Hasil dari kedua metode ini akan dilihat nilai keakuratan dan kemiripannya dengan kategorisasi yang terdapat pada *e-commerce* Aliexpress.com. Bahasa pemrograman yang akan digunakan dalam penelitian ini yaitu R dengan menggunakan aplikasi *Rstudio*.

### 1.4 Manfaat

Manfaat dari penelitian ini antara lain sebagai berikut:

1. Melakukan penerapan metode *Latent Dirichlet Allocation* (LDA) dalam pemodelan topik secara *unigram* dan *bigram*.
2. Melakukan penerapan metode *Naïve Bayes Classifier* dalam pengklasifikasian data.
3. Melakukan kategorisasi produk *apparel* sesuai dengan judul dan spesifikasi produk.
4. Mengetahui nilai keakuratan dari hasil kategorisasi produk *apparel*.
5. Melakukan komparasi antara hasil pengelompokan dengan pendekatan *supervised learning* dan *unsupervised learning*.
6. Penelitian ditujukan ke web *e-commerce* dalam hal ini *aliexpress.com*, dan user yang dituju yaitu penjual yang akan memilih kategori barang jualannya di *website aliexpress.com*.

## **BAB II**

### **TINJAUAN PUSTAKA**

#### **2.1 E-Commerce**

*E-commerce* adalah sebuah layanan *online* yang bergerak di bidang perdagangan dan dapat digunakan sebagai tempat transaksi jual beli melalui media elektronik. Kemudahan dalam pembayaran dan hak akses membuat platform ini banyak digunakan oleh masyarakat Indonesia. Proses jual-beli yang dilakukan mulai dari pencarian barang yang ingin dibeli, pemesanan barang, proses pembayaran dilakukan secara elektronik. Pemanfaatan *e-commerce* dapat dilakukan dengan adanya transaksi kepada konsumen (*Business-to-Customer*) maupun transaksi kepada sesama pegiat bisnis (*Business-to-Business*). Untuk dapat mengakses *e-commerce*, pengguna *e-commerce* baik penjual maupun pembeli membutuhkan akses internet. (Shaik *et al.*, 2019)

#### **2.2 Machine Learning**

*Machine Learning* adalah salah satu varian dari sistem kecerdasan buatan yang memungkinkan komputer dapat belajar tanpa diprogram secara eksplisit, sehingga computer atau mesin tersebut akan memiliki suatu kecerdasan dari hasil belajar yang dilakukan. *Machine Learning* dapat membantu mengenali, mengidentifikasi, atau memprediksi data tertentu dengan mengekstrak pengetahuan dari data yang belum diketahui hubungannya dan mempelajari dengan algoritma tertentu. (Nurhayati *et al.*, 2019)

#### **2.3 Supervised Learning dan Unsupervised Learning**

*Supervised Learning* dan *Unsupervised Learning* adalah salah satu algoritma dari *machine learning* dalam mengenali dan memproses suatu data hingga diperoleh suatu pengetahuan. *Supervised Learning* adalah salah satu algoritma yang menggunakan *training dataset* untuk membuat sebuah prediksi dari kumpulan data. *Unsupervised Learning* adalah salah satu algoritma yang digunakan untuk memperoleh pengetahuan dari *dataset* tanpa label. (Nurhayati *et al.*, 2019)

#### **2.4 Topic Modelling**

*Topic Modelling* adalah sebuah topik terdiri dari kata-kata tertentu yang Menyusun topik tersebut, dan dalam satu dokumen memiliki kemungkinan terdiri dari beberapa topik dengan probabilitas masing-masing. Namun secara pemahaman manusia, dokumen-dokumen merupakan objek yang dapat diamati, sedangkan topik-topik merupakan yang terdapat dalam dokumen tersebut merupakan salah satu struktur yang tersembunyi. Dalam hal ini, *topic modelling* bertujuan untuk menemukan topik dari kata-kata yang terdapat pada dokumen tersebut. (Putra *et al.*, 2017)

#### **2.5 Klasifikasi**

Klasifikasi merupakan suatu proses yang bertujuan untuk menentukan suatu obyek kedalam suatu kelas atau kategori yang sudah ditentukan sebelumnya. Klasifikasi data ataupun dokumen juga dapat dimulai dari membangun aturan klasifikasi tertentu yang menggunakan *data training* yang sering disebut sebagai

tahapan pembelajaran dan pengujian digunakan sebagai *data testing*. (Raharjo *et al.*, 2014).

Beberapa tugas dari klasifikasi yang melibatkan proses pembangunan terhadap model yang dibentuk untuk melakukan prediksi target atau variabel dari dataset yang sudah jelas, ataupun variabel independent. (Puspitasari *et al.*, 2018)

## 2.6 Latent Dirichlet Allocation (LDA)

*Latent Dirichlet Allocation* (LDA) adalah salah satu *unsupervised method* yang tidak membutuhkan pembelajaran sebelumnya sebagai acuan kebenarannya. Metode ini merupakan salah satu model untuk pemodelan topik. Tujuan dari metode ini adalah untuk menyimpulkan informasi utama yang tersembunyi dalam suatu dokumen. (Fitrianto, 2017)

*Latent Dirichlet Allocation* (LDA) mengasumsikan setiap dokumen memiliki topik dan dibentuk dari kata-kata yang berkaitan dengan topik tersebut, sehingga suatu dokumen dapat direpresentasikan sebagai campuran dari topik-topik tersembunyi dengan proporsi yang berbeda-beda. Teknik *Gibbs Sampling* digunakan dalam pengaplikasian LDA, dengan tahapan sebagai berikut:

1. Tentukan jumlah topik dan jumlah iterasi.
2. Untuk setiap kata yang ada dalam suatu *tweet*, masukan kata tersebut ke dalam suatu topik secara acak.
3. Pilih satu *tweet*,
4. Pilih satu kata dalam *tweet*,
5. Hitung nilai probabilitas kata tersebut terhadap setiap topik yang ada dengan menggunakan persamaan (1),
6. Masukkan kata ke dalam topik yang memiliki nilai tertinggi,
7. Lakukan tahap 5 dan 6 untuk setiap kata dalam *tweet* hingga seluruh *tweet* telah terproses,
8. Ulangi tahap 4 sampai dengan tahap 8 sebanyak iterasi yang ditentukan.

$$P(j|w_i, d_i) = \frac{C_{w_i j}^{WT}}{\sum_{w=1}^W C_{w j}^{WT}} \cdot \frac{C_{d_i j}^{DT}}{\sum_{t=1}^T C_{d_i t}^{DT}}$$

Dengan keterangan sebagai berikut:

$j$  = topik yang sedang dikalkulasi

$w_i$  = kata yang sedang dikalkulasi

$d_i$  = dokumen yang sedang dikalkulasi

$C^{WT}$  = matriks jumlah kata dalam suatu dokumen

$C^{DT}$  = matriks jumlah topik dalam suatu dokumen

$C_{w_i j}^{WT}$  = jumlah kata  $w_i$  dalam topik  $j$

$C_{w j}^{WT}$  = jumlah kata  $w$  dalam topik  $j$

$C_{d_i j}^{DT}$  = jumlah kata dalam  $d_i$  termasuk topik  $j$

$C_{d_i t}^{DT}$  = jumlah kata dalam  $d_i$  termasuk topik  $t$

Dengan menggunakan persamaan (1), probabilitas suatu dokumen masuk kedalam suatu topik dihitung dengan melihat jumlah topik pada suatu dokumen. Probabilitas suatu kata termasuk ke dalam topik apa, dihitung dengan melihat jumlah kata pada suatu topik. (Al-khairi *et al.*, 2018)

## 2.7 Naïve Bayes Classifier

*Naïve Bayes classifier* adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class*. *Naïve Bayes* sebagai *classifier* memiliki kemampuan klasifikasi serupa dengan *decision tree* dan *neural network*. Metode *Bayes* merupakan pendekatan statistik untuk melakukan inferensi induksi pada persoalan klasifikasi. Teorema *Bayes* memiliki bentuk umum seperti pada persamaan (5) berikut. (Annur, 2018)

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (5)$$

Keterangan:

X	= Data dengan <i>class</i> yang belum diketahui
H	= Hipotesis data X merupakan suatu <i>class</i> spesifik
P(H X)	= Probabilitas hipotesis H berdasarkan kondisi X
P(H)	= Probabilitas hipotesis H
P(X H)	= Probabilitas X berdasarkan kondisi tersebut
P(X)	= Probabilitas dari X

## 2.8 Text Mining

*Text Mining* merupakan suatu proses pengolahan dan pengambilan informasi dari data tekstual yang memiliki format tidak teratur. *Text Mining* adalah penemuan dari pengetahuan yang menarik pada dokumen teks. (Taufik, 2018)

Alur *text mining* terdiri atas pengumpulan teks, *text pre-processing*, *text transformation*, *feature selection*, *data mining*, dan interpretasi. *Text pre-processing* terdiri dari *case folding*, *non alpha numeric removal*, *stop words removal*, dan *stemming*. Tahap *case folding* bertujuan untuk mengubah bentuk teks kedalam bentuk huruf kecil. *Non alpha numeric removal* bertujuan untuk menghilangkan karakter selain karakter alfanumerik pada teks. *Stop words removal* adalah suatu proses penghilangan kata-kata yang bukan sebagai ciri dari kata tersebut seperti penggunaan imbuhan atau kata kepemilikan. *Stemming* adalah suatu proses pengembalian kata kedalam bentuk dasarnya tanpa memiliki imbuhan apapun. (Siringoringo *et al.*, 2019)

## 2.9 Silhouette Analysis

*Silhouette Analysis* adalah salah satu pengujian dalam proses *clustering* yang dapat digunakan untuk mempelajari jarak yang dihasilkan antar *cluster*. *Silhouette plot* dapat menampilkan sejumlah nilai *cluster* yang optimal, diambil dari nilai rata-rata dari nilai kedekatan setiap titik pada satu *cluster* dengan titik pada *cluster* lainnya, sehingga dapat membantu memperoleh nilai K yang dapat digunakan untuk proses *clustering* di tahap selanjutnya. (Herwanto, 2018)

## 2.10 Tahap Evaluasi

Tahap evaluasi merupakan proses untuk menguji hasil komparasi dari metode LDA dan *Naïve Bayes* yang dilakukan. Metode evaluasi yang digunakan adalah *Accuracy*, *Precision*, *Recall*, dan *F1-score*.

*Accuracy* merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data, seperti pada persamaan (1) berikut.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (1)$$

*Precision* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif, seperti pada persamaan (2) berikut.

$$Precision = \frac{(TP)}{(TP+FP)} \quad (2)$$

*Recall* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif, seperti pada persamaan (3) berikut.

$$Recall = \frac{(TP)}{(TP+FN)} \quad (3)$$

*F1-score* merupakan perbandingan rata-rata presisi dan recall yang dibobotkan, seperti pada persamaan (4) berikut. (Waasiu *et al.*, 2021)

$$F1 - score = \frac{2 * (Recall * Precision)}{(Recall + Precision)} \quad (4)$$

## 2.11 Penelitian Terdahulu

1. Nama : Debby Ratna Daniel, Ahmad Saifuddin Zuhri (2013)  
 Judul : Perancangan *Expert System* Untuk Klasifikasi Barang Pada Kantor Bea Cukai Tanjung Perak Surabaya  
 Isi : Penelitian ini merancang sebuah *expert system* untuk klasifikasi barang impor pada kantor Bea Cukai Tanjung Perak, Surabaya. Penelitian ini menggunakan pendekatan *forward chaining*, perancangan *prototype*, dan pembuatan *user interface* dirancang menggunakan bahasa pemrograman PHP, pengelolaan database dengan MySQL, dan menggunakan sistem operasi Linux. *Prototype* tersebut diujikan kepada 16 orang PFPD sebagai user. Hasil survei dari penelitian ini didapatkan bahwa responden PFPD menyatakan bahwa rancangan *expert system* tersebut dapat memenuhi kebutuhan penetapan tarif, membantu klasifikasi barang secara tepat dan cepat, dapat merasa dibantu oleh ahli klasifikasi, serta mudah dioperasikan.
2. Nama : Ali Cevahir, Koji Murakami (2016)  
 Judul : *Large-scale Multi-class and Hierarchical Product Categorization for an E-commerce Giant*  
 Isi : Penelitian ini mengusulkan alat klasifikasi otomatis untuk memprediksi kategori yang cocok dengan judul dan deskripsi produk yang diberikan dengan studi kasus pada *e-commerce* Rakuten Ichiba. Penelitian ini menggunakan metode jaringan saraf tiruan *Deep Belief Network* (DBN) dan *Deep Autoencoder* (DAE), kemudian diimplementasikan pada *Graphics Processing Unit* (GPU) dan *library* CUDA dengan C++. Pelatihan model menggunakan 150 juta produk dengan *taxonomy tree* sebanyak 5 (lima) *level*, dan memiliki 28.338 kategori. Hasil pengujian menunjukkan bahwa 81%

prediksi kategori yang ditentukan sesuai dengan pemilihan kategori yang dilakukan oleh penjual.

3. Nama : Srinivasu Gottipati, Mumtaz Vauhkonen (2016)

Judul : *E-Commerce Product Categorization*

Isi : Penelitian ini melakukan kategorisasi produk yang sesuai dengan judul dan deskripsi produk pada *e-commerce website* Amazon. Penelitian ini menggunakan pemodelan topik *unsupervised* yaitu *Latent Dirichlet Allocation* (LDA) dan menggunakan 3 (tiga) *classifier* yaitu *Naïve Bayes*, *Support Vector Machine* (SVM), dan *K-Nearest Neighbors* (K-NN). Penelitian ini juga menggunakan ekstraksi fitur untuk pengolahan kata *unigram* dan *bigram* dengan *split* sebesar 50%. Proses pengolahan data diimplementasikan menggunakan *library Apache Lucene*, dan penghitungan *chi-square* menggunakan aplikasi WEKA. Hasil pengujian menunjukkan bahwa pada model *chi-square* memiliki nilai akurasi terbaik pada *unigram* dengan menggunakan *classifier Naïve Bayes*, diikuti oleh SVM dan K-NN.

## 2.12 Perbandingan Penelitian Terdahulu

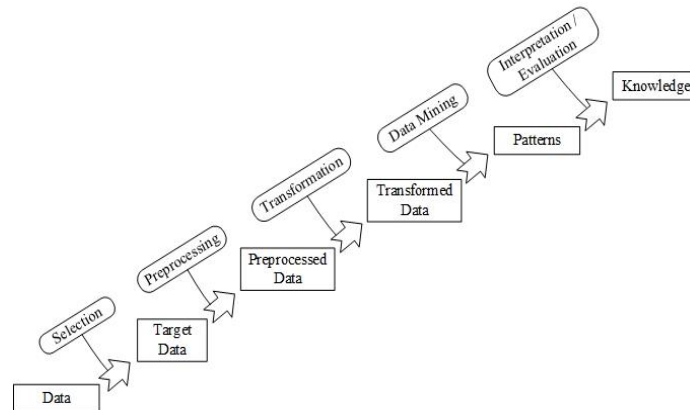
**Tabel 1.** Perbandingan Penelitian Terdahulu

Nama dan Tahun	Objek Penelitian	Pendekatan		Bahasa Pemrograman			Aplikasi		Metode		
		<i>Rule-based</i>	<i>Machine Learning</i>	<i>P H P</i>	<i>C++</i>	<i>R</i>	WEKA	Rstudio	<i>Forward Chaining</i>	<i>DBN</i>	<i>LDA</i>
Daniel <i>et al</i> (2013)	Barang Impor Kantor Bea Cukai	√	-	√	-	-	-	-	√	-	-
Cevahir <i>et al</i> (2016)	Rakuten Ichiba	-	√	-	√	-	-	-	-	√	-
Gottipati <i>et al</i> (2016)	Amazon	-	√	-	-	-	√	-	-	-	√
Mega Inez Syafira (2021)	Aliexpress	-	√	-	-	√	-	√	-	-	√

## BAB III METODE PENELITIAN

### 3.1 Metode Penelitian

Metode penelitian yang digunakan dalam penelitian ini menggunakan model *Knowledge Discovery and Data Mining* (KDD). Model ini digunakan karena mempunyai tahapan yang lengkap dan terstruktur. *Flowchart* yang digunakan dalam sistem ini ditampilkan dalam Gambar 1, sebagai berikut :



**Gambar 1.** Metode KDD

#### 3.1.1 Pengumpulan Data

Tahap pengumpulan data adalah salah satu tahap pengambilan data ataupun dokumen sebagai bahan awal penelitian yang akan dianalisis atau digunakan untuk mendapatkan hasil yang dibutuhkan.

#### 3.1.2 Seleksi dan Pembersihan Data

Tahap seleksi dan pembersihan data diperlukan untuk memilih data yang penting dan sesuai yang dibutuhkan dalam penelitian, dan menghapus data yang bersifat *noise* dan tidak penting yang kemungkinan dapat mempengaruhi keakuratan hasil penelitian.

#### 3.1.3 Transformasi Data

Tahap transformasi data dilakukan untuk mengubah data yang sudah diseleksi sebelumnya ke dalam sebuah bentuk atau format yang diperlukan untuk tahap pengolahan data.

#### 3.1.4 Pemrosesan Data

Tahap pemrosesan data bertujuan untuk melakukan pembersihan data secara menyeluruh, seperti mengurangi dan mengubah data yang inkonsisten serta melakukan normalisasi menyeluruh pada data, sehingga menghasilkan data bersih yang dapat diolah pada tahap *data mining*.

##### 3.1.4.1 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF sebagai metode pemberian bobot dan ekstraksi fitur yang dapat digunakan untuk mengambil fitur yang penting dan membuang fitur yang tidak diperlukan, untuk memudahkan proses pengolahan data di tahap selanjutnya.

