Ekstraksi Informasi Transaksi *Online* pada Twitter

Masayu Leylia Khodra
Institut Teknologi Bandung
masayu@stei.itb.ac.id

Ayu Purwarianti
Institut Teknologi Bandung
ayu@stei.itb.ac.id

ABSTRAK

Penelitian ini mengeksplorasi analisis konten tweet untuk mendapatkan informasi transaksi *online* di Indonesia yang datanya masih minim saat ini. Untuk itu, aplikasi SaFE-F dikembangkan yang melakukan pencarian (***S****earch*) dan **F**ilter tweet yang relevan, **E**kstraksi informasi transaksi online, dan menyimpan hasil ekstraksinya (***F****illing*). Dengan menggunakan pendekatan ekstraksi informasi berbasis klasifikasi, dilakukan klasifikasi tweet dan klasifikasi token. Oleh karena itu, korpus tweet bahasa Indonesia dikonstruksi untuk pembangunan model klasifikasi. Eksperimen model klasifikasi tweet untuk tahapan filter menunjukkan bahwa model terbaik dengan akurasi 85.09% didapatkan dengan menggunakan algoritma pembelajaran C4.5, fitur trigram, dan tanpa praproses. Eksperimen model klasifikasi token untuk tahapan ekstraksi menunjukkan bahwa model terbaik dengan akurasi 81.49% didapatkan dengan menggunakan algoritma pembelajaran IBk (*Instance-based learning*) dan set 7 fitur terbaik dengan *gain ratio*.

Kata Kunci

Ekstraksi informasi, transaksi online, twitter, tweet, klasifikasi, bahasa Indonesia, algoritma pembelajaran

1. PENDAHULUAN

Informasi transaksi *online* di Indonesia didapatkan dari survei yang hanya melibatkan pembeli *online* dalam jumlah kecil (DailySocial, 2012). Pengumpulan informasi transaksi dari para penjual *online* tidak efektif karena jumlah penjual yang banyak dan informasi tersebut merupakan data rahasia bagi para penjual. Makalah ini mengeksplorasi analisis konten tweet dari twitter[1] untuk mengumpulkan secara otomatis informasi transaksi *online* di Indonesia. Pengguna twitter dari Indonesia telah mencapai 30 juta pada Juli 2012 (Semiocast, 2012). *Microblog* ini telah digunakan para penggunanya untuk menulis berbagai aktifitas termasuk aktifitas dalam melakukan transaksi *online*. Selain itu, transaksi *online* dilakukan pengguna twitter lebih banyak 24% dibandingkan rata-rata pengguna internet lainnya (Boorstin, 2012).

Berikut adalah contoh tweet berbahasa Indonesia yang berisi aktifitas pengiriman buku telah sampai ke pembeli dari penjual *online*. Dari tweet ini, didapatkan bahwa produk yang dibeli adalah #HOPE, sedangkan penjualnya adalah @bukabuku.

```
Setelah 3 minggu, akhirnya #HOPE nya @arsyilrahman
nyampe juga, thanks @bukabuku :))
http://t.co/dJOB3Ryt
```

Analisis konten tweet berbahasa Indonesia sudah dilakukan untuk berbagai kepentingan seperti analisis opini dan sentimen (Romelta, 2012; Sunni & Widyantoro, 2012; Aliandu, 2012), klasifikasi tweet kemacetan lalu lintas (Rodiyansyah, 2012), ekstraksi informasi kemacetan lalu lintas (Hasby & Khodra, 2013; Endarnoto dkk, 2011), ataupun peringkasan untuk menjelaskan *trending topic* pada twitter Indonesia (Winatmoko & Khodra, 2013).

Walaupun berbagai penelitian analisis konten tweets telah dilakukan, ekstraksi informasi transaksi *online* di tweet berbahasa Indonesia belum pernah dilakukan sebelumnya. Makalah ini bertujuan mengekstraksi informasi transaksi *online* di Indonesia dengan melakukan klasifikasi tweet dan ekstraksi informasi. Klasifikasi tweet merupakan aktifitas menentukan label atau kategori dari suatu tweet, misalnya opini positif, negatif, atau netral. Ekstraksi informasi merupakan proses yang mengumpulkan informasi target dari kumpulan teks yang tidak terstruktur ke dalam bentuk yang lebih terstruktur (Manning, 2012). Contoh informasi target dari tweet adalah nama jalan, dan kondisi jalan (Hasby & Khodra, 2013). Penelitian sebelumnya biasanya melakukan hanya klasifikasi tweet saja untuk mendapatkan tweet yang relevan (Rodiyansyah, 2012) atau ekstraksi informasi tweet langsung dengan asumsi kumpulan tweet yang diproses sudah relevan.