### 3.1.5 Penggalian Data

Tahap penggalian data adalah tahap pengolahan data dengan tujuan untuk mengetahui karakteristik dari data yang diolah dan menemukan pola tertentu dari suatu data, untuk diterapkan kepada fungsi *data mining* yang diperlukan.

#### 3.1.5.1 Pemodelan Topik dengan *Latent Dirichlet Allocation* (LDA)

Tahap pemodelan topik dengan menggunakan metode *Latent Dirichlet Allocation* (LDA) dilakukan untuk membentuk dan memperoleh sejumlah topik berisi distribusi kata acak dari kumpulan data atau dokumen, dengan tujuan untuk mengetahui apa saja topik utama dan deskripsi topik dari kumpulan data tersebut.

#### 3.1.5.2 Klasifikasi dengan *Naïve Bayes* (NB)

Tahap klasifikasi dengan menggunakan algoritma *Naïve Bayes* (NB) dilakukan untuk memprediksi, mengelompokkan dan menentukan kategori pada data, dengan menggunakan perhitungan nilai probabilitas terhadap variabel dalam sebuah kelas sebagai penentu klasifikasi.

### 3.1.6 Evaluasi Pola

Tahap ini dilakukan untuk mengevaluasi keluaran yang ditemukan berupa pola menarik atau model prediksi untuk dinilai, apakah hasil tersebut dapat dengan jelas merepresentasikan pengetahuan yang dibutuhkan.

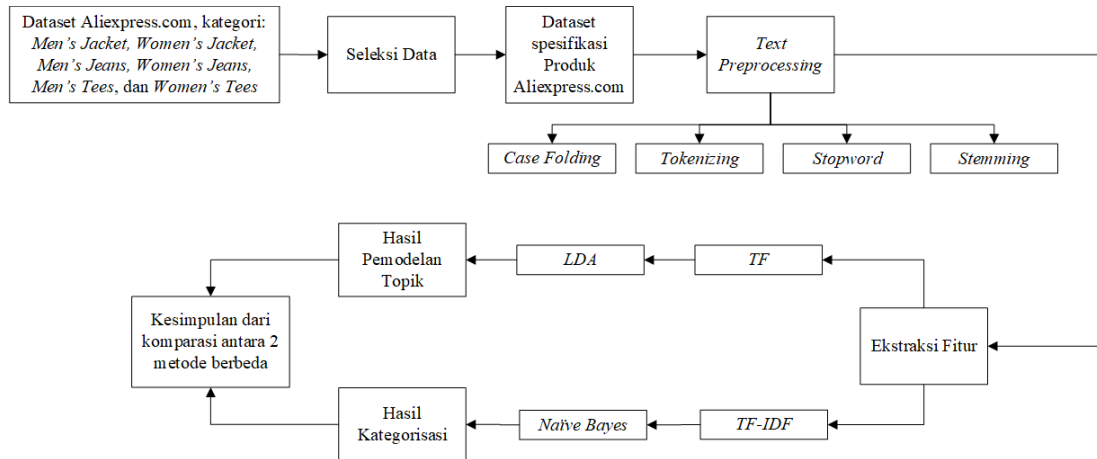
### 3.1.7 Representasi Pengetahuan

Tahap ini dilakukan sebagai tahap terakhir dari *data mining*, yaitu dengan menampilkan keluaran berupa visualisasi untuk memaparkan interpretasi hasil dari analisis dan pengolahan data yang telah dilakukan sebelumnya.

## 3.2 Gambaran Umum Penelitian

Gambaran umum penelitian menampilkan suatu bagan untuk memberikan penjelasan mengenai alur kerja atau urutan dari satu proses ke proses lainnya secara menyeluruh dalam suatu sistem yang akan dilakukan. Penelitian dimulai dengan memperoleh dan melakukan seleksi terhadap *dataset* produk *apparel* dari situs *marketplace* Aliexpress.com, sehingga diperoleh data berupa judul dan spesifikasi produk *apparel*. Selanjutnya dilakukan pembersihan dan pemrosesan data meliputi 4 (empat) tahap yaitu *case folding*, *tokenizing*, *stopword removal*, dan *stemming*. Setelah itu dilakukan ekstraksi fitur pada *dataset* dengan menggunakan metode TF-IDF dan melakukan tokenisasi data kedalam dua jenis yaitu data *unigram* dan *bigram*. Dari kedua data tersebut dilakukan proses pengelompokan data secara terpisah dengan pendekatan *unsupervised learning* yaitu dengan melakukan pemodelan topik menggunakan metode LDA, dan *supervised learning* dengan melakukan kategorisasi menggunakan metode *Naïve Bayes*, sehingga diperoleh hasil pengelompokan dari dua pendekatan berbeda. Gambaran umum penelitian ditampilkan dalam Gambar 2, sebagai berikut:





**Gambar 2.** Gambaran Umum Penelitian

### 3.3 Alat dan Bahan

#### 3.3.1 Alat

Alat yang digunakan untuk melakukan penelitian ini berupa perangkat keras (*Hardware*) dan perangkat lunak (*Software*), yaitu sebagai berikut:

##### 1. Perangkat Keras

- Laptop Acer Aspire E1-470G.
- Processor Intel® Core™ I3-3217U (CPU 1.8GHz, 3MB L3 Cache).
- ROM 500GB HDD.
- RAM 4GB DDR3 L Memory.
- VGA Intel® HD Graphics 4000.
- VGA NVIDIA GeForce 820M.

##### 2. Perangkat Lunak

- Sistem Operasi Windows 10 pro 64 bit.
- Rstudio.
- Anaconda Prompt.
- Python.
- Google Chrome.
- Microsoft Word 2019.
- Microsoft Visio 2019.
- Data Scraper Extension for Chrome.

#### 3.3.2 Bahan

Bahan yang digunakan untuk keperluan penelitian ini yaitu sebagai berikut :

- Data produk *apparel* untuk kategori *Men's Jacket, Women's Jacket, Men's Jeans, Women's Jeans, Men's Tees, dan Women's Tees* pada situs Aliexpress.com.
- Sejumlah referensi seperti buku dan jurnal yang terkait dengan penelitian.

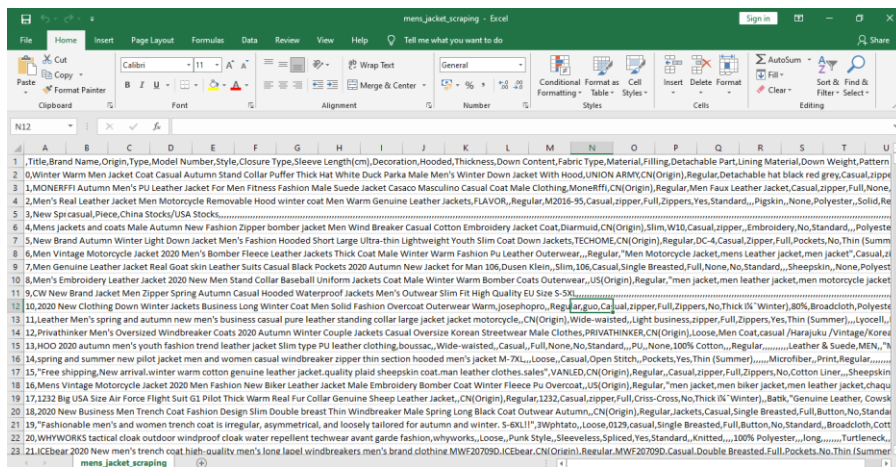
## BAB IV PERANCANGAN DAN IMPLEMENTASI

### 4.1 Tahap Pelaksanaan Penelitian

Tahap ini melakukan analisa terhadap kebutuhan selama penelitian, dengan menjelaskan prosedur yang harus dilalui dalam pelaksanaan penelitian. Pada tahap ini terdapat bagan terstruktur yang menjelaskan beberapa tahap penelitian seperti darimana sumber data berasal, bagaimana pengolahan data yang telah diperoleh sebelumnya, hingga mendapatkan kesimpulan dari hasil penelitian. Analisis kebutuhan pengerjaan sistem ditampilkan pada Lampiran 12.

### 4.2 Tahap Pengumpulan Data

Pengumpulan data dilakukan sebagai tahap awal dalam pengolahan data sebelum tahap penggalian informasi dimulai. Data awal diperoleh dengan melakukan proses *scraping* data dari situs *marketplace* Aliexpress.com menggunakan aplikasi *Data Scraper Extension for Chrome* untuk pengambilan judul dan *link* produk. *Scraping* data dilanjutkan dengan menggunakan bahasa pemrograman *Python* dan aplikasi *Anaconda Prompt*, dengan mengambil data spesifikasi produk pada kategori *Men's Jacket*, *Women's Jacket*, *Men's Jeans*, *Women's Jeans*, *Men's Tees*, dan *Women's Tees*. Proses *scraping* dan hasil *scraping* data ditampilkan dalam Lampiran 2 dan Gambar 3 sebagai berikut:



1	Title,Brand Name,Origin,Type,Model Number,Style,Closure Type,Sleeve Length(cm),Decoration,Hooded,Thickness,Down Content,Fabric Type,Material,Filling,Detachable Part,Lining Material,Down Weight,Pattern
2	0,Winter Warm Men Jacket Casual Autumn Stand Collar Puffer Thick Hat White Duck Parka Male Men's Winter Down Jacket With Hood,UNION ARMY,CN(Origin),Regular,Detachable hat black red grey,Casual,Zippe
3	1,MONIFFI Autumn Men's PU Leather Jacket For Men Fitness Fashion Male Suede Jacket Casaco Masculino Casual Coat Male Clothing,Monoffi,CN(Origin),Regular,Men Faux Leather jacket,Casual,zipper,Full,None,
4	2,Men's Real Leather Jacket Men Motorcycle Removable Hood winter coat Men Warm Genuine Leather Jackets,FLAVOR,,Regular,M2018-95,Casual,zipper,Full,Zippers,Yes,Standand,,Pigskin,,None,Polyester,,Solid,Re
5	3,New Sprcasual,Piece,China Stocks/USA Stocks
6	4,Mens jackets and coats Male Autumn New Fashion Zipper bomber jacket Men Wind Breaker Casual Cotton Embroidery Jacket Coat,Diarmuid,CN(Origin),Slim,W10,Casual,zipper,,Embroidery,No,Standard,,Polyeste
7	5,New Brand Autumn Winter Light Down Jacket Men's Fashion Hooded Short Large Ultra-thin Lightweight Youth Slim Coat Down Jackets,TECHOME,CN(Origin),Regular,DC-4,Casual,Zipper,Full,Pockets,No,Thin(Summ
8	6,Men Vintage Motorcycle Jacket 2020 Men's Bomber Fleece Leather Jackets Thick Coat Male Winter Warm Fashion Pu Leather Outerwear,,Regular,,Men Motorcycle jacket,mens Leather jacket,mens jacket,,Casual,Zi
9	7,Men Genuine Leather Jacket Real goat skin Leather Suits Casual Black Pockets 2020 Autumn New Jacket for Man 106,Dusen Klein,,Slim,106,Casual,Single Breasted,Full,None,No,Standard,,Sheepskin,,None,Polyest
10	8,Men's Embroidery Leather Jacket 2020 New Men Stand Collar Baseball Uniform Jackets Coat Male Winter Warm Bomber Coats Outerwear,,US(Origin),Regular,,men jacket,mens leather jacket,mens motorcycle jacket
11	9,CW New Brand Jacket Men Zipper Spring Autumn Casual Hooded Waterproof Jackets Men's Outerwear Slim Fit High Quality EU Size S-5XL
12	10,2020 New Clothing Down Winter Jackets Business Long Winter Coat Men Solid Fashion Overcoat Outerwear Warm,Josephopro,,Regular,guo,Casual,zipper,Full,Zippers,No,Thick 0x Winter,80%,Broadcloth,Polyeste
13	11,Leather Men's spring and autumn new men's business casual pure leather standing collar large jacket motorcycle,,CN(Origin),Wide-waisted,,Light business,zipper,Full,Zippers,Yes,Thin(Summer,,Lycocel,,
14	12,Privathtinker Men's Oversized Windbreaker Coats 2020 Autumn Winter Couple Jackets Casual Oversize Korean Streetwear Male Clothes,PRIVATHINKER,CN(Origin),Loose,Men Coat,Casual/Harajuku/Vintage/Koree
15	13,HOO 2020 autumn men's youth fashion trend leather jacket Slim type PU leather clothing,boussac,,Wide-waisted,,Casual,,Full,None,No,Standard,,PU,None,100% Cotton,,Regular,,Leather & Suede,MEN,,H
16	14, Spring and summer new pilot jacket men and women casual windbreaker zipper thin section hooded men's jacket M-7XL,,Loose,,Casual,Open Stitch,,Pockets,Yes,Thin(Summer,,Microfiber,,Print,Regular,,micro
17	15,,Free shipping,New arrival,winter warm cotton genuine leather jacket,quality plaid sheepskin coat,man leather clothes,sales',VANLED,CN(Origin),Regular,,Casual,zipper,Full,Zippers,No,Cotton Liner,,Sheepskin
18	16,Mens Vintage Motorcycle Jacket 2020 Men Fashion New Biker Leather Jacket Male Embroidery Bomber Coat Winter Fleece Pu Overcoat,,US(Origin),Regular,,men jacket,mens biker jacket,mens leather jacket,chaqu
19	17,1232 Big USA Size Air Force Flight Suit G1 Pilot Thick Warm Real Fur Collar Genuine Sheep Leather Jacket,,CN(Origin),Regular,1232,Casual,zipper,Full,Cross-Cross,No,Thick 0x Winter,,Bath,,Genuine Leather,,Cowki
20	18,2020 New Business Men Trench Coat Fashion Design Slim Double breast Thin Windbreaker Male Spring Long Black Coat Outerwear Autumn,,CN(Origin),Regular,Jackets,Casual,Single Breasted,Full,Button,No,Stand
21	19,,Fashionable men's and women trench coat is irregular, asymmetrical, and loosely tailored for autumn and winter. S-6XL11',3Wphtato,,Loose,0129,casual,Single Breasted,Full,Button,No,Standard,,Broadcloth,Cott
22	20,,WHYWORKS tactical cloak outdoor windproof cloak water repellent techwear avant garde fashion,whyworks,,Loose,,Punk Style,,Sleeveless,Spliced,Yes,Standard,,Knitted,,100% Polyester,,long,,Turtleneck,,
23	21,,ICEbear 2020 New men's trench coat high-quality men's lone laeol windbreakers men's brand clothing MWF20709D,ICEbear,CN(Origin),Regular,MWF20709D,Casual,Double Breasted,Full,Pockets,No,Thin(Summer

Gambar 3. Hasil Scraping Data

### 4.3 Tahap Seleksi dan Pembersihan Data

Hasil *scraping* memperoleh 5.700 data dari 6 (enam) kategori *apparel* pada situs Aliexpress.com. Seleksi dan pembersihan data dilakukan secara manual menggunakan aplikasi *Microsoft Excel 2019*, dengan menghapus *missing value*, membuang atribut yang kurang penting untuk pengolahan data selanjutnya, dan penambahan *class*, dengan hasil pemrosesan seperti pada Lampiran 3.

### 4.4 Tahap Transformasi Data

Tahap transformasi data dilakukan dengan menggunakan aplikasi *Microsoft Excel 2019* dengan menyamaratakan atribut yang akan digunakan dari setiap kategori dengan cara menggabungkan seluruh atribut kedalam satu *dataframe*, dimana satu

baris atribut sama dengan satu dokumen. Hasil dari transformasi data ditampilkan pada Lampiran 4.