Aplikasi analisis yang dikembangkan disebut dengan SaFE-F yang melakukan pencarian tweet (***Search***) dengan kata kunci tertentu, memfilter konten tweet yang relevan dengan kegiatan transaksi *online* (**F**ilter), mengekstraksi informasi transaksi *online* (**E**kstraksi), dan menyimpan hasil ekstraksi dalam bentuk yang lebih terstruktur (***Filling***). Selain mengembangkan aplikasi SaFE-F, penelitian ini juga berkontribusi dalam mengkonstruksi korpus tweet yang telah dilabeli informasi transaksi *online*, dan mengembangkan model ekstraksi berbasis klasifikasi untuk mengekstraksi informasi transaksi *online* pada tweet berbahasa Indonesia (Hasby & Khodra, 2013).

Pada bagian selanjutnya, akan dibahas mengenai aplikasi SaFE-F dan setiap tahapan dalam memproses twitter sampai mendapatkan informasi hasil ekstraksi. Bagian 3 akan membahas korpus yang dikonstruksi untuk membangun model klasifikasi, sedangkan bagian 4 membahas eksperimen yang dilakukan. Pada bagian terakhir, dibahas kesimpulan dan penelitian selanjutnya yang akan dilakukan.

2. Ekstraksi Informasi

Untuk mendapatkan informasi terstruktur dari teks yang tidak terstruktur, hal pertama yang perlu didefinisikan adalah informasi target sebagai informasi terstruktur yang akan diekstrak. Informasi ini dapat berupa entitas ataupun relasi antar entitas. Secara umum, entitas dapat berupa orang, perusahaan, organisasi, atau lokasi. Oleh karena itu, kegiatan utama dalam ekstraksi informasi adalah pengenalan entitas (*named-entity recognition*) dan ekstraksi relasinya (Jiang, 2012).

[1] https://twitter.com/

Pengenalan entitas dapat dilakukan dengan memanfaatkan pola kemunculan entitas tersebut pada teks. Pola ini dapat didefinisikan secara manual oleh pakar ataupun didapatkan secara otomatis dengan pembelajaran mesin. Oleh karena itu, terdapat dua pendekatan dalam pengenalan entitas, yaitu pendekatan berbasis aturan dan pendekatan berbasis pembelajaran (Jiang, 2012).

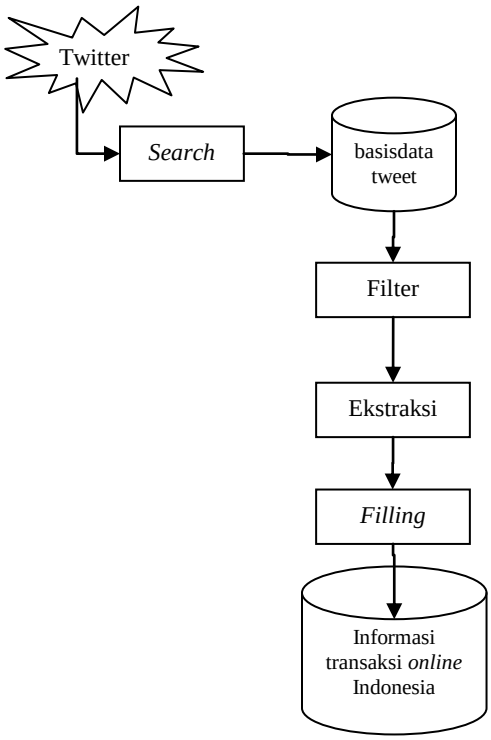
Setelah pengenalan entitas selesai dilakukan, kegiatan berikutnya adalah ekstraksi relasi antar entitas. Dengan mendefinisikan relasi semantik yang mungkin, entitas menjadi argumen dari relasi tersebut. Pendekatan yang paling umum dalam ekstraksi relasi adalah klasifikasi (Jiang, 2012). Berbagai pendekatan dibedakan oleh fitur yang digunakan (Hasby, 2013).