#### 4.5 Tahap Pemrosesan Data (*Preprocessing*)

*Preprocessing* atau data pra-proses merupakan tahapan untuk membersihkan kumpulan data agar dapat lebih mudah untuk diolah. Tahapan ini mencakup perbaikan kesalahan pada data seperti pembuangan data yang terdapat duplikasi atau inkonsisten data. Tahapan *preprocessing* data terdiri atas *Case Folding*, *Tokenizing*, *Stopword Removal*, *Stemming*, dan Inisialisasi *N-gram*. Contoh data spesifikasi produk yang akan diolah kedalam tahap *preprocessing* dapat dilihat pada Gambar 4 berikut:

Style	Clothing Length	Material	Decoration
Casual	Short	Polyester	Zipper

**Gambar 4.** Contoh Data Spesifikasi Produk *Men's Jacket*

##### 4.5.1 *Case Folding*

*Case Folding* adalah salah satu tahapan *preprocessing data* untuk merubah semua huruf kapital (*uppercase*) yang terdapat pada dokumen menjadi huruf kecil (*lowercase*). Hal ini bertujuan agar seluruh kata yang terdapat pada dokumen tersebut memiliki standar yang sama, sehingga antara satu kata yang sama yang memiliki huruf kapital dengan kata lain yang memiliki huruf kecil tidak memiliki arti yang berbeda. *Case folding* hanya merubah ukuran pada karakter tertentu yaitu huruf a sampai huruf z. Contoh hasil *case folding* dapat dilihat pada Gambar 5 berikut:

Style	Clothing Length	Material	Decoration
casual	short	polyester	zipper

**Gambar 5.** Hasil *Case Folding*

##### 4.5.2 *Tokenizing*

*Tokenizing* adalah salah satu proses pemotongan suatu kalimat pada sebuah dokumen tertentu menjadi sebuah kata per kata yang terpisah. Kalimat yang melalui proses *tokenizing* akan dipisah menjadi sekumpulan *array* berisi kata dari kalimat tersebut. *Tokenizing* bertujuan untuk mengidentifikasi setiap kata yang dipisahkan dengan karakter tertentu, seperti spasi atau karakter lainnya. Contoh hasil *tokenizing* dapat dilihat pada Tabel 2 berikut:

**Tabel 2.** Hasil *tokenizing*

casual	short	polyester	zipper
--------	-------	-----------	--------

##### 4.5.3 *Stopword Removal*

*Stopword* adalah kata-kata umum yang biasanya tidak memiliki makna yang penting dalam suatu kalimat dan muncul dalam jumlah yang banyak. *Stopword removal* adalah salah satu tahapan untuk menghilangkan kata-kata umum yang bertujuan untuk mengurangi jumlah kata yang akan diproses ke tahap selanjutnya, sehingga hanya akan diambil kata-kata yang bernilai penting untuk dokumen tersebut. Contoh hasil *stopword removal* dapat dilihat pada Gambar 6 berikut:

Style	Clothing Length	Material	Decoration
casual	short	polyester	zippers

**Gambar 6.** Hasil *Stopword Removal*

#### 4.5.4 *Stemming*

*Stemming* adalah tahapan terakhir dari *preprocessing* data, yaitu dengan menghilangkan imbuhan yang terdapat pada setiap kata sehingga dokumen tersebut hanya menggunakan kata dasar saja. *Stemming* dilakukan pada saat pembuatan *index*, yang bertujuan untuk memperkecil jumlah *index* yang berbeda dan memudahkan pengelompokan kata dasar yang memiliki arti serupa namun memiliki bentuk yang berbeda. Contoh hasil *stemming* dapat dilihat pada Gambar 7 berikut:

Style	Clothing Length	Material	Decoration
casual	short	polyester	zipper

**Gambar 7.** Hasil *Stemming*

#### 4.5.5 *N-gram*

*N-gram* adalah pemecahan sebuah kalimat ke-dalam kata-kata ke dalam suatu nilai N, dimana nilai N bisa berupa 1 (satu) kata yang dikenal sebagai *unigram*, 2 (dua) kata yang dikenal sebagai *bigram*, dan 3 (tiga) kata yang dikenal sebagai *trigram*. *N-gram* dapat digunakan untuk pembentukan pemodelan topik sehingga hasil berupa kumpulan topik yang terbentuk akan lebih beragam dan mudah teridentifikasi. Hasil *stemming* yang terdapat pada Gambar 7 akan diubah ke dalam *unigram* dan *bigram*. Hasil *unigram* dan *bigram* ditampilkan pada Tabel 3 dan Tabel 4 berikut:

**Tabel 3.** Hasil *unigram*

'casual'	'short'	'polyester'	'zipper'
----------	---------	-------------	----------

**Tabel 4.** Hasil *bigram*

'casual short'	'short polyester'	'polyester zipper'
----------------	-------------------	--------------------

#### 4.5.6 *Term Frequency-Inverse Document Frequency (TF-IDF)*

TF-IDF adalah salah satu metode *feature selection* dan *term weighting* dengan cara memberikan bobot pada setiap kata dalam suatu dokumen. TF-IDF bertujuan untuk meminimalisir terjadinya *human error* terhadap pengkategorisasian dan mempermudah pengolahan data agar dapat diteruskan ke tahap selanjutnya.

Perhitungan TF-IDF dimulai dengan menentukan dokumen yang akan digunakan, kemudian mengambil kata-kata penting yang terdapat pada setiap dokumen untuk diberikan bobot sebesar 1, dan dikumpulkan ke dalam *bag of words*. Kalimat yang terdapat pada setiap dokumen diambil dari hasil *preprocessing* data. Selanjutnya menentukan nilai df, IDF, dan bobot dari setiap dokumen dengan proses perhitungan yang dapat dilihat pada Lampiran 11, persamaan (6), (7), dan (8) berikut:

$$df = \text{jumlah } tf \quad (6)$$

$$IDF = \log \frac{D}{df} \quad (7)$$

$$W = IDF * tf \quad (8)$$

#### 4.6 Tahap Penggalian Data (*Data Mining*)

*Data Mining* atau penggalian data adalah salah satu tahapan penting untuk proses pencarian pola atau informasi dalam data yang sudah ditransformasikan menggunakan algoritma atau metode tertentu. Tahapan ini juga dapat digunakan untuk metode pencarian pengetahuan (*knowledge*).

##### 4.6.1 *Latent Dirichlet Allocation (LDA)*

*Latent Dirichlet Allocation (LDA)* adalah metode *topic modelling* yang akan menentukan topik berdasarkan distribusi kata pada sebuah dokumen. LDA berjalan dengan cara memberikan topik acak ke setiap kata yang terdapat pada *corpus*, menentukan jumlah topik yang akan dicari, dan menghasilkan parameter berupa penentuan distribusi jumlah topik dalam dokumen dan penentuan distribusi kata dari sebuah topik. Parameter tersebut akan membantu dalam penentuan topik yang telah ditentukan jumlahnya.

Hasil perhitungan TF-IDF pada data spesifikasi produk yang telah dilakukan sebelumnya akan digunakan sebagai *document-term matrix* untuk menentukan topik dalam sebuah *corpus*. Penentuan topik dilakukan dengan mengambil nilai probabilitas kata yang paling besar, yang akan dibagi kedalam beberapa topik secara *unigram* dan *bigram*, dengan nilai K yang diperoleh dari hasil penentuan jumlah *cluster* dengan menggunakan *Silhouette Analysis*. Contoh dari hasil pemodelan topik untuk data *unigram* dan *bigram* ditampilkan pada Tabel 5 dan Tabel 6 berikut:

**Tabel 5.** Pemodelan Topik *unigram*

Topik 1	Topik 2	Topik 3	Topik 4
Casual	Short	Polyester	Zipper

**Topik 1** = 0,6020\*Casual

**Topik 2** = 0,6021\*Short

**Topik 3** = 0,5283\*Polyester

**Topik 4** = 0,6021\*Zipper

**Tabel 6.** Pemodelan Topik *bigram*

Topik 1	Topik 2	Topik 3
Casual short	Short polyester	Polyester zipper

**Topik 1** = 0,6021\*Casual short

**Topik 2** = 0,6021\*Short polyester

**Topik 3** = 0,6021\*Polyester zipper

#### 4.6.2 Naïve Bayes Classifier

*Naïve Bayes classifier* mengandalkan probabilitas dari *record* dalam suatu *dataset*. Penggunaan *Naïve Bayes classifier* dalam penelitian ini adalah untuk memberikan prediksi klasifikasi dari kumpulan data *training* terhadap data *testing* yang berupa data baru. Contoh data *training* dan data *testing* ditunjukkan pada Gambar 8 dan Gambar 9 berikut:

Style	Clothing Length	Material	Decoration	Class
casual	short	polyester	zipper	Men's Jacket
casual	regular	pu	none	Men's Jacket
streetwear	regular	wool	fur	Women's Jacket

**Gambar 8.** Contoh Data *Training*

Style	Clothing Length	Material	Decoration	Class
casual	long	wool	fur	?

**Gambar 9.** Contoh Data *Testing*

Perhitungan dilakukan dengan mencari probabilitas setiap atribut dengan kelas berbeda yaitu *Men's Jacket* dan *Women's Jacket*, menggunakan 3 (tiga) data *testing* dan 1 (satu) data *training* yang belum diketahui klasifikasinya. *Output* yang diharapkan dari perhitungan ini adalah data *testing* dapat dikategorisasikan ke dalam salah satu dari kumpulan kategorisasi yang terdapat pada dataset.

$$P | C = \frac{\text{Jumlah data yang sesuai dengan } C}{\text{Jumlah data keseluruhan yang terdapat pada kelas tersebut}}$$

#### Perhitungan kelas *Men's Jacket*:

$$\text{Kategori : } P | C \text{ Men's Jacket} = \frac{2}{3}$$

$$\text{Style : } P | C \text{ casual} = \frac{2}{2} = 1$$

$$\text{Clothing Length : } P | C \text{ long} = \frac{0}{2}$$

$$\text{Material : } P | C \text{ wool} = \frac{0}{2}$$

$$\text{Decoration : } P | C \text{ fur} = \frac{0}{2}$$

$$\frac{2}{3} \times \frac{2}{2} \times \frac{0}{2} \times \frac{0}{2} \times \frac{0}{2} = 0$$

**Perhitungan kelas *Women's Jacket*:**

$$\text{Kategori : } P | C \text{ Women's Jacket} = \frac{1}{3}$$

$$\text{Style : } P | C \text{ casual} = \frac{0}{1}$$

$$\text{Clothing Length : } P | C \text{ long} = \frac{0}{1}$$

$$\text{Material : } P | C \text{ wool} = \frac{1}{1}$$

$$\text{Decoration : } P | C \text{ fur} = \frac{1}{1}$$

$$\frac{1}{3} \times \frac{0}{1} \times \frac{0}{1} \times \frac{1}{1} \times \frac{1}{1} = 0$$

**4.6.2.1 Laplacian Correction**

Perhitungan untuk *Men's Jacket* dan *Women's Jacket* keduanya menghasilkan 0, dikarenakan ada salah satu probabilitas yang bernilai 0 maka ketika dilakukan perkalian untuk semua probabilitas akan menghasilkan 0. Untuk mengubah hasilnya agar tidak 0, maka dilakukan metode tambahan yaitu *Laplacian Correction*.

*Laplacian Correction* digunakan untuk mengubah hasil probabilitas yang bernilai 0, dengan cara menambahkan tupel ke *class* dan atribut yang terdapat pada dataset yang bernilai 0 sesuai dengan jumlah *class* dan atributnya.

Penggunaan *Laplacian Correction* pada data di atas menambahkan 1 tupel untuk masing-masing *class* yaitu *Men's Jacket* dan *Women's Jacket*, dan menambahkan 10 tupel untuk masing-masing atribut. Sehingga perhitungan probabilitas dengan *Laplacian Correction* menjadi seperti berikut:

**Perhitungan kelas *Men's Jacket*:**

$$\begin{aligned} & \frac{2}{3} \times \frac{3}{12} \times \frac{1}{12} \times \frac{1}{12} \times \frac{1}{12} = 0 \\ & = 0,667 \times 0,25 \times 0,083 \times 0,083 \times 0,083 \\ & = 0,00009534548225 \end{aligned}$$

**Perhitungan kelas *Women's Jacket*:**

$$\begin{aligned} & \frac{1}{3} \times \frac{1}{11} \times \frac{1}{11} \times \frac{2}{11} \times \frac{2}{11} = 0 \\ & = 0,333 \times 0,091 \times 0,091 \times 0,182 \times 0,182 \\ & = 0,000091341848052 \end{aligned}$$

**Perhitungan probabilitas dokumen:**

$$\begin{aligned} P(d) &= P(d|C \text{ Men's Jacket}) + P(d|C \text{ Women's Jacket}) \\ &= 0,00009534548225 + 0,000091341848052 \\ &= 0,000186687330302 \end{aligned}$$

**Perhitungan probabilitas Men's Jacket:**

$$\begin{aligned} P(C \text{ Men's Jacket}|d) &= \frac{P(d|C \text{ Men's Jacket})}{P(d)} = \frac{0,00009534548225}{0,000186687330302} \\ &= 0,5107228331765295 = 51,072\% \end{aligned}$$

**Perhitungan probabilitas Women's Jacket:**

$$\begin{aligned} P(C \text{ Women's Jacket}|d) &= \frac{P(d|C \text{ Women's Jacket})}{P(d)} = \frac{0,000091341848052}{0,000186687330302} \\ &= 0,4892771668234705 = 48,928\% \end{aligned}$$

Diperoleh hasil perhitungan probabilitas *Men's Jacket* terhadap data *testing* adalah sebesar 51,072% dan probabilitas *Women's Jacket* terhadap data *testing* adalah sebesar 48,928%. Sehingga dapat disimpulkan bahwa data *testing* tersebut masuk ke dalam kategori *Men's Jacket*. Hasil kategorisasi data *testing* dapat dilihat pada Tabel 7 berikut:

**Tabel 7.** Hasil Kategorisasi

Data	Type Data	Probabilitas	Hasil Kategorisasi
Data 1	Data training		<i>Men's Jacket</i>
Data 2	Data training		<i>Men's Jacket</i>
Data 3	Data training		<i>Women's Jacket</i>
Data 4	Data testing	<i>Men's Jacket</i> = 51,072% <i>Women's Jacket</i> = 48,928%	<i>Men's Jacket</i>

**4.7 Evaluasi Pola (Pattern Evaluation)**

Evaluasi pola pada penelitian ini dilakukan untuk menemukan pola pada hasil *data mining* sebelumnya yang akan merepresentasikan pengetahuan. Evaluasi pola dilakukan dengan menggunakan *Confusion Matrix*, yang akan menampilkan visualisasi dari kinerja model klasifikasi untuk mengidentifikasi kebenaran atau kekeliruan pada hasil klasifikasi. Tabel *Confusion Matrix*, tabel hasil kategorisasi data dengan *Naïve Bayes*, dan tabel *Confusion Matrix* untuk hasil kategorisasi dapat dilihat pada Tabel 8, Tabel 9, dan Tabel 10 berikut.



**Tabel 8.** *Confusion Matrix*

	<b>Prediksi</b>	
<b>Aktual</b>	TRUE	FALSE
TRUE	<i>True Positive (TP)</i>	<i>False Positive (FP)</i>
FALSE	<i>False Negative (FN)</i>	<i>True Negative (TN)</i>

Keterangan:

*True Positive (TP)* = data dengan kelas TRUE, berhasil diprediksi dengan benar dan bernilai kelas TRUE.

*True Negative (TN)* = data dengan kelas FALSE berhasil diprediksi dengan benar dan bernilai kelas FALSE.

*False Positive (FP)* = data dengan kelas TRUE gagal diprediksi dengan benar, sehingga bernilai kelas FALSE.

*False Negative (FN)* = data dengan kelas FALSE gagal diprediksi dengan benar, sehingga bernilai kelas TRUE.

**Tabel 9.** Hasil Kategorisasi Aktual dan Prediksi dengan *Naïve Bayes*

<b>Data</b>	<b>Aktual</b>	<b>Prediksi</b>
Data 1	<i>Men's Jacket</i>	<i>Men's Jacket</i>
Data 2	<i>Men's Jacket</i>	<i>Men's Jacket</i>
Data 3	<i>Women's Jacket</i>	<i>Women's Jacket</i>
Data 4	<i>Men's Jacket</i>	<i>Men's Jacket</i>

**Tabel 10.** *Confusion Matrix* untuk Hasil Kategorisasi

	<b>Prediksi</b>	
<b>Aktual</b>	<i>Men's Jacket</i>	<i>Women's Jacket</i>
<i>Men's Jacket</i>	3	0
<i>Women's Jacket</i>	0	1

Nilai *True Positive (TP)* pada tabel 14 sama dengan 3, yang berarti 3 (tiga) buah data spesifikasi produk dengan kelas *Men's Jacket* berhasil terprediksi dengan benar dan bernilai kelas *Men's Jacket*. Nilai *True Negative (TN)* pada tabel 14 sama dengan 1, yang berarti 1 (satu) buah spesifikasi produk dengan kelas *Women's Jacket* berhasil terprediksi dengan benar dan bernilai kelas *Women's Jacket*. Sementara untuk nilai *False Positive (FP)* dan *False Negative (FN)* masing-masing bernilai 0, yang berarti tidak ada data spesifikasi produk yang gagal terprediksi dengan benar. Dengan

demikian, hasil perhitungan dari *precision*, *accuracy*, dan *recall* dari data hasil kategorisasi adalah sebagai berikut:

$$Precision = \frac{TP}{FP+TP} = \frac{3}{0+3} = \frac{3}{3} = 1 = 100\%$$

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} = \frac{3+1}{(3+1+0+0)} = \frac{4}{4} = 1 = 100\%$$

$$Recall = \frac{TP}{FN+TP} = \frac{3}{0+3} = \frac{3}{3} = 1 = 100\%$$

#### 4.8 Representasi Pengetahuan (*Knowledge Representation*)

Tahap ini akan menyajikan pengetahuan berupa visualisasi yang mudah dipahami faktanya, sehingga dapat membantu proses penalaran dari informasi yang telah didapat sebelumnya.

##### 4.8.1 Tahap Perancangan

Tahap ini akan membuat rancangan sistem secara keseluruhan dengan menggunakan visualisasi berupa bagan yang akan menggambarkan alur proses yang terjadi pada sistem yang akan dibuat. Analisis kebutuhan pengerjaan sistem ditampilkan dalam Lampiran 12, dan tahap perancangan dalam penelitian ini ditampilkan dalam Lampiran 13.

##### 4.8.2 Tahap Implementasi

Tahap ini merupakan tahap penerapan sistem yang sudah dirancang sebelumnya seperti pada Lampiran 13, untuk dibangun kedalam sistem yang sebenarnya.

###### 4.8.2.1 Proses *Scraping*

Implementasi untuk proses *scraping* data yaitu melakukan pengambilan data spesifikasi produk Aliexpress.com sebanyak 5.700 data, menggunakan bahasa pemrograman *Python* dan aplikasi *Anaconda Prompt*, seperti yang ditampilkan pada Gambar 10 berikut:

```

Anaconda Prompt - python scrape.py --dataset "data aliexpress/men tees 960 d...
selenium.webdriver.remote.errorhandler.py", line 242, in check_response
    raise exception_class(message, screen, stacktrace)
selenium.common.exceptions.InvalidArgumentException: Message: invalid argument
(Session info: chrome=85.0.4183.121)

****

The above exception was the direct cause of the following exception:

Traceback (most recent call last):
  File "scrape.py", line 108, in <module>
    for result in pool.imap_unordered(func=scrape, iterable=urls_chunked):
  File "C:\Users\ACER\AppData\Local\conda\conda\envs\scraping\lib\multiprocessin
g\pool.py", line 748, in next
    raise value
selenium.common.exceptions.InvalidArgumentException: Message: invalid argument
(Session info: chrome=85.0.4183.121)

<scraping> D:\aaaaa\python scrape.py --dataset "data aliexpress/men tees 960 dat
a link.csv" --save_to a.csv --processes 2
Running into 2 batch...
0%: | 0/480 [00:00<?, ?it/s]
0%: | 1/480 [00:33<4:26:38, 33.46s/it]
0%: | 1/480 [00:32<4:23:16, 32.78s/it]

```

Gambar 10. Proses *Scraping*

###### 4.8.2.2 Proses *Preprocessing* dan *Data Mining*

Implementasi untuk proses *preprocessing* dengan 2 (dua) tahap, yaitu secara manual seperti yang ditampilkan pada Lampiran 3, dan menggunakan aplikasi *Rstudio* dengan *library tm*, seperti yang ditampilkan pada Lampiran 4 dan Gambar 11 berikut:

```

1 library("NLP") # untuk scan_tokenizer (tokenisasi)
2
3 # ===== PREPROCESSING RAW =====
4
5 # membaca file dataset dengan ekstensi .csv
6 # set working directory ke E:
7 setwd("E:/")
8 docs <- read.csv(file = "E:/aaaaa/dataset_apparel_alieexpress.csv", header = TRUE,
9 docs
10
11 # melihat struktur dari dokumen
12 str(docs)
13
14 # membuat corpus untuk 4 atribut: Style, Clothing Length, Material, Decoration
15 # corpus = kumpulan teks yang bisa dijadikan sebagai sumber data
16 corpus1 <- vCorpus(vectorSource(docs$Style)) # creating corpus
17 corpus2 <- vCorpus(vectorSource(docs$Clothing.Length)) # creating corpus
18 corpus3 <- vCorpus(vectorSource(docs$Material)) # creating corpus
19 corpus4 <- vCorpus(vectorSource(docs$Decoration)) # creating corpus
20
21
22
23

```

```

Attaching package: 'ggplot2'

The following object is masked from 'package:NLP':

  annotate

Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3wBa
warning message:
package 'factoextra' was built under R version 4.0.5
>
> # scaling the attributes values because of uneven scaling
> Midf_scale <- scale(Midf)
> View(Midf_scale)

```

**Gambar 11.** Proses *Preprocessing*

Untuk proses *data mining*, dilakukan dengan menggunakan aplikasi *Rstudio* dengan *library Rweka* untuk membuat fungsi tokenisasi *unigram* dan *bigram*, merubah bentuk dataset kedalam sebuah *DocumentTermMatrix* (DTM) untuk membentuk *terms* dan mendapatkan nilai TF-IDF, seperti yang ditampilkan pada Gambar 12 berikut:

```

169
170
171 # ===== DTM saja =====
172
173 # membuat function untuk tokenisasi unigram & bigram
174 library("Rweka")
175 untokenizer <- function(x) NgramTokenizer(x, weka_control(min=1, max=1))
176 bitokenizer <- function(x) NgramTokenizer(x, weka_control(min=2, max=2))
177
178 control_list_unigram = list(tokenize = untokenizer,
179                             weighting = function(x)
180                               weightTF(x)
181                           )
182
183 control_list_bigram = list(tokenize = bitokenizer,
184                             weighting = function(x)
185                               weightTF(x)
186                           )
187
188
189
20347 DTM saja

```

```

> dtm_unigram = tm::DocumentTermMatrix(corpus, control_list_unigram)
> dtm_ngram = tm::DocumentTermMatrix(corpus, control_list_bigram)
> dtm_unigram
<<DocumentTermMatrix (documents: 3779, terms: 288)>>
Non-/sparse entries: 18717/1069635
Sparsity: 98%
Maximal term length: 23
weighting: term frequency (tf)
> dtm_ngram
<<DocumentTermMatrix (documents: 3779, terms: 1000)>>
Non-/sparse entries: 15298/3763702
Sparsity: 100%
Maximal term length: 31

```

**Gambar 12.** Proses *Data Mining*

#### 4.8.2.3 Proses Pemodelan Topik

Implementasi untuk proses pemodelan topik dengan metode *Latent Dirichlet Allocation* (LDA) menggunakan aplikasi *Rstudio*, dengan menggunakan *library cluster, factoextra, pacman, topicmodels, dplyr, tidyr, igrph, devtools, LDavis, ggplot2, topicdoc* dan menggunakan fungsi *topicmodels2LDavis* dan *optimal\_k*, seperti yang ditampilkan pada Gambar 13 berikut:

```

220
221 # ----- LDA -----
222 |
223 # install.packages("cluster")
224 # install.packages("factoextra")
225 library(cluster) # Algoritma klastering
226 library(factoextra) # Algoritma klastering dan visualisasi
227
228 # scaling the attributes values because of uneven scaling
229 M1df_scale <- scale(M1df)
230 M2df_scale <- scale(M2df)
231
232 fviz_nbclust(M1df_scale, kmeans, method = "silhouette")+
233   labs(subtitle = "silhouette method") # metode silhouette
234
235 fviz_nbclust(M2df_scale, kmeans, method = "silhouette")+
236   labs(subtitle = "silhouette method") # metode silhouette
237
238 ## Install dan menggunakan library yang sudah diinstall sebelumnya
239
222:1 LDA
R Script

```

```

Attaching package: 'ggplot2'

The following object is masked from 'package:NLP':

  annotate

welcome! want to learn more? See two factoextra-related books at https://goo.gl/ve3wba
warning message:
package 'factoextra' was built under R version 4.0.5
> # scaling the attributes values because of uneven scaling
> M1df_scale <- scale(M1df)
> view(M1df_scale)
> |

```

Gambar 13. Proses Pemodelan Topik

#### 4.8.2.4 Proses Kategorisasi

Implementasi untuk proses kategorisasi dengan algoritma *Naïve Bayes* menggunakan aplikasi *Rstudio*, dengan menggunakan *library e1071*, *SparseM*, *tm*, *caret*, *caTools*, *gmodels*, dan *rsample*, seperti yang ditampilkan pada Gambar 14 berikut:

```

308
309- #### KLASIFIKASI KATEGORI KE 1 ####
310
311 df.k <- M1[, c("documents", "category")]
312 # colnames(df.k) <- c("words", "class")
313
314 set.seed(123)
315 rows <- sample(nrow(df.k)) # data random
316 df.k <- df.k[rows, ]
317
318 k_split <- initial_split(df.k, prop = .80, strata = "category") # bagi training set 80%
319 k_train <- training(k_split)
320 k_test <- testing(k_split)
321
322 k_features <- setdiff(names(k_train), "category")
323 kx_train <- k_train[, k_features]
324 ky_train <- k_train$category
325
326 ktest_features <- setdiff(names(k_test), "category")
327
234:23 LDA
R Script

```

```

>
> #### KLASIFIKASI KATEGORI KE 1 ####
>
> df.k <- M1[, c("documents", "category")]
> # colnames(df.k) <- c("words", "class")
>
> set.seed(123)
> rows <- sample(nrow(df.k)) # data random
> df.k <- df.k[rows, ]
>
> k_split <- initial_split(df.k, prop = .80, strata = "category") # bagi training set 80%
> k_train <- training(k_split)
> k_test <- testing(k_split)
>
> k_features <- setdiff(names(k_train), "category")

```

Gambar 14. Proses Kategorisasi

## BAB V HASIL DAN PEMBAHASAN

### 5.1 Hasil

Tahap ini akan memaparkan hasil dari sistem yang telah dibangun berupa *output* dari pemodelan topik dan kategorisasi berdasarkan data spesifikasi produk *apparel* yang telah dilakukan berdasarkan perancangan sebelumnya.

#### 5.1.1 Kondisi Dataset

Setelah *scraping* data dilakukan dari situs Aliexpress.com dengan 6 (enam) kategori *apparel* yaitu *Men's Jacket*, *Women's Jacket*, *Men's Jeans*, *Women's Jeans*, *Men's Tees*, dan *Women's Tees*, didapat sejumlah data berisi spesifikasi setiap produk *apparel* menggunakan Bahasa Inggris, dengan rincian kondisi *dataset* yang ditampilkan pada Tabel 11 berikut.

**Tabel 11.** Kondisi *Dataset*

No.	Kategori	Jumlah Data Hasil <i>Scraping</i>	Jumlah Data Setelah <i>Preprocessing</i> Manual	Jumlah Data Setelah <i>Preprocessing</i> Rstudio
1.	<i>Men's Jacket</i>	960	748	717
2.	<i>Women's Jacket</i>	900	754	738
3.	<i>Men's Jeans</i>	960	581	528
4.	<i>Women's Jeans</i>	960	904	890
5.	<i>Men's Tees</i>	960	288	244
6.	<i>Women's Tees</i>	960	669	662
<b>Total</b>		5.700 data	3.944 data	3.779 data

#### 5.1.2 Hasil *Preprocessing* dan Transformasi Data

Tahap *preprocessing* dan transformasi data dilakukan dengan 2 (dua) proses, proses secara manual menggunakan aplikasi *Microsoft Excel 2019* untuk menghilangkan *missing value*, memilih dan mengambil atribut penting, dan menambahkan *class*. Proses selanjutnya menggunakan aplikasi *Rstudio* untuk pembersihan data dan pengubahan data kedalam format *dataframe* agar memudahkan proses pengolahan data. Hasil *preprocessing* ditampilkan pada Gambar 15 dan Gambar 16 berikut.

Style	Clothing.Length	Material	Decoration	Class
4 Casual	Regular	Polyester	Embroidery	Men Jacket
5 Casual	Regular	Cotton	Pockets	Men Jacket
6 Casual	Regular	PU	Zippers	Men Jacket
7 Casual	Regular	Sheepskin	None	Men Jacket
8 Casual	Regular	PU	Embroidery	Men Jacket
9 Casual	Regular	Polyester	Zippers	Men Jacket
10 Light business	Short	Lyocell	Zippers	Men Jacket
11 casual /Harajuku /Vintage/Korean Style/Fashion/Japan/stree...	long	Polyester	Embroidery	Men Jacket
12 Casual	Regular	Sheepskin	Zippers	Men Jacket
13 Casual	Regular	PU	Zippers	Men Jacket
14 Casual	Short	Genuine Leather: Cowskin	Cross-Cross	Men Jacket

Showing 3 to 14 of 3,944 entries, 5 total columns

**Gambar 15.** Sebelum *Preprocessing* dan Transformasi Data

	text_clean	documents
1	casual short polyester zipper	casual short polyester zipper
2	casual regular pu none	casual regular pu none
3	casual regular pigskin zipper	casual regular pigskin zipper
4	casual regular cotton pocket	casual regular cotton pocket
5	casual regular pu zipper	casual regular pu zipper
6	casual regular sheepskin none	casual regular sheepskin none
7	casual regular polyester zipper	casual regular polyester zipper
8	light business short lyocell zipper	light business short lyocell zipper
9	casual regular sheepskin zipper	casual regular sheepskin zipper
10	casual regular pu zipper	casual regular pu zipper
11	casual short genuine leather cowskin crisscross	casual short genuine leather cowskin crisscross

Showing 1 to 12 of 3,779 entries, 2 total columns

**Gambar 16.** Sesudah *Preprocessing* dan Transformasi Data

### 5.1.3 Hasil *Term Frequency-Inverse Document Frequency (TF-IDF)*

Metode *Term Frequency-Inverse Document Frequency (TF-IDF)* dalam penelitian ini digunakan untuk pembobotan kata dan ekstraksi fitur pada setiap dokumen dengan melakukan tokenisasi secara *unigram* dan *bigram* dan menjadikan hasil tokenisasi sebagai *term*. Hasil pembobotan kata dan ekstraksi fitur menggunakan TF-IDF ditampilkan pada Gambar 17 dan Gambar 18 berikut.

	documents	acetate	acrylic	active	adjustable	america	american	ankle	applique
1	casual short polyester zipper	0	0	0	0	0	0	0	0
2	casual regular pu none	0	0	0	0	0	0	0	0
3	casual regular pigskin zipper	0	0	0	0	0	0	0	0
4	casual regular cotton pocket	0	0	0	0	0	0	0	0
5	casual regular pu zipper	0	0	0	0	0	0	0	0
6	casual regular sheepskin none	0	0	0	0	0	0	0	0
7	casual regular polyester zipper	0	0	0	0	0	0	0	0
8	light business short lyocell zipper	0	0	0	0	0	0	0	0
9	casual regular sheepskin zipper	0	0	0	0	0	0	0	0
10	casual regular pu zipper	0	0	0	0	0	0	0	0

Showing 1 to 11 of 3,779 entries, 291 total columns

**Gambar 17.** Hasil TF-IDF *unigram*

	documents	acetate.button	acetate.cotton	acetate.fake	acetate.fur	acetate.hole	acetate.none
1	casual short polyester zipper	0	0	0	0	0	0
2	casual regular pu none	0	0	0	0	0	0
3	casual regular pigskin zipper	0	0	0	0	0	0
4	casual regular cotton pocket	0	0	0	0	0	0
5	casual regular pu zipper	0	0	0	0	0	0
6	casual regular sheepskin none	0	0	0	0	0	0
7	casual regular polyester zipper	0	0	0	0	0	0
8	light business short lyocell zipper	0	0	0	0	0	0
9	casual regular sheepskin zipper	0	0	0	0	0	0
10	casual regular pu zipper	0	0	0	0	0	0

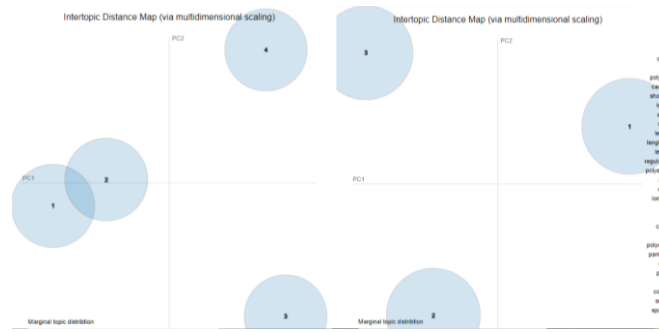
Showing 1 to 11 of 3,779 entries, 1003 total columns

**Gambar 18.** Hasil TF-IDF *bigram*

### 5.1.4 Hasil *Topic Modelling*

Pemodelan topik dilakukan dengan terlebih dahulu menentukan jumlah topik yang ingin dibuat, dimana jumlah topik ditentukan oleh hasil plotting dari pendekatan *Silhouette*, yang menghasilkan  $K = 7, 2, 4,$  dan  $3$  topik untuk *unigram* dan  $K = 3, 5,$

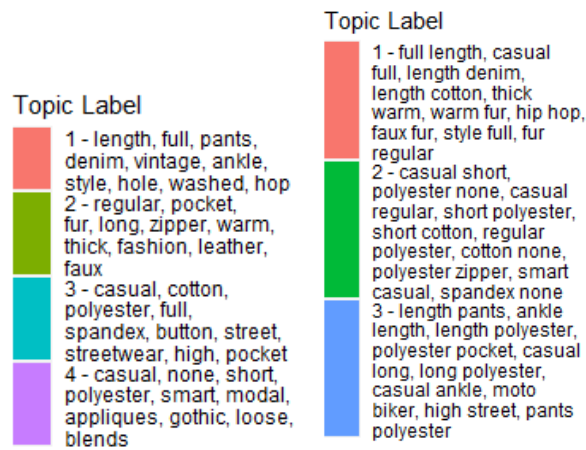
dan 9 topik untuk *bigram*. *Library ggplot2* dan *LDAvis* digunakan untuk menampilkan visualisasi pemodelan topik yang berhasil terbentuk, dimana nilai K yang digunakan harus  $\Rightarrow 2$ . Hasil pemodelan topik dan distribusi kata untuk setiap topik *unigram* dan *bigram* dengan nilai K yang berbeda-beda dapat dilihat pada lampiran 5. Hasil pemodelan topik dan distribusi kata tiap topik *unigram* dengan  $K = 4$  dan *bigram* dengan  $K = 3$  ditampilkan pada Gambar 19 dan Gambar 20, dan hasil interpretasi setiap topik *unigram* dan *bigram* ditampilkan pada Tabel 12 dan Tabel 13 berikut.