3. Aplikasi SAFE-F

Sesuai namanya yang telah dijelaskan pada bagian Pendahuluan, SAFE-F terdiri atas 4 tahapan utama yaitu *search*, filter, ekstraksi, dan *filling*. Keempat tahapan ini ditunjukkan oleh Gambar 1.

3.1 Tahap *Search*

Tahap *search* diawali dengan menentukan kata kunci pencarian. Kata kuncinya berupa nama akun twitter dari toko *online* Indonesia yang popular yaitu *kutukutubuku*, *tokopedia*, *bukabuku*, *tokobagus*, *bukalapak*, *juale*, *berniagaIndo*, *dmarketID*, *zaloraID*, *ngomik*, and *bhinnekacom*. Twitter API digunakan untuk tahapan pertama ini. Setiap 15 menit aplikasi SaFE-F melakukan pencarian dan menyimpannya ke dalam basis data tweet.



Gambar 1. Tahapan utama pada SaFE-F

Hasil pencarian berbasis kata kunci tersebut masih menghasilkan konten yang beragam. Terdapat tiga kategori tweet yang dihasilkan yaitu:

1. Kategori relevan, jika tweet tersebut berisi informasi aktifitas sebelum pembelian (berminat, memesan, dan membatalkan), aktifitas transaksi pembelian (terutama verifikasi pembayaran), dan aktifitas pengiriman barang. Pada contoh tweet berikut, contoh (a) menunjukkan aktifitas pemesanan, contoh (b) menunjukkan aktifitas pembelian, dan contoh (c) menunjukkan aktifitas pengiriman barang.

   a. `kk admin in bed with model nya masih stock ga? @bukabuku`

   b. `@bukabuku tolong donks verifikasi pembayaran saya terimakasih`

   c. `Just arrived!! Woahh, thanks @bukabuku :-D http://t.co/F5QbxMd6`

2. Kategori iklan, jika tweet tersebut berisi iklan produk yang dijual, contohnya:

   `Pasmina Sutra http://t.co/hMh3u1CU lewat @tokobagus`

3. Kategori tidak-relevan, jika tweet tersebut tidak mengandung informasi relevan ataupun iklan, contohnya:

   `ada pr gak yaaaaaaaa-,,,,-\nlg mls bukabuku wk`

3.2 Tahap Filter

Tahapan filter melakukan analisis konten tweet untuk mengklasifikasinya ke dalam satu dari tiga kategori tweet di atas. Tahapan ini menggunakan model klasifikasi yang merupakan hasil *supervised learning*. Algoritma pembelajaran mesin ini menghasilkan model klasifikasi sebagai fungsi estimasi yang mampu memetakan konten tweet ke salah satu kategori tersebut. Fungsi estimasi ini merupakan pola pemetaan yang ada pada data pembelajaran, yaitu korpus pasangan tweet dan kategorinya. Setiap tweet yang didapatkan dari tahap search menjadi input dari tahap filter untuk ditentukan kategorinya.

Setiap tweet direpresentasikan sebagai vektor fitur. Penelitian ini memanfaatkan kata dan n-gram sebagai fitur leksikal. Tantangan dari representasi leksikal adalah mencari kumpulan term (kata atau n-gram) yang paling representatif dari tweet yang diproses. Semakin banyak term yang digunakan (sering disebut dimensi vektor), semakin lengkap informasi tweet yang direpresentasikan tetapi hal ini akan membutuhkan waktu pemrosesan yang semakin besar. Besarnya dimensi dipengaruhi oleh praproses yang dilakukan seperti pembuangan kata yang tidak bermakna (stopword), penggunaan huruf kecil atau kapital (case folding), atau seleksi fitur dengan membuang term yang memiliki frekuensi kemunculan yang rendah. Selain besarnya dimensi vektor, kualitas model klasifikasi juga dipengaruhi pembobotan term yang digunakan.