(a)

(b)

**Gambar 19.** Hasil Visualisasi LDA (a) 4 topik *unigram* dan (b) 3 topik *bigram*



(a)

(b)

**Gambar 20.** Distribusi Kata LDA (a) 4 topik *unigram* dan (b) 3 topik *bigram*

**Tabel 12.** Interpretasi Topik LDA *unigram*

Topik	Distribusi Kata Tiap Topik	Interpretasi Topik
1.	<i>Length full pants denim vintage ankle style hole washed hop</i>	Gaya ( <i>style</i> )
2.	<i>Regular pocket fur long zipper warm thick fashion leather faux</i>	Jaket ( <i>jacket</i> )
3.	<i>Casual cotton polyester full spandex button street streetwear high pocket</i>	Gaya jalanan ( <i>streetwear</i> )
4.	<i>Casual none short polyester smart modal appliques gothic loose blends</i>	Gaya gotik ( <i>gothic</i> )

**Tabel 13.** Interpretasi Topik LDA *bigram*

Topik	Distribusi Kata Tiap Topik	Interpretasi Topik
1.	<i>Full length casual full length denim length cotton thick warm warm fur hip hop faux fur style full fur regular</i>	Bulu ( <i>fur</i> )
2.	<i>Casual short polyester none casual regular short polyester short cotton regular polyester cotton none polyester zipper smart casual spandex none</i>	Polyester ( <i>polyester</i> )
3.	<i>Length pants ankle length length polyester polyester pocket casual long long polyester casual ankle moto biker high street pants polyester</i>	Panjangnya ( <i>length / long</i> )

### 5.1.5 Hasil Kategorisasi

Kategorisasi dilakukan berdasarkan kumpulan data spesifikasi produk *apparel* yang terbagi atas data latih dan data uji, untuk mengetahui dan mengelompokkan setiap produk kedalam kategori dan sub-kategori yang sesuai dengan menggunakan algoritma *Naïve Bayes*. Hasil kategorisasi untuk kategori dan sub-kategori produk *apparel* ditampilkan pada Gambar 21 berikut.

	name	actual_category	predict_category	actual_sub_category	predict_sub_category
1	vintage regular sheepskin none	Men	Men	Women Jeans	Women Jeans
2	double faced fur regular raccoon dog	Women	Women	Women Jeans	Women Jeans
3	casual regular polyester none	Women	Men	Women Jeans	Women Jeans
4	streetwear full length spandex hole	Women	Women	Women Jacket	Women Jacket
5	inflation hip hop men jeans full length cotton polyester poc...	Men	Women	Men Jacket	Men Jacket
6	casual full length denim pocket	Men	Men	Women Jeans	Women Jeans
7	casual regular polyester pattern	Men	Men	Men Jeans	Women Jeans
8	casual regular pu none	Men	Men	Men Jeans	Men Jeans
9	smart casual short cotton button	Men	Men	Men Jacket	Men Jacket
10	streetwear full length polyester tie dye	Women	Women	Women Jeans	Women Jeans

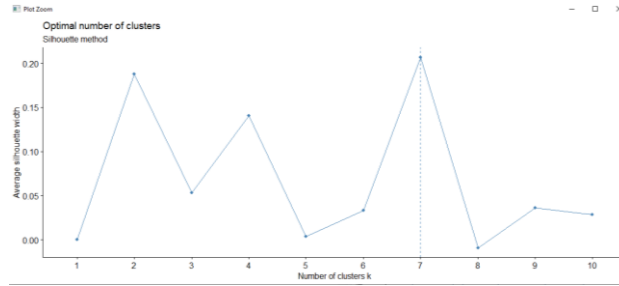
Showing 1 to 11 of 754 entries, 5 total columns

**Gambar 21.** Hasil Kategorisasi

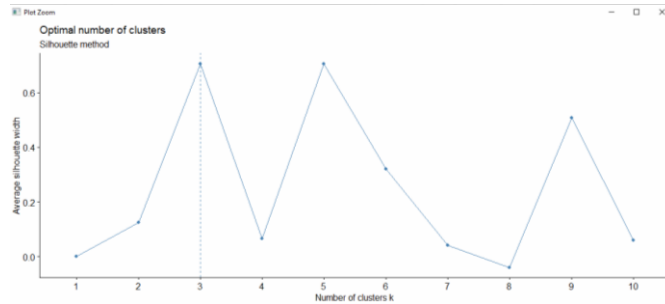
### 5.1.6 Hasil Silhouette Analysis

*Silhouette Analysis* dibutuhkan dalam penelitian ini untuk menentukan nilai K yang optimal sebelum dilakukan pemodelan topik menggunakan LDA. Data *DocumentTermMatrix* yang diubah kedalam bentuk *dataframe* digunakan untuk menghitung nilai K dengan nilai rata-rata paling tinggi untuk *unigram* dan *bigram*. Hasil *plotting* dari *Silhouette Analysis* ditampilkan pada Gambar 22 dan Gambar 23 berikut.





**Gambar 22.** Hasil *Silhouette Analysis* untuk data *unigram*



**Gambar 23.** Hasil *Silhouette Analysis* untuk data *bigram*

### 5.1.7 Hasil Evaluasi

Evaluasi dari metode *naïve bayes* dengan menggunakan perhitungan *confusion matrix* diperlukan untuk mengetahui distribusi pengelompokan dari kategorisasi yang telah dilakukan, dan diperlukan perhitungan *accuracy*, *precision*, *recall*, dan *F1-score* untuk melihat nilai keakuratan dari hasil kategorisasi data spesifikasi produk *apparel*. Hasil *confusion matrix* untuk kategori dan sub-kategori *apparel* dapat dilihat pada Gambar 24, dan hasil *accuracy*, *precision*, *recall*, dan *F1-score* untuk setiap kategori dapat dilihat pada Tabel 14 berikut.

```
> sk.cm.test
Confusion Matrix and Statistics
```

		Predicted							
		Men Jacket	Men Jeans	Men Tees	Women Jacket	Women Jeans	Women Tees		
Actual	Men Jacket	99	0	0	11	33	1		
	Men Jeans	0	66	0	0	38	0		
	Men Tees	0	0	26	1	9	4		
	Women Jacket	12	0	0	87	52	0		
	Women Jeans	0	0	0	0	181	0		
	Women Tees	0	0	0	2	16	116		

```
> k.cm.test
Confusion Matrix and Statistics
```

		Predicted	
		Men	Women
Actual	Men	188	109
	women	17	440

(a)

(b)

**Gambar 24.** Hasil *Confusion Matrix* (a) Kategori dan (b) Sub-kategori *Apparel*

**Tabel 14.** Nilai *accuracy*, *precision*, *recall* dan *F1-score* Kategorisasi

	<b>Nama Kategori</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>
<b>Kategori</b>	<i>Men &amp; Women</i>	83,29%	63,30%	91,71%	74,90%
<b>Sub-kategori</b>	<i>Men's Jacket</i>	91,10%	68,75%	89,19%	77,65%
	<i>Women's Jacket</i>	88,17%	57,62%	86,14%	69,05%
	<i>Men's Jeans</i>	97,24%	63,46%	100%	77,65%
	<i>Women's Jeans</i>	77,51%	100%	55,02%	70,98%
	<i>Men's Tees</i>	99,04%	65%	100%	78,79%
	<i>Women's Tees</i>	96,51%	86,57%	95,87%	90,98%
<b>RATA-RATA</b>		90,41%	72,1%	88,28%	77,14%

## 5.2 Pembahasan

Dataset hasil *scraping* berbentuk unik, memiliki karakter atau simbol yang tidak terlalu penting dan bersifat *noise*, terdapat *missing value* pada beberapa atribut yang cukup banyak, terdapat beberapa atribut-atribut tambahan yang tidak terlalu penting dan tidak berdampak secara langsung pada hasil penelitian, serta setiap kategori memiliki nama atribut yang berbeda namun memiliki kemiripan antara satu atribut dengan atribut lainnya. Berdasarkan kondisi dataset yang ditampilkan pada Tabel 11, dataset awal dengan jumlah 5.700 data berkurang menjadi 3.944 data setelah dilakukan pemilihan data dan atribut yang penting, menghapus *missing value* serta pemberian kelas pada setiap baris data secara manual menggunakan *Microsoft Excel 2019*, dan berkurang kembali menjadi 3.779 data setelah dilakukan pembersihan isi data dengan menggunakan tahapan *preprocessing* dan melakukan transformasi data kedalam sebuah *dataframe* pada aplikasi *Rstudio*, dan menggabungkan data dari 4 atribut menjadi 1 dokumen menggunakan *Microsoft Excel 2019*. Hasil pembersihan secara manual menghasilkan data dengan 4 (empat) atribut dengan *missing value* paling sedikit dan mewakili spesifikasi dari setiap produk, yaitu atribut *Style*, *Clothing Length*, *Material*, dan *Decoration*, 1 (satu) kelas untuk 2 (dua) kategori yaitu *Men* dan *Women*, dan 1 (satu) kelas untuk 6 (enam) sub-kategori yaitu *Men's Jacket*, *Women's Jacket*, *Men's Jeans*, *Women's Jeans*, *Men's Tees*, dan *Women's Tees*. Isi data spesifikasi produk pada setiap atribut bersifat pendek (*short text*) dan kurang tepat untuk dijadikan dokumen untuk setiap atribut karena kurang efektif untuk proses ekstraksi fitur dan kategorisasi, untuk itu data ditransformasikan menggunakan *Microsoft Excel 2019* kedalam sebuah dokumen, dimana satu baris data spesifikasi produk sama dengan satu dokumen, seperti yang ditampilkan pada Gambar 25 berikut.

	<b>Style</b>	<b>Clothing.Length</b>	<b>Material</b>	<b>Decoration</b>	<b>documents</b>
1	Casual	Short	Polyester	Zippers	casual short polyester zipper
2	Casual	Regular	PU	None	casual regular pu none
3	Casual	Regular	Pigskin	Zippers	casual regular pigskin zipper

(a)

(b)

**Gambar 25.** Hasil (a) Sebelum dan (b) Sesudah Transformasi Data

Proses ekstraksi fitur dengan menggunakan metode TF-IDF dilakukan dengan terlebih dahulu memecah dokumen kedalam setiap satu suku kata untuk *unigram* dan dua suku kata untuk *bigram* untuk dijadikan *terms*, dan pembobotan pada setiap kata *unigram* dan *bigram* dengan jumlah dokumen sebanyak 3.799 dokumen, 288 *terms* untuk *unigram*, dan 1000 *terms* untuk *bigram*. Kondisi dataset awal yang bersifat dokumen pendek (*short text*) tidak memungkinkan hasil dataset dengan TF-IDF untuk diteruskan ke proses selanjutnya yaitu pada pemodelan topik menggunakan LDA, karena dataset *short text* memiliki persentase kekosongan data (*sparsity*) yang besar karena jumlah kata di setiap dokumennya terlalu sedikit, sehingga pembobotan kata hanya menggunakan pembobotan dengan metode *Term Frequency* (TF), dan berlanjut ke proses pemodelan topik menggunakan LDA. (Likhitha *et al.*, 2019)

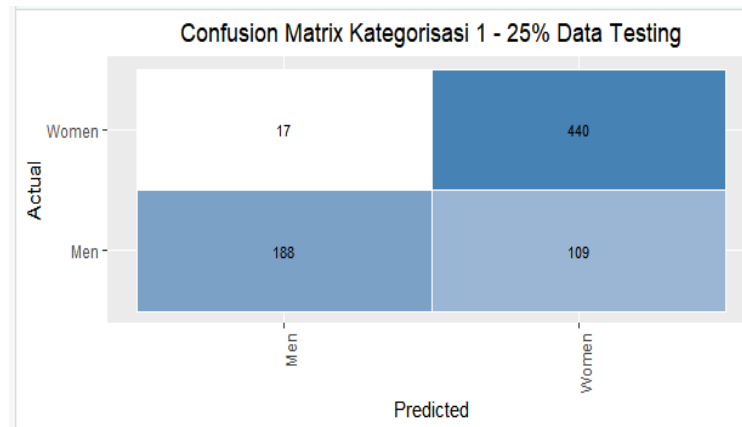
Untuk menghasilkan pemodelan topik dan kategorisasi pada penelitian yang dilakukan, kedua proses tersebut dilakukan secara terpisah dengan mengambil sumber data yang sama, yaitu data hasil pembobotan kata menggunakan metode *Term Frequency* (TF). Data hasil pembobotan kata tersebut kemudian diubah kedalam 2 (dua) format, yaitu format *DocumentTermMatrix* untuk digunakan pada proses selanjutnya yaitu pemodelan topik dengan metode LDA, dan format *DocumentTermMatrix* yang diubah kedalam *dataframe* untuk digunakan pada proses selanjutnya yaitu kategorisasi dengan algoritma *Naïve Bayes*.

Sebelum melakukan pemodelan topik, diperlukan penentuan jumlah topik berupa nilai K yang akan digunakan pada metode LDA agar memperoleh topik yang optimal. Penentuan jumlah topik dilakukan dengan menggunakan metode *Silhouette Analysis*, dimana data dengan format *DocumentTermMatrix* yang telah ditransformasikan kedalam bentuk *dataframe* akan melewati proses *scaling* yaitu pemberian nilai untuk data pada setiap atribut, selanjutnya penerapan metode *Silhouette* pada aplikasi *Rstudio* dilakukan pada data hasil *scaling* dan menghasilkan dua buah plot untuk *unigram* dan *bigram* dan menampilkan hasil K optimal yang dilihat dari nilai rata-rata tertinggi seperti pada Gambar 23 dan Gambar 24. Diperoleh nilai K tertinggi yaitu K = 7, 2, 4, dan 3 untuk data *unigram* dan K = 3, 5, dan 9 untuk data *bigram*.

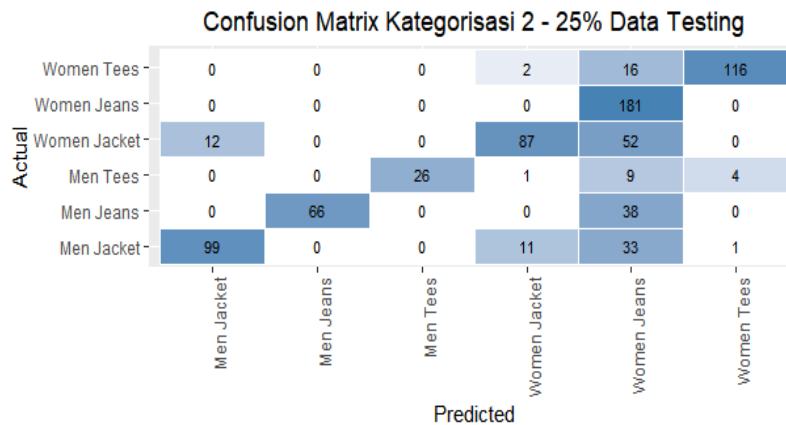
Pemodelan topik dilakukan dengan menerapkan nilai K pada data *unigram* dan *bigram* sebagai jumlah topik yang akan dihasilkan. Proses LDA yang dibutuhkan pada penelitian ini menggunakan aplikasi *Rstudio* dan hanya bisa menggunakan nilai  $K \geq 2$ , sehingga pemodelan topik dilanjutkan dengan K = 7, 4, dan 3 untuk data *unigram* dan K = 3, 5, dan 9 untuk data *bigram*. Hasil pemodelan topik dengan K = 7 untuk *unigram* seperti pada lampiran 5 memberikan hasil yang kurang baik dengan cukup banyak irisan antar topik, mengakibatkan beberapa topik memiliki beberapa distribusi kata yang sama dengan topik lainnya. Sehingga pemodelan topik pada penelitian ini mengambil dari nilai K optimal *unigram* tertinggi selanjutnya yaitu K = 4 yang irisan antar topiknya lebih sedikit dibandingkan dengan nilai K lain, dan K optimal *bigram* tertinggi pertama yaitu K = 3. Pada pemodelan topik *unigram*, Topik 1 menginterpretasikan isi topik tentang gaya pakaian atau *style*, Topik 2 meninterpretasikan isi topik tentang salah satu jenis pakaian yaitu *jacket*, Topik 3

meninterpretasikan isi topik tentang salah satu gaya pakaian yaitu *streetwear*, dan Topik 4 menginterpretasikan isi topik tentang salah satu gaya pakaian yaitu *gothic*. Sedangkan pada pemodelan topik *bigram*, Topik 1 menginterpretasikan isi topik tentang bahan pakaian dari bulu yaitu *fur*, Topik 2 meninterpretasikan isi topik tentang bahan pakaian yaitu *polyester*, dan Topik 3 meninterpretasikan isi topik tentang panjang pakaian atau *length/long*.

Pada proses kategorisasi, aplikasi *Rstudio* tidak dapat melakukan klasifikasi secara hirarkis (*hierarchical classification*), maka dari itu dalam melakukan kategorisasi spesifikasi data produk dilakukan dengan 2 (dua) proses, yaitu kategorisasi kategori dan kategorisasi sub-kategori. Kategorisasi kategori berisikan dua kategori berdasarkan jenis kelamin, yaitu *Men* dan *Women*, dan kategorisasi sub-kategori berisikan enam kategori berdasarkan jenis *apparel*, yaitu *Men's Jacket*, *Women's Jacket*, *Men's Jeans*, *Women's Jeans*, *Men's Tees*, dan *Women's Tees*. Proses kategorisasi menggunakan algoritma *Naïve Bayes* dengan data latih sebesar 80% yaitu sebanyak 3.025 data, dan data uji sebesar 20% yaitu sebanyak 754 data dari total 3.779 data spesifikasi produk. Visualisasi untuk hasil kategorisasi dapat dilihat pada Gambar 26 dan Gambar 27 berikut.



**Gambar 26.** Hasil Kategorisasi Kategori



**Gambar 27.** Hasil Kategorisasi Sub-kategori

Berdasarkan Gambar 26 untuk hasil kategorisasi kategori, dari 754 data yang diujikan terdapat 628 data yang berhasil terprediksi dengan sesuai, yaitu kategori *Men* sebanyak 188 data dan kategori *Women* sebanyak 440 data. Dan terdapat 126 data yang gagal terprediksi, yaitu kategori *Men* sebanyak 17 data yang masuk ke kategori *Women*, dan kategori *Women* sebanyak 109 data yang masuk ke kategori *Men*. Pengujian dengan penambahan 1 (satu) data baru tanpa diketahui kelasnya yaitu “*casual regular polyester button*” juga dilakukan dengan model *Naïve Bayes* yang telah dibuat, dan menghasilkan data tersebut masuk kedalam kategori *Women*, seperti pada Gambar 28 berikut.

```
> # test menggunakan data yang baru diisi
> apparel.k <- data.frame(documents = c("casual regular polyester button"));
> predict(k_model, apparel.k) # predict dengan 75% : 25%
[1] women
Levels: Men women
```

**Gambar 28.** Hasil Kategorisasi Kategori Data Baru

Berdasarkan Gambar 27 untuk hasil kategorisasi sub-kategori, dari 754 data yang diujikan terdapat 575 data yang berhasil terprediksi dengan sesuai, yaitu kategori *Men's Jacket* sebanyak 99 data, *Women's Jacket* sebanyak 87 data, *Men's Jeans* sebanyak 66 data, *Women's Jeans* sebanyak 181 data, *Men's Tees* sebanyak 26 data, dan *Women's Tees* sebanyak 116 data. Dan terdapat 179 data yang gagal terprediksi, yaitu kategori *Men's Jacket* sebanyak 12 data masuk ke kategori *Women's Jacket*. Kategori *Women's Jacket* sebanyak 11 data masuk ke kategori *Men's Jacket*, 1 data masuk ke kategori *Men's Tees*, dan 2 data masuk ke kategori *Women's Tees*. Kategori *Women's Jeans* sebanyak 33 data masuk ke kategori *Men's Jacket*, 38 data masuk ke kategori *Men's Jeans*, 9 data masuk ke kategori *Men's Tees*, 52 data masuk ke kategori *Women's Jacket*, dan 16 data masuk ke kategori *Women's Tees*. Kategori *Women's Tees* sebanyak 1 data masuk ke kategori *Men's Jacket*, dan 4 data masuk ke kategori *Men's Tees*. Untuk kategori *Men's Jeans* dan *Men's Tees* berhasil terprediksi seluruhnya. Pengujian dengan penambahan 1 (satu) data baru tanpa diketahui kelasnya yaitu “*casual regular polyester button*” juga dilakukan dengan model *Naïve Bayes* yang telah dibuat, dan menghasilkan data tersebut masuk kedalam kategori *Women's Jeans*, seperti pada Gambar 29 berikut.

```
> # test menggunakan data yang baru diisi
> apparel.sk <- data.frame(documents = c("casual regular polyester button"));
> predict(sk_model, apparel.sk) # predict dengan 75% : 25%
[1] women Jeans
Levels: Men Jacket Men Jeans Men Tees Women Jacket women Jeans women Tees
```

**Gambar 29.** Hasil Kategorisasi Sub-kategori Data Baru

Hasil kategorisasi kategori dan sub-kategori kemudian digabung kedalam satu *dataframe* berisi 5 (lima) atribut yaitu dokumen spesifikasi produk, kategori awal, kategori hasil prediksi, sub-kategori awal, dan sub-kategori prediksi seperti pada Gambar 13.