Hanya tweet dengan kategori relevan yang akan masuk ke tahapan selanjutnya. Tweet dengan kategori iklan belum akan diproses dalam penelitian ini.

3.3 Tahap Ekstraksi

Untuk tahapan ekstraksi, terdapat 10 jenis informasi yang akan diekstraksi dari tweet yaitu: produk yang dibeli (PP: purchased

product), produk yang diminati (PI: product of interest), tempat belanja (SP: shopping place), jumlah produk yang dibeli (NP: number of purchase), cara pembayaran (PM: payment method), kepuasan pelanggan (CS: customer satisfaction), ketidakpuasan pelanggan (CI: customer inconvenience), haparan pelanggan (CE: customer expectation), lokasi pengguna (UL: user location), dan harga produk (PR: product price). Seperti penelitian Hasby & Khodra (2013), tahapan ekstraksi mengaplikasikan pendekatan ekstraksi informasi berbasis model klasifikasi.

Klasifikasi yang dilakukan pada tahap ekstraksi berbeda dengan klasifikasi pada tahap filter. Jika tahapan filter mengklasifikasi setiap tweet ke dalam kategori relevansinya dengan transaksi *online*, tahapan ekstraksi tidak mengklasifikasi tweet tetapi mengklasifikasi setiap token atau kata pada tweet. Model klasifikasi untuk tahap ekstraksi ini juga dibangun secara otomatis dengan algoritma pembelajaran mesin.

Berdasarkan penjelasan sebelumnya, terdapat 10 jenis informasi yang akan diekstraksi dari setiap tweet. Dengan menggunakan notasi BIO (Begin In Other), setiap jenis informasi terdiri atas dua kategori yaitu kategori `begin-<jenis informasi>` untuk token pertama yang mengandung informasi tersebut dan `in-<jenis informasi>` untuk token kedua dan berikutnya yang mengandung informasi tersebut. Kategori tambahan `other` didefinisikan untuk token lain yang tidak berlabel. Total kategori untuk 10 jenis informasi yang telah didefinisikan tersebut adalah 10*2+1=21 kategori. Model klasifikasi akan menganalisis setiap token pada tweet dan menentukan kategori token tersebut. Contoh berikut merupakan contoh tweet pada korpus yang setiap tokennya telah ditentukan kategorinya.

Setelah/B-CS **3**/I-CS **minggu,**/I-CS **akhirnya**/I-CS **#HOPE**/B-PP nya/O @arsyilrahman/O nyampe/O juga,/O thanks/O **@bukabuku**/B-SP :))/O http://t.co/dJOB3Ryt/O

Terdapat 3 jenis informasi pada tweet tersebut yaitu CS (customer satisfaction), PP (purchased product), dan SP (shopping place). Token pertama setiap informasi mendapat awalan B (begin) seperti kategori B-CS (begin-CS), B-PP (begin-PP), dan B-SP (begin-SP). Jika informasi tersebut mengandung lebih dari satu token, kategori yang digunakan diawali dengan I (in) seperti I-CC (in-CS).

Berbeda dengan vektor fitur untuk model klasifikasi tweet pada tahap filter, vektor fitur untuk token didefinisikan berdasarkan atribut leksikal token tersebut dan tetangganya.

Tabel 1. Set fitur untuk setiap token pada tweet

Fitur	Deskripsi
currentWord	Leksikal token yang diproses
currentTag	POS tag dari token yang diproses
Bef1Word	Leksikal token sebelum token yang diproses
Bef1Tag	POS tag dari token sebelum token yang diproses
Bef1Class	Kategori dari token sebelum token yang diproses
Bef2Word	Leksikal token dengan gap 2 dari token yang diproses
Bef2Tag	POS tag dari token dengan gap 2 dari token yang diproses
Bef2Class	Kategori dari token dengan gap 2 dari token yang diproses
IsLink	Apakah token tersebut *link* ?
isNumber	Apakah token tersebut angka ?
isMention	Apakah token tersebut *mention*, yaitu token yang diawali dengan @ sebagai karakter pertamanya ?
isTag	Apakah token tersebut *hastag*, yaitu token yang diawali dengan # sebagai karakter pertamanya ?
isPrice	Apakah token tersebut harga?
isDate	Apakah token tersebut tanggal ?

3.4 Tahap Filling

Untuk setiap jenis informasi X, tahapan *filling* menggabungkan semua token dengan kategori B-X dan I-X yang berurutan sampai mendapatkan token dengan kategori berbeda. Tahapan ini akan menyimpan hasil ekstraksi dalam bentuk yang lebih terstruktur. Pada contoh tweet berikut, hasil ekstraksi menentukan kategori setiap token, lalu hasil filling berupa kumpulan informasi dengan jenis informasi yang sesuai hasil ekstraksi.

Tweet: `Ga sabar nunggu 2 brg yg w pesen .. Buku Casual Vacancy dr Bukabuku .. N jaket varsity Bigbang dr Sevenstar ..`

Hasil ekstraksi: **Ga**/B-CI **sabar**/I-CI **nunggu**/I-CI **2**/I-CI **brg**/I-CI **yg**/I-CI **w**/I-CI **pesen**/I-CI .. **Buku**/B-PP **Casual**/B-PP **Vacancy**/I-PP **dr**/O **Bukabuku**/B-SP .. **N**/O **jaket**/B-PP **varsity**/B-PP **Bigbang**/I-PP **dr**/O **Sevenstar**/B-PP ..

Hasil *filling*:

- Purchased product: {`buku, casual vacancy, jaket, varsity bigbang`}
- Shopping place: {`bukabuku, sevenstar`}
- Customer inconvenience: {`ga sabar nunggu 2 brg yg w pesen`}

Pada contoh ini, masih terdapat kesalahan ekstraksi token 'Casual' dan 'Varsity' yang diklasifikasikan sebagai kategori B-PP. Kedua token tersebut seharusnya diklasifikasikan sebagai kategori I-PP, sehingga hasil fillingnya akan menjadi '`buku casual vacancy`' dan '`jaket varsity bigbang`'. Diskusi mengenai kinerja model klasifikasi ekstraksi akan dibahas lebih lanjut pada bagian Eksperimen.

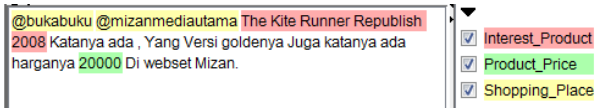
4. Korpus Tweet Transaksi *Online*

Berdasarkan penjelasan pada bagian Aplikasi SaFE-F, terdapat dua korpus tweet yang dibutuhkan, yaitu korpus filter untuk pembangunan model klasifikasi pada tahapan filter, dan korpus ekstraksi untuk pembangunan model klasifikasi pada tahapan ekstraksi. Kedua korpus ini dianotasi oleh manusia (*annotator*) dengan menganalisis secara manual relevansi konten tweet dan informasi yang terkandung di dalam tweet yang relevan. Karena informasi yang akan dilabeli bersifat umum, annotator merupakan pengguna twitter yang biasa melakukan transaksi *online*.

Korpus filter berisi 5000 tweet berlabel yang terdiri atas 1875 tweet berkategori tidak relevan, 369 tweet berkategori relevan, dan 2756 tweet berkategori iklan.

Korpus ekstraksi berisi 3455 token berlabel yang diambil dari 266 tweet relevan. Anotasi dilakukan dengan menggunakan GATE (Cunningham dkk, 2011) sebagai perangkat anotasi. Setiap tweet direpresentasikan sebagai satu dokumen, dan anotasi dilakukan per tweet. Korpus ini berupa kumpulan 266 xml yang dibangkitkan oleh GATE. **Error! Reference source not found.** shows labeled tweet in GATE GUI and xml format.

```
@bukabuku @mizanmediautama The Kite Runner Republish 2008 Katanya ada, Yang versi goldenya Juga katanya ada harganya 20000 Di webset Mizan.
```