Dari perhitungan *confusion matrix* yang telah didapat seperti pada Tabel 14, didapatkan nilai rata-rata dari *accuracy* sebesar 90,41%, *precision* sebesar 72,1%, *recall* sebesar 88,28%, dan *F1-score* sebesar 77,14%. Sehingga dapat disimpulkan

bahwa algoritma *Naïve Bayes* dapat melakukan kategorisasi data spesifikasi produk *apparel* Aliexpress dengan cukup baik, namun dengan nilai *F1-score* yang relatif kecil dibawah 80% menandakan bahwa kemungkinan terdapat sejumlah faktor yang mempengaruhi hasil penelitian sehingga nilai *F1-score* belum maksimal, seperti keragaman dan *imbalanced data*, hasil *preprocessing* yang kurang bersih atau terdapat kata-kata penting yang terbuang selama proses pembersihan data, dan faktor lainnya.

Setelah dilakukan kategorisasi dengan menggunakan pendekatan *unsupervised learning* dan *supervised learning*, keduanya memiliki hasil yang berbeda dimana data hasil pendekatan *supervised learning* memiliki kinerja yang lebih baik dibandingkan dengan pendekatan *unsupervised learning*. Kategorisasi yang dilakukan dengan metode *Naïve Bayes* memiliki hasil berupa akurasi yang cukup baik dikarenakan *dataset* berisi data kategorikal dan sebelumnya sudah diberikan label untuk setiap data yang ada, didukung dengan kecepatan dalam pembuatan model sehingga pengelompokan menjadi lebih mudah dan memiliki hasil pengelompokan yang cukup akurat. Sementara kategorisasi yang dilakukan dengan metode LDA kurang mendapat hasil yang baik, yang dapat dipengaruhi oleh *dataset* tanpa label yang digunakan, dan dilihat dari hasil interpretasi dari distribusi kata yang terdapat pada setiap topik tidak langsung merujuk kepada suatu kategori dari produk *apparel*, melainkan hanya bisa menginterpretasi topik secara garis besar berdasarkan distribusi kata didalam topik tersebut.

Adapun keterbatasan dari penelitian ini yaitu penggunaan dataset dalam jumlah sedikit dan kategori produk yang terbatas, belum bisa melakukan pemodelan topik menggunakan LDA untuk data dengan nilai TF-IDF, kategorisasi masih dilakukan secara terpisah bukan secara hirarkis, masih banyak hasil kategorisasi yang belum terkategori dengan baik dan sesuai, dan hasil penelitian hanya dipaparkan dalam bentuk visualisasi gambar dan belum diterapkan secara dinamis seperti dalam bentuk sistem informasi.

## BAB VI KESIMPULAN

### 6.1 Kesimpulan

Penelitian menggunakan 5.700 data hasil *scraping* dari data spesifikasi produk *apparel* pada situs Aliexpress.com, dan menghasilkan 3.779 data setelah dilakukan proses pembersihan data dengan isi data meliputi 4 (empat) atribut yaitu *Style*, *Clothing Length*, *Material*, dan *Decoration*, 1 (satu) kelas untuk 2 (dua) kategori yaitu *Men* dan *Women*, dan 1 (satu) kelas untuk 6 (enam) sub-kategori yaitu *Men's Jacket*, *Women's Jacket*, *Men's Jeans*, *Women's Jeans*, *Men's Tees*, dan *Women's Tees*. Data tersebut untuk selanjutnya digunakan pada tahap pemodelan topik menggunakan *Latent Dirichlet Allocation* (LDA) dan kategorisasi dengan algoritma *Naïve Bayes*.

Dari hasil penelitian dapat disimpulkan dengan menggunakan metode LDA untuk pemodelan topik berhasil membagi dokumen kedalam 4 (empat) buah topik untuk *unigram* dengan interpretasi setiap topik yaitu mengenai gaya pakaian (*style*), salah satu jenis pakaian yaitu jaket (*jacket*), salah satu gaya pakaian yaitu *streetwear* (*streetwear*), dan salah satu gaya pakaian lainnya yaitu *gothic* (*gothic*), dan 3 (tiga) buah topik untuk *bigram* dengan interpretasi setiap topik yaitu mengenai bahan pakaian dari bulu (*fur*), bahan pakaian (*polyester*), dan panjang pakaian (*length/long*).

Kategorisasi dengan algoritma *Naïve Bayes* dilakukan dengan dua proses yaitu kategorisasi kategori berdasarkan jenis kelamin yaitu *Men* dan *Women*, dan kategorisasi sub-kategori berdasarkan jenis *apparel* yaitu *Men's Jacket*, *Women's Jacket*, *Men's Jeans*, *Women's Jeans*, *Men's Tees*, dan *Women's Tees*. Kategorisasi dilakukan dengan menguji 20% yaitu sebanyak 754 data dari 3.779 data spesifikasi produk. Kategorisasi kategori menghasilkan 628 data yang terprediksi dengan tepat, yaitu kategori *Men* sebanyak 188 data dan kategori *Women* sebanyak 440 data, dan data yang gagal terprediksi dengan tepat sebanyak 126 data. Kategorisasi sub-kategori menghasilkan 575 data yang berhasil terprediksi dengan tepat, yaitu kategori *Men's Jacket* sebanyak 99 data, *Women's Jacket* sebanyak 87 data, *Men's Jeans* sebanyak 66 data, *Women's Jeans* sebanyak 181 data, *Men's Tees* sebanyak 26 data, dan *Women's Tees* sebanyak 116 data, dan data yang gagal terprediksi dengan tepat sebanyak 179 data. Pengujian dengan data baru yang tidak diketahui kelasnya yaitu "*casual regular polyester button*" juga dilakukan untuk kategori dan sub-kategori, menghasilkan data tersebut masuk ke kategori *Women* dan sub-kategori *Women's Jeans*.

Hasil evaluasi penelitian dengan pendekatan *supervised learning* menggunakan *Confusion Matrix* memberikan nilai rata-rata untuk *accuracy* sebesar 90,41%, *precision* sebesar 72,1%, *recall* sebesar 88,28%, dan *F1-score* sebesar 77,14%. Hasil persentase *F1-Score* yang relatif rendah yaitu sebesar 77,14% dipengaruhi oleh *imbalance* data antar kelas.

Proses kategorisasi dengan 2 (dua) pendekatan yaitu *unsupervised learning* dan *supervised learning* memberikan hasil yang berbeda. Pendekatan *unsupervised learning* memiliki hasil berupa interpretasi pada setiap topik yang tidak langsung merujuk kepada suatu kategori dari produk *apparel* melainkan hanya bisa menginterpretasi isi topik secara garis besar berdasarkan distribusi kata pada topik tersebut. Sementara hasil *supervised learning* memiliki hasil yang lebih baik

dikarenakan penggunaan *dataset* yang sudah berisi data kategorikal dan memiliki label pada setiap data, sehingga pengelompokan menjadi lebih mudah dan memiliki hasil yang cukup akurat dilihat dari presentase evaluasi model menggunakan *confusion matrix*.

## 6.2 Saran

Dalam penelitian Pemodelan Kategorisasi Produk *Apparel* Pada *Marketplace* Dengan *Machine Learning* ini masih ada sejumlah kekurangan yang diharapkan dapat diperbaiki dan dikembangkan pada penelitian selanjutnya, sebagai berikut:

1. Dapat menggunakan objek penelitian dari kategori lain menggunakan deskripsi setiap produk, atau objek penelitian dari situs *marketplace* lainnya.
2. Pada tahap *preprocessing* sebaiknya diperhatikan kata-kata yang terbuang pada saat proses pembersihan data menggunakan *stopword list*, karena dapat mempengaruhi hasil penelitian.
3. Penggunaan dataset dalam jumlah banyak dan dengan isi dokumen yang panjang yang dapat memberikan keragaman pada hasil pemodelan topik dan kategorisasi.
4. Menggunakan metode dan algoritma lainnya untuk pemodelan topik dan kategorisasi.
5. Kategorisasi secara kompleks dan mendetail untuk pengembangan penelitian selanjutnya bisa dicoba dengan melakukan kategorisasi hirarkis *multi-level*.
6. Penggabungan dengan citra untuk kategorisasi lebih baik.
7. Penerapan penelitian kedalam sebuah sistem informasi menggunakan *software* seperti *website*, aplikasi *desktop*, maupun aplikasi *mobile*.




## DAFTAR PUSTAKA

- Aco, A., & Endang, A. H. (2017). Analisis Bisnis E-Commerce pada Mahasiswa Universitas Islam Negeri Alauddin Makassar. *Jurnal INSYPRO*, 2.
- Al-khairi, Y. U., Wibisono, Y., & Putro, B. L. (2018). Deteksi Topik Fashion Pada Twitter Dengan Latent Dirichlet Allocation. *Jurnal Aplikasi dan Teori Ilmu Komputer (JATIKOM)*, 1, 1-8.
- Annur, H. (2018). Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes. *ILKOM Jurnal Ilmiah*, 160-165.
- Cevahir, A., & Murakami, K. (2016). Large-scale Multi-class and Hierarchical Product Categorization for an E-commerce Giant. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, (pp. 525-535).
- Daniel, D. R., & Zuhri, A. S. (2014, April 2014). Perancangan Expert Sytem Untuk Klaifikasi Barang Pada Kantor Bea Cukai Tanjung Perak Surabaya. *Jurnal Ekonomi dan Bisnis*, 17-31.
- Fitrianto, E. P. (2017). *Pemeringkatan Perguruan Tinggi Menggunakan Metode Labeled Latent Dirichlet Allocation (L-LDA) Untuk Mengukur Tingkat Kesiapterapan Teknologi Perguruan Tinggi di Indonesia*. Jurusan Teknik Informatika, Fakultas Teknologi Informasi. Surabaya: Institut Teknologi Sepuluh Nopember.
- Gottipati, S., & Vauhkonen, M. (2016). E-Commerce Product Categorization. Stanford University.
- Herwanto, G. B. ( 2018, September). DOCUMENT CLUSTERING DENGAN LATENT DIRICHLET ALLOCATION DAN WARD HIERARICHAL CLUSTERING. *Jurnal Pseudocode*, V, 29 - 37.
- Likhitha, S., Harish, B. S., & Kumar, H. M. (2019, Agustus). A Detailed Survey on Topic Modelling for Document and Short Text Data. *International Journal of Computer Applications*, 178, 1-9.
- Nurhayati, Busman, & Iswara, R. P. (2019). Pengembangan Algoritma Unsupervised Learning Technique Pada Big Data Analysis Di Media Sosial Sebagai Media Promosi Online Bagi Masyarakat. *Jurnal Teknik Informatika*, 12 No. 1, 79 - 96.
- Prihatini, P. M., Suryawan, I. K., & Mandia, I. N. (2017). Metode Latent Dirichlet Allocation Untuk Ekstraksi Topik Dokumen. *Jurnal Rancang Bangun dan Teknologi (LOGIC)*, 153-157.


- Puspitasari, A. M., Ratnawati, D. E., & Widodo, A. W. (2018, Februari). Klasifikasi Penyakit Gigi dan Mulut Menggunakan Metode Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2, 802-810.
- Putra, I. K., & Kusumawardani, R. P. (2017). Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA). *Jurnal Teknik ITS*, 6, 2337-3520.
- Raharjo, S., & Winarko, E. (2014). Klasterisasi, Klasifikasi, dan Peringkasan Teks Berbahasa Indonesia. *Prosiding Seminar Ilmiah Nasional Komputer dan Sistem Intelijen (KOMMIT)*, 8, pp. 391-401. Depok.
- Shaik, I., Nittela, S. S., Hiwarkar, T., & Nalla, S. (2019, September). K-means Clustering Algorithm Based on E-Commerce Big Data. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8(11), 1910-1914.
- Siringoringo, R., & Jamaluddin. (2019, April). Text Mining dan Klasterisasi Sentimen Pada Ulasan Produk Toko Online. *Jurnal Penelitian Teknik Informatika Universitas Prima Indonesia (UNPRI)*, 2.
- Taufik, A. (2018). Komparasi Algoritma Text Mining Untuk Klasifikasi Review Hotel. *Jurnal Teknik Komputer AMIK BSI*, IV, 112-118.
- Waasiu, A., B, A. I., & Lawi, A. (2021). Klasifikasi Audio Cats andDogsMenggunakan Model Artifical Neural Network Multi-perceptron. *Konferensi Nasional Ilmu Komputer (KONIK)*, (p. 58). Makassar.

# **LAMPIRAN**

## Lampiran 1. Surat Keputusan (SK) Tugas Akhir



YAYASAN PAKUAN SILIWANGI  
**Universitas Pakuan**  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
*Keagungan, Keindahan & Keberagaman Dalam Cita-cita NECTA*



---

**SURAT KEPUTUSAN DEKAN**  
**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**  
**UNIVERSITAS PAKUAN**  
**NO : 522 /SK/D/FMIPA/II/2020**

**TENTANG**

**PENGGANGKATAN PEMBIMBING TUGAS AKHIR**  
**PADA PROGRAM STUDI ILMU KOMPUTER**  
**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**  
**UNIVERSITAS PAKUAN**  
**DEKAN FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**  
**UNIVERSITAS PAKUAN**

**Menimbang** : 1. bahwa setiap mahasiswa tingkat akhir Program Strata Satu (SI) harus melaksanakan Tugas Akhir sebagaimana tercantum dalam kurikulum setiap Jurusan di lingkungan Fakultas MIPA Universitas Pakuan.  
2. bahwa untuk pelaksanaan Tugas Akhir diperlukan pengawasan dari pembimbing.  
3. bahwa sehubungan dengan point 1 dan 2 di atas perlu dituangkan ke dalam suatu keputusan dekan.

**Meningat** : 1. Undang - Undang RI No. 2 Tahun 2003 tentang Sistem Pendidikan Nasional.  
2. Peraturan Pemerintah No. 60 Tahun 1999 tentang Pendidikan Tinggi.  
3. Statuta Universitas Pakuan Tahun 2017.  
4. Keputusan Rektor Universitas Pakuan Nomor : 067 / KEP / REK / VIII / 2015 tentang Pemberhentian dan Pengangkatan Dekan Fakultas di lingkungan Universitas Pakuan Masa Bakti 2015 - 2020.  
5. Ketentuan Akademik yang tercantum dalam Buku Panduan Studi Fakultas MIPA - Universitas Pakuan Tahun 2018.

**Memperhatikan** : Usulan dari Ketua Program Studi Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Pakuan.

**MEMUTUSKAN**

**Menetapkan** :  
**Pertama** : Mengangkat pembimbing yang namanya tersebut di bawah ini :

1. Pembimbing Utama : Prihastuti Harsani, M.Si.
2. Pembimbing Pendamping : Sufiatul Maryana, S.Kom., M.Kom.

Untuk membimbing dalam rangka melaksanakan tugas akhir bagi

Nama : Mega Inez Syafira  
NPM : 065116051  
Program Studi : Ilmu Komputer  
Judul Skripsi : Kategorisasi Otomatis Produk Apparel Pada Marketplace

Jl. Pakuan P.O. Box 452, Bogor 16143, Telp./Fax. (0251) 8375547  
Website : <http://www.fmipa-unpak.net>

- Kedua : Kepada para pembimbing diharapkan dapat menjalankan tugasnya pembimbing dengan sebaik-baiknya.
- Ketiga : Dalam waktu 1 (satu) bulan setelah diterbitkan SK ini, mahasiswa wajib melaksanakan Seminar Rencana Penelitian yang diselenggarakan oleh jurusan dengan dihadiri oleh pembimbing dan penguji.
- Keempat : Dana untuk honorarium pembimbing dibebankan kepada mahasiswa yang ketentuannya diatur oleh Fakultas MIPA.
- Kelima : Surat Keputusan ini berlaku untuk jangka waktu 1 (satu) tahun sejak tanggal ditetapkan sampai dengan mahasiswa tersebut lulus sidang skripsi dengan ketentuan akan diadakan perubahan / perbaikan sebagaimana mestinya bila di kemudian hari terdapat kekeliruan dalam penetapannya.

Ditetapkan di : Bogor  
Pada tanggal : 03 Maret 2020



Dr. Prasetyorini, MS.

TEMBUSAN :

1. Yth. Pembantu Rektor I
2. Yth. Ketua Program Studi Ilmu Komputer
3. Yth. Prihastuti Harsani, M.Si.
4. Yth. Sufiatul Maryana, S.Kom., M.Kom.
5. Arsip

## Lampiran 2. Hasil Data Scraping Awal

	Title	Brand.Name	Origin	Type	Model.Number	Style	Closure.Type	Sleeve.Length.cm.	
0	Winter Warm Men Jacket Coat Casual Autumn Stand Collar ...	UNION ARMY	CN(Origin)	Regular	Detachable hat blac...	Casual	zipper	Full	Zippers
1	MONERFFI Autumn Men's PU Leather Jacket For Men Fitness...	MoneRFFI	CN(Origin)	Regular	Men Faux Leather Ja...	Casual	zipper	Full	None
2	Men's Real Leather Jacket Men Motorcycle Removable Hood...	FLAVOR		Regular	M2016-95	Casual	zipper	Full	Zippers
3	New Spring Autumn Mens Trench Coat Jacket Plus Size 4XL ...	Dihope	CN(Origin)			Casual		Full	
4	Mens Jackets and coats Male Autumn New Fashion Zipper b...	Diarmuid	CN(Origin)	Slim	W10	Casual	zipper		Embroidery
5	New Brand Autumn Winter Light Down Jacket Men's Fashio...	TECHOME	CN(Origin)	Regular	DC-4	Casual	Zipper	Full	Pockets
6	Men Vintage Motorcycle Jacket 2020 Men's Bomber Fleece ...			Regular	Men Motorcycle Jac...	Casual	zipper	Full	Zippers
7	Men Genuine Leather Jacket Real Goat skin Leather Suits Ca...	Dusen Klein		Slim	106	Casual	Single Breasted	Full	None
8	Men's Embroidery Leather Jacket 2020 New Men Stand Coll...		US(Origin)	Regular	men jacketmen leat...	Casual	zipper	Full	Embroidery
9	CW New Brand Jacket Men Zipper Spring Autumn Casual H...								
10	2020 New Clothing Down Winter Jackets Business Long Win...	Josephopro		Regular	guo	Casual	zipper	Full	Zippers
11	Leather Men's spring and autumn new men's business casu...		CN(Origin)	Wide-walsted	Light business		zipper	Full	Zippers
12	Privathinker Men's Oversized Windbreaker Coats 2020 Autu...	PRIVATHINKER	CN(Origin)	Loose	Men Coat	casual /Harajuku /V...	Single Breasted	Full	Embroidery
13	HOO 2020 autumn men's youth fashion trend leather jacket...	boussac		Wide-walsted		Casual		Full	None
14	spring and summer new pilot jacket men and women casual...			Loose			Open Strch		Pockets
15	Free shipping/New arrival.winter warm cotton genuine leath...	VANLED	CN(Origin)	Regular		Casual	zipper	Full	Zippers
16	Mens Vintage Motorcycle Jacket 2020 Men Fashion New Bik...		US(Origin)	Regular	men jacketmen bik...	Casual	zipper	Full	Zippers
17	1232 Big USA Size Air Force Flight Suit G1 Pilot Thick Warm ...		CN(Origin)	Regular	1232	Casual	zipper	Full	Cross-Cross
18	2020 New Business Men Trench Coat Fashion Design Slim D...		CN(Origin)	Regular	Jackets	Casual	Single Breasted	Full	Button
19	Fashionable men's and women trench coat is irregular, asym...	3Wphtato		Loose	0129	casual	Single Breasted	Full	Button
20	WHYWORKS tactical cloak outdoor windproof cloak water r...	whyworks		Loose		Punk Style		Sleeveless	Spliced

Showing 1 to 21 of 960 entries, 761 total columns

	Title	Brand.Name	Origin	Type	Model.Number	Style	Closure.Type	Sleeve.Length.cm.	
20	WHYWORKS tactical cloak outdoor windproof cloak water r...	whyworks		Loose		Punk Style		Sleeveless	Spliced
21	ICEbear 2020 New men's trench coat high-quality men's lon...	ICEbear	CN(Origin)	Regular	MWF20709D	Casual	Double Breasted	Full	Pockets
22	2020 New Men Solid Color Long-sleeved Coat Business Gen...		CN(Origin)	Slim		Casual	Single Breasted	Full	None
23	Russia 90% white duck down long jackets men Winter long ...	Brieuces		Regular	PIN7805	Casual	zipper	Full	Pockets
24	Autumn and Winter New Fleece Patchwork Men's Jacket ...	QRWR	CN(Origin)	Loose	MYD392	Casual	zipper		Spliced
25	90%Down Jackets Men Winter Jacket Men Fashion Thick Wa...	JUNGLE ZONE		Regular	877	Hooded down jacket	zipper	Full	Fur
26	New Men's Biker Leather Jacket Fur Collar Detachable Faux ...	AKSR		Regular		Casual	zipper	Full	Fur
27	Spring Autumn Motorcycle Leather Jacket Men Slim Fit Obli...			Motorcycle leather...	Pu leather-01	Casual	zipper	Full	Draped
28	Winter Men Men's Down Jacket Real Sheepskin Leather Jack...			Slim	1926	Casual	zipper	Full	Pockets
29	JiUBL Mens Leather Jackets Motorcycle Stand Collar Zipper ...	JiUBL	CN(Origin)	Regular	HMT-CP1923	Moto & Biker	zipper		Fur
30	New Men's Leather Jacket Brand Slim Fit Motorcycle Leather...	DOYU		Slim	PY26	Moto & Biker	zipper	Full	Zippers
31	2020 Autumn Winter Plus Size Men's Leather Jacket Vintage...		US(Origin)	Regular	men leather jacket...	Casual	zipper	Full	Zippers
32	Leather Jacket Men Winter fleece Motorcycle PU Leather Jac...	qiqichen		Regular	6693 leather jacket ...	Casual Fashion Mill...	zipper	Full	Pockets
33	large size 7XL 8XL suede casaco Men's Real Leather Jacket ...			Loose	large size 7XL 8XL s...	Casual	zipper	Full	
34	Punk MJ Michael Jackson Germany Military Cool Georgeous...			Slim	2554	England Style	zipper		Sequined
35	Autumn Jacket Men New Slim Retro Winter Jackets Male PU...		CN(Origin)	Regular	S27713	Casual	zipper	Full	Pockets
36	Men's White Duck Down Jacket Warm Hooded Thick Puffer ...	UNION ARMY		Regular		Casual	zipper	Sleeveless	None
37	Vintage Medieval Robe Cosplay Costume Vintage Men's Tre...	Elijevans		Regular	1015	vintage	Single Breasted	Full	None
38	2018 Brand Autumn Spring Casual Zipper Leather Jacket Mo...	NIBESSER		Slim	Men Jacket	Hip Hop	zipper	Full	None
39	New men's casual hooded bomber jacket spring and autumn...		CN(Origin)	Slim	JK70	Casual	zipper		Pockets
40	Page Not Found - Aliexpress.com								

Showing 21 to 41 of 960 entries, 761 total columns



## Lampiran 4. Hasil *Preprocessing* Data *Rstudio*

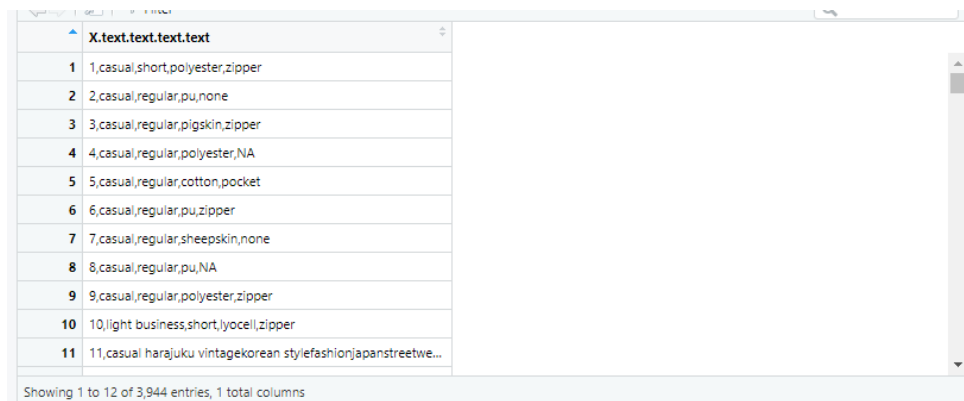
A) Sebelum pembersihan dan transformasi data:



	Style	Clothing.Length	Material	Decoration	Class
6	Casual	Regular	PU	Zippers	Men Jacket
7	Casual	Regular	Sheepskin	None	Men Jacket
8	Casual	Regular	PU	Embroidery	Men Jacket
9	Casual	Regular	Polyester	Zippers	Men Jacket
10	Light business	Short	Lyocell	Zippers	Men Jacket
11	casual /Harajuku /Vintage/Korean Style/Fashion/Japan/stree...	long	Polyester	Embroidery	Men Jacket
12	Casual	Regular	Sheepskin	Zippers	Men Jacket
13	Casual	Regular	PU	Zippers	Men Jacket
14	Casual	Short	Genuine Leather, Cowskin	Criss-Cross	Men Jacket
15	Casual	X-Long	Polyester	Button	Men Jacket
16	casual	long	Cotton	Button	Men Jacket

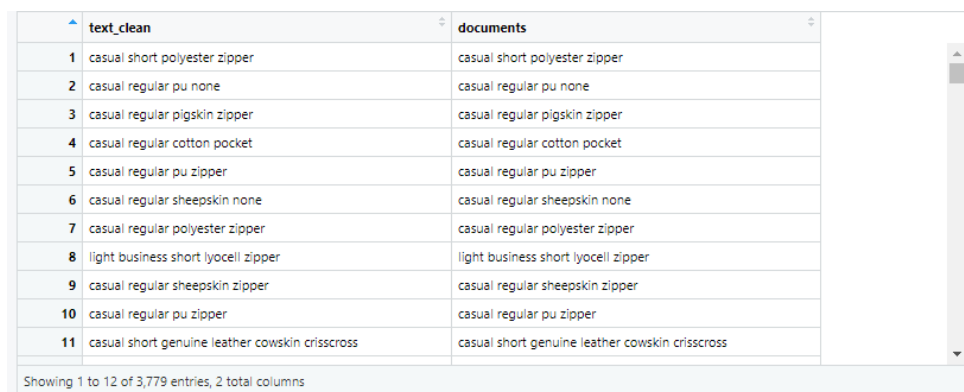
Showing 6 to 17 of 3,944 entries, 5 total columns

B) Sesudah pembersihan dan transformasi data:



	X.text.text.text.text
1	1,casual,short,polyester,zipper
2	2,casual,regular,pu,none
3	3,casual,regular,pigskin,zipper
4	4,casual,regular,polyester,NA
5	5,casual,regular,cotton,pocket
6	6,casual,regular,pu,zipper
7	7,casual,regular,sheepskin,none
8	8,casual,regular,pu,NA
9	9,casual,regular,polyester,zipper
10	10,light business,short,lyocell,zipper
11	11,casual harajuku vintagekorean stylefashionjapanstreetwe...

Showing 1 to 12 of 3,944 entries, 1 total columns



	text_clean	documents
1	casual short polyester zipper	casual short polyester zipper
2	casual regular pu none	casual regular pu none
3	casual regular pigskin zipper	casual regular pigskin zipper
4	casual regular cotton pocket	casual regular cotton pocket
5	casual regular pu zipper	casual regular pu zipper
6	casual regular sheepskin none	casual regular sheepskin none
7	casual regular polyester zipper	casual regular polyester zipper
8	light business short lyocell zipper	light business short lyocell zipper
9	casual regular sheepskin zipper	casual regular sheepskin zipper
10	casual regular pu zipper	casual regular pu zipper
11	casual short genuine leather cowskin crisscross	casual short genuine leather cowskin crisscross

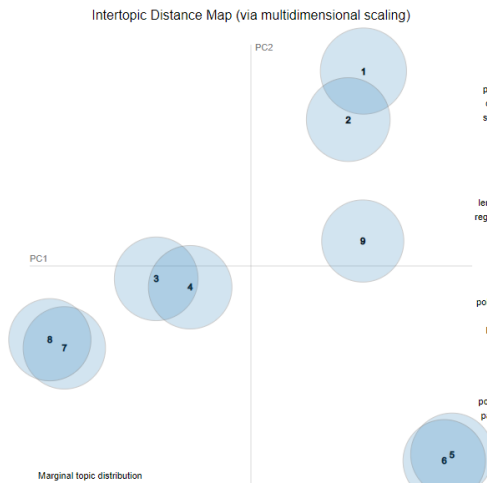
Showing 1 to 12 of 3,779 entries, 2 total columns



Lampiran 5. Hasil Pemodelan Topik *Unigram* dan *Bigram*

	Plot	Label
<b>Unigram (7 Topik)</b>	<p>Intertopic Distance Map (via multidimensional scaling)</p> <p>PC1</p> <p>PC2</p> <p>Marginal topic distribution</p>	<p>Topic Label</p> <ul style="list-style-type: none"> <li>1 - length, pants, vintage, ankle, spandex, streetwear, office, washed, ripped, lace</li> <li>2 - fur, long, thick, fashion, warm, faux, pocket, regular, fox, spliced</li> <li>3 - full, length, denim, hole, hop, tencel, hip, club, fake, scratched</li> <li>4 - cotton, full, street, style, length, button, high, appliques, england, print</li> <li>5 - casual, short, polyester, none, modal, hollow, hip, cashmere, bleached, half</li> <li>6 - casual, none, polyester, short, button, acetate, pocket, zipper, blends, tassel</li> <li>7 - regular, pocket, zipper, polyester, casual, leather, biker, moto, cow, sheepskin</li> </ul>
<b>Unigram (3 Topik)</b>	<p>Intertopic Distance Map (via multidimensional scaling)</p> <p>PC1</p> <p>PC2</p> <p>Marginal topic distribution</p>	<p>Topic Label</p> <ul style="list-style-type: none"> <li>1 - length, full, pants, denim, vintage, ankle, spandex, street, style, streetwear</li> <li>2 - casual, none, short, polyester, cotton, button, appliques, club, modal, gothic</li> <li>3 - regular, pocket, polyester, fur, long, casual, zipper, fashion, warm, thick</li> </ul>
<b>Bigram (5 Topik)</b>	<p>Intertopic Distance Map (via multidimensional scaling)</p> <p>PC1</p> <p>PC2</p> <p>Marginal topic distribution</p>	<p>Topic Label</p> <ul style="list-style-type: none"> <li>1 - full length, casual full, length denim, length polyester, hip hop, denim pocket, length tencel, hop full, denim none, denim hole</li> <li>2 - casual regular, regular polyester, casual long, polyester pocket, long polyester, polyester zipper, style full, polyester button, regular pu, club full</li> <li>3 - polyester none, short polyester, thick warm, warm fur, faux fur, casual full, fur regular, fur long, regular faux, fur short</li> <li>4 - casual short, short cotton, cotton none, moto biker, cow leather, spandex none, smart casual, short spandex, regular cow, cotton x</li> <li>5 - length pants, ankle length, length cotton, casual ankle, high street, pants cotton, pants polyester, cotton pocket, street full, polyester vintage</li> </ul>

**Bigram (9  
Topik)**



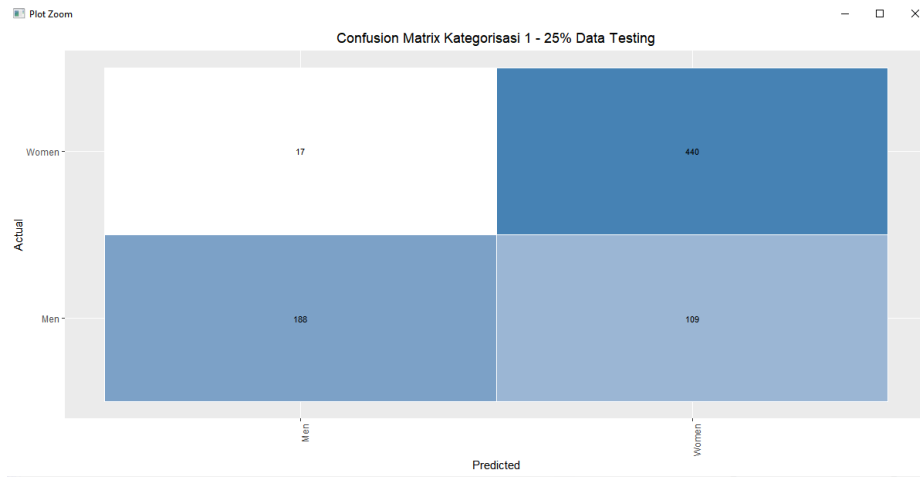
- 1 - full length, length denim, casual full, hip hop, denim pocket, hop full, denim none, length tencel, denim hole, tencel pocket
- 2 - full length, casual full, length polyester, high street, street full, length spandex, fashion full, polyester ripped, office full, smart casual
- 3 - casual regular, regular polyester, polyester pocket, polyester zipper, polyester button, regular pu, regular nylon, regular cotton, pu pocket, casual xlong
- 4 - casual short, short cotton, cotton none, cotton x, full cotton, short modal, modal none, offset print, cotton offset, patchwork fur
- 5 - thick warm, warm fur, faux fur, fur regular, fur long, regular faux, fur short, long faux, fashion slim, fur none
- 6 - polyester none, short polyester, casual short, spandex none, short spandex, streetwear full, full polyester, full spandex, club full, polyester appliques
- 7 - casual full, moto biker, cow leather, smart casual, regular cow, leather pocket, biker regular, biker full, regular sheepskin, england style
- 8 - length cotton, casual long, long polyester, full length, style full, cotton pocket, street style, cotton button, cotton vintage, office full
- 9 - length pants, ankle length, casual ankle, pants cotton, pants polyester, polyester

## Lampiran 6. Hasil Visualisasi Kategorisasi

### A) Kategori: *Men* dan *Women*

```
> # pembagian data actual & predicted
> k.cm.test <- confusionMatrix(table(ky_test, ky_predict, dnn = c("Actual","Predicted")));
> k.cm.test
Confusion Matrix and Statistics
```

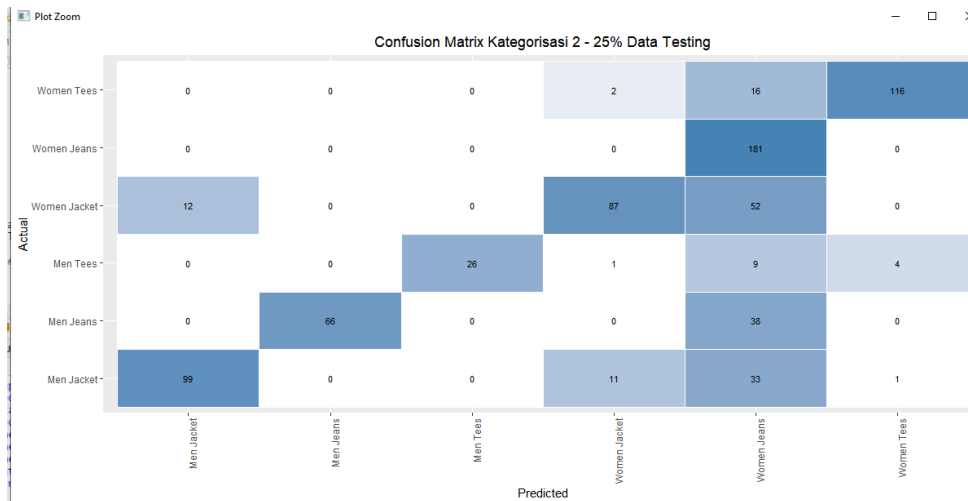
	Predicted	
Actual	Men	Women
Men	188	109
Women	17	440



### B) Sub-kategori: *Men's Jacket*, *Women's Jacket*, *Men's Jeans*, *Women's Jeans*, *Men's Tees*, dan *Women's Tees*

```
> # pembagian data actual & predicted
> sk.cm.test <- confusionMatrix(table(sky_test, sky_predict, dnn = c("Actual","Predicted")));
> sk.cm.test
Confusion Matrix and Statistics
```

	Predicted					
Actual	Men Jacket	Men Jeans	Men Tees	Women Jacket	Women Jeans	Women Tees
Men Jacket	99	0	0	11	33	1
Men Jeans	0	66	0	0	38	0
Men Tees	0	0	26	1	9	4
Women Jacket	12	0	0	87	52	0
Women Jeans	0	0	0	0	181	0
Women Tees	0	0	0	2	16	116



## Lampiran 7. Hasil Kategorisasi *Actual* dan *Predicted*

	name	actual_category	predict_category	actual_sub_category	predict_sub_category
1	vintage regular sheepskin none	Men	Men	Women Jeans	Women Jeans
2	double faced fur regular raccoon dog	Women	Women	Women Jeans	Women Jeans
3	casual regular polyester none	Women	Men	Women Jeans	Women Jeans
4	streetwear full length spandex hole	Women	Women	Women Jacket	Women Jacket
5	infation hip hop men jeans full length cotton polyester poc...	Men	Women	Men Jacket	Men Jacket
6	casual full length denim pocket	Men	Men	Women Jeans	Women Jeans
7	casual regular polyester pattern	Men	Men	Men Jeans	Women Jeans
8	casual regular pu none	Men	Men	Men Jeans	Men Jeans
9	smart casual short cotton button	Men	Men	Men Jacket	Men Jacket
10	streetwear full length polyester tie dye	Women	Women	Women Jeans	Women Jeans
11	casual short polyester button	Men	Women	Men Jacket	Men Jacket
12	smart casual full length tencel none	Men	Men	Men Tees	Men Tees
13	patchwork fur long faux fur spliced	Women	Women	Women Jeans	Women Jeans
14	casual regular pu pocket	Men	Men	Men Jeans	Men Jeans
15	office ankle length pants microfiber fake zipper	Women	Women	Men Tees	Men Tees
16	hip hop regular acetate none	Men	Women	Men Jacket	Women Jacket
17	casual short cotton none	Men	Women	Women Jacket	Women Jeans
18	style long polyester none	Men	Men	Men Jacket	Women Jeans
19	vintage patchwork track coat regular polyester pocket	Men	Men	Men Jeans	Men Jeans
20	casual regular polyester zipper	Men	Women	Women Jacket	Women Jacket
21	fashion slim long faux fur flocking	Women	Women	Men Jacket	Women Jeans

Showing 1 to 21 of 754 entries, 5 total columns

	name	actual_category	predict_category	actual_sub_category	predict_sub_category
21	fashion slim long faux fur flocking	Women	Women	Men Jacket	Women Jeans
22	office full length cotton pocket	Women	Women	Women Jeans	Women Jeans
23	vintage full length cotton pocket	Women	Men	Men Tees	Men Tees
24	casual short modal none	Women	Women	Women Jacket	Women Jeans
25	casual regular cotton none	Men	Men	Women Jacket	Women Jacket
26	club full cotton none	Women	Women	Men Jacket	Men Jacket
27	casual regular polyester zipper	Men	Women	Men Jeans	Men Jeans
28	vintage full length polyester none	Women	Women	Women Jeans	Women Jeans
29	streetwear long polyester zipper	Women	Women	Women Tees	Women Tees
30	spring autumn winter undershirt half cotton none	Women	Women	Women Jacket	Women Jacket
31	high street full length denim none	Men	Men	Men Tees	Men Tees
32	vintage full length spandex pocket	Women	Women	Women Tees	Women Tees
33	casual short cotton none	Women	Women	Men Tees	Women Jeans
34	casual full length polyester bleached	Women	Women	Men Jeans	Men Jeans
35	punk style full length cotton fake zipper	Men	Women	Women Jacket	Women Jacket
36	casual short cotton none	Women	Women	Men Jacket	Men Jacket
37	casual full length spandex none	Women	Women	Men Jeans	Men Jeans
38	elegant fashion full length cotton vintage	Women	Women	Women Jacket	Women Jacket
39	casual short polyester zipper	Women	Women	Women Jacket	Women Jacket
40	casual ankle length pants cotton fake zipper	Women	Women	Men Jacket	Women Jeans
41	casual short cow leather pocket	Men	Men	Men Jacket	Men Jacket

Showing 21 to 41 of 754 entries, 5 total columns

## Lampiran 8. Hasil Evaluasi Kategori dan Sub-kategori

### A) Kategori

```

                Predicted
Actual  Men Women
Men    188  109
Women  17   440

Accuracy : 0.8329
95% CI : (0.8043, 0.8588)
No Information Rate : 0.7281
P-value [Acc > NIR] : 7.689e-12

Kappa : 0.63

McNemar's Test P-value : 5.192e-16

Sensitivity : 0.9171
Specificity : 0.8015
Pos Pred Value : 0.6330
Neg Pred Value : 0.9628
Prevalence : 0.2719
Detection Rate : 0.2493
Detection Prevalence : 0.3939
Balanced Accuracy : 0.8593

'Positive' Class : Men

```

### B) Sub-kategori

```

                Predicted
Actual  Men Jacket Men Jeans Men Tees Women Jacket Women Jeans Women Tees
Men Jacket    99         0         0         11         33         1
Men Jeans     0         66         0         0         38         0
Men Tees      0         0        26         1         9         4
Women Jacket  12         0         0         87         52         0
Women Jeans   0         0         0         0         181        0
Women Tees    0         0         0         2         16        116

Overall Statistics

Accuracy : 0.7626
95% CI : (0.7306, 0.7926)
No Information Rate : 0.4363
P-value [Acc > NIR] : < 2.2e-16

Kappa : 0.7025

McNemar's Test P-value : NA

Statistics by Class:

                Class: Men Jacket Class: Men Jeans Class: Men Tees
Sensitivity          0.8919          1.00000          1.00000
Specificity          0.9300          0.94477          0.98077
Pos Pred Value       0.6875          0.63462          0.65000
Neg Pred Value       0.9803          1.00000          1.00000
Prevalence           0.1472          0.08753          0.03448
Detection Rate       0.1313          0.08753          0.03448
Detection Prevalence 0.1910          0.13793          0.05305
Balanced Accuracy    0.9110          0.97238          0.99038

                Class: women Jacket Class: women Jeans Class: women Tees
Sensitivity          0.8614          0.5502          0.9587
Specificity          0.9020          1.0000          0.9716
Pos Pred Value       0.5762          1.0000          0.8657
Neg Pred Value       0.9768          0.7417          0.9919
Prevalence           0.1340          0.4363          0.1605

                Detection Rate          0.1154          0.2401          0.1538
                Detection Prevalence     0.2003          0.2401          0.1777
                Balanced Accuracy        0.8817          0.7751          0.9651
> |

```

**Lampiran 9.** Perbandingan Data Latih dan Data Uji

	Nama Kategori	Jumlah Data	Data Latih (80%)	Data Uji (20%)
<b>Kategori</b>	<i>Men</i>	1.489 (39,4%)	1.192 (39,4%)	297 (39,4%)
	<i>Women</i>	2.290 (60,6%)	1.833 (60,6%)	457 (60,6%)
<b>Jumlah</b>		<b>3.779</b> <b>(100%)</b>	<b>3.025</b> <b>(100%)</b>	<b>754</b> <b>(100%)</b>
<b>Sub-kategori</b>	<i>Men's Jacket</i>	717 (18,97%)	573 (18,94%)	144 (19,1%)
	<i>Women's Jacket</i>	738 (19,53%)	587 (19,41%)	151 (20%)
	<i>Men's Jeans</i>	528 (13,97%)	424 (14,02%)	104 (13,8%)
	<i>Women's Jeans</i>	890 (23,55%)	709 (23,44%)	181 (24%)
	<i>Men's Tees</i>	244 (6,46%)	204 (6,75%)	40 (5,3%)
	<i>Women's Tees</i>	662 (17,52%)	528 (17,44%)	134 (17,8%)
<b>Jumlah</b>		<b>3.779</b> <b>(100%)</b>	<b>3.025</b> <b>(100%)</b>	<b>754</b> <b>(100%)</b>

**Lampiran 10. Stopword List Bahasa Inggris**

I	Has	Wouldn't	Up
Me	Had	Shan't	Down
Myself	Having	Shouldn't	In
We	Do	Can't	Out
Our	Does	Cannot	On
Ours	Did	Couldn't	Off
Ourselves	Doing	Mustn't	Over
You	Would	Let's	Under
Your	Should	That's	Again
Yours	Could	Who's	Further
Yourself	Ought	What's	Then
Yourselves	I'm	Here's	Once
He	You're	There's	Here
Him	He's	When's	There
His	She's	Where's	When
Himself	It's	Why's	Where
She	We're	How's	Why
Her	They're	A	How
Hers	I've	An	All
Herself	You've	The	Any
It	We've	And	Both
Its	They've	But	Each
Itself	I'd	If	Few
They	You'd	Or	More
Them	He'd	Because	Most
Their	She'd	As	Other
Theirs	We'd	Until	Some
Themselves	They'd	While	Such
What	I'll	Of	No
Which	You'll	At	Nor
Who	He'll	Bye	Not
Whom	She'll	For	Only
This	We'll	With	Own
That	They'll	About	Same
These	Isn't	Against	So
Those	Aren't	Between	Then
Am	Wasn't	Into	Too
Is	Weren't	Through	Very
Are	Hasn't	During	
Was	Haven't	Before	
Were	Hadn't	After	
Be	Doesn't	Above	
Been	Don't	Below	
Being	Didn't	To	
Have	Won't	From	

### Lampiran 11. Perhitungan Manual TF-IDF

Penentuan dokumen yang didapat pada beberapa spesifikasi produk dicontohkan sebagai berikut :

**Dokumen 1 (D1):** casual short polyester zipper

**Dokumen 2 (D2):** casual regular pu none

**Dokumen 3 (D3):** vintage long polyester none

**Dokumen 4 (D4):** sports regular polyester spliced

Selanjutnya menentukan nilai df, IDF, dan bobot dari setiap dokumen yang dapat dilihat pada persamaan (6), (7), dan (8). Hasil pembobotan setiap kata menggunakan proses TF-IDF ditampilkan pada tabel berikut.

$$df = \text{jumlah } tf \quad (6)$$

$$IDF = \log \frac{D}{df} \quad (7)$$

$$W = IDF * tf \quad (8)$$

Tabel Perhitungan df *unigram*

BoW	tf				df
	D1	D2	D3	D4	
Casual	1	1			2
Short		1			1
Polyester	1		1	1	3
Zipper				1	1

Tabel Perhitungan IDF dan W *unigram*

BoW	Df	D/df	IDF	Bobot (W)			
				D1	D2	D3	D4
Casual	2	2	0,3010	<b>0,3010</b>	<b>0,3010</b>		
Short	1	4	0,6021		<b>0,6021</b>		
Polyester	3	1.5	0,1761	<b>0,1761</b>		<b>0,1761</b>	<b>0,1761</b>
Zipper	1	4	0,6021				<b>0,6021</b>

Tabel Perhitungan df *bigram*

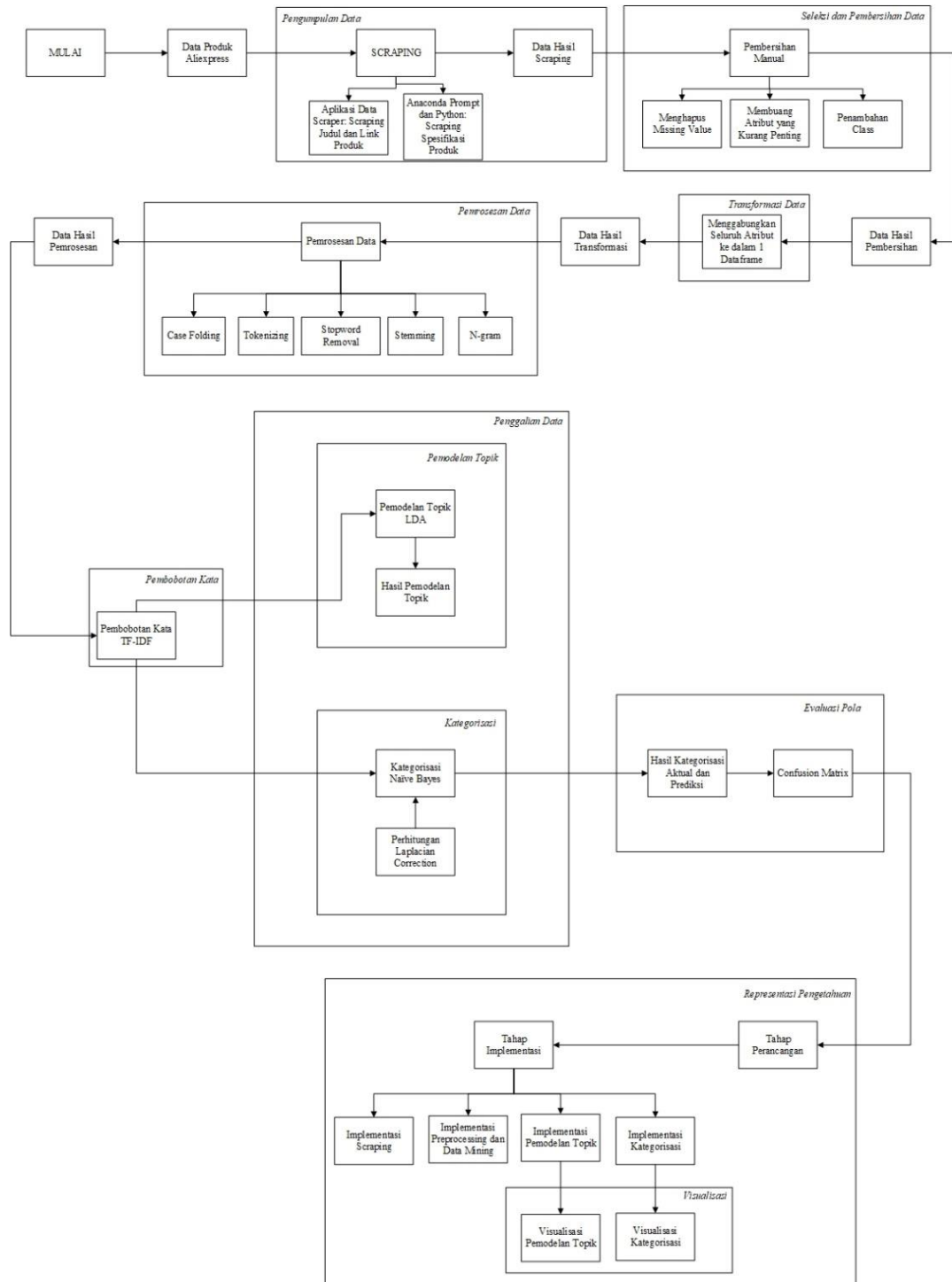
BoW	tf				df
	D1	D2	D3	D4	
Casual short	1				1
Short polyester	1				1
Polyester zipper	1				1

Tabel Perhitungan IDF dan W *bigram*

BoW	Df	D/df	IDF	Bobot (W)			
				D1	D2	D3	D4
Casual short	1	4	0,6021	<b>0,6021</b>			
Short polyester	1	4	0,6021	<b>0,6021</b>			
Polyester zipper	1	4	0,6021	<b>0,6021</b>			



## Lampiran 12. Analisis Kebutuhan Pengerjaan Sistem



### Lampiran 13. Tahap Perancangan Sistem