Gambar 2. Hasil anotasi informasi yang akan diekstraksi dari suatu tweet pada antarmuka GATE

```
<?xml version='1.0' encoding='UTF-8'?>
<GateDocument>
<!-- The document's features-->
<GateDocumentFeatures>
<Feature>
  <Name
className="java.lang.String">gate.SourceURL</Name>
  <Value className="java.lang.String">created from
String</Value>
</Feature>
</GateDocumentFeatures>
<!-- The document content area with serialized nodes
-->
<TextWithNodes><Node id="0" />@bukabuku<Node id="9"
/> <Node id="10" />@mizanmediautama<Node id="26" />
<Node id="27" />The Kite Runner Republish 2008<Node
id="57" /> Katanya ada , Yang Versi goldenya Juga
katanya ada harganya <Node id="118" />20000<Node
id="123" /> Di webset Mizan.</TextWithNodes>
<!-- The default annotation set -->
<AnnotationSet><Annotation Id="0"
Type="Shopping_Place" StartNode="0"
EndNode="9"></Annotation>
<Annotation Id="1" Type="Shopping_Place"
StartNode="10" EndNode="26"></Annotation>
<Annotation Id="2" Type="Interest_Product"
StartNode="27" EndNode="57"></Annotation>
<Annotation Id="3" Type="Product_Price"
StartNode="118"
EndNode="123"></Annotation></AnnotationSet>
</GateDocument>
```

Gambar 3. Hasil anotasi tweet dalam format xml

5. Eksperimen

Eksperimen bertujuan untuk mendapatkan model klasifikasi terbaik untuk tahapan filter dan tahapan ekstraksi. Seperti yang dijelaskan sebelumnya, model klasifikasi untuk tahapan filter mengklasifikasi setiap tweet ke dalam kategori relevan, sedangkan model klasifikasi untuk tahapan ekstraksi mengklasifikasi setiap token pada tweet ke dalam kategori BIO-<jenis informasi>.

Eksperimen dilakukan dengan menggunakan Weka 3.7.9 (Hall dkk, 2009). Algoritma pembelajaran yang digunakan telah disediakan Weka yaitu Naive Bayes (NB), *Instance-based learning* (IBk), dan C4.5. Model NB berupa model probabilitas setiap atribut terhadap suatu kelas, dan klasifikasi suatu data menggunakan model NB dilakukan dengan mencari kelas yang memiliki probabilitas maksimum berdasarkan atribut dari data tersebut (Mitchell, 1997). Ibk tidak menghasilkan model pembelajaran karena bersifat *lazy learning*, tetapi hanya menyimpan semua data pembelajaran yang ada. Klasifikasi suatu data pada Ibk dilakukan dengan mencari kelas mayoritas dari k-data tetangga terdekat dengan data yang diklasifikasikan (Mitchell, 1997). Model C4.5 berupa pohon keputusan, dan klasifikasi suatu data dilakukan dengan menginferensi pohon sampai mencapai daun yang merepresentasikan kategori yang dicari (Mitchell, 1997).

5.1 Klasifikasi Tweet untuk Filter

Eksperimen ini menggunakan korpus filter dengan skema 66% *percentage split*, artinya 66% data dipilih secara random menjadi data pembelajaran, dan sisanya menjadi data pengujian. Tabel 2 menunjukkan jumlah tweet untuk setiap kategori pada data pembelajaran dan pengujian.

Tabel 2. Jumlah tweet pada korpus filter untuk eksperimen

Kategori relevansi	#tweet data pembelajaran	# tweet data pengujian	Total tweet
Tidak relevan	1252	623	1875
Relevan	230	139	369
Iklan	1868	888	2756
Total tweets	**3350**	**1650**	**5000**

Seperti yang telah disebutkan sebelumnya, data pembelajaran yang berisi kumpulan tweet dipraproses menjadi vektor fitur agar didapatkan pola untuk memetakannya ke kategori relevansi. Eksperimen ini memanfaatkan filter StringToWordVector yang telah disediakan oleh Weka. Beragam pilihan praproses dan pembobotan merupakan parameter dari filter ini.

Pada Tabel 3, ditunjukkan hasil eksperimen untuk setiap setting, yaitu jenis fitur (kata/unigram, n-gram), praproses (pembuangan stopword, case folding, frekuensi minimum), pembobotan (biner, tf.idf), dan algoritma pembelajaran yang digunakan (NB, IBk, dan C4.5). Representasi tweet dengan 3-gram lebih baik daripada unigram (satu kata). Semua praproses tidak berhasil memperbaiki kinerja model tanpa praproses, bahkan kinerjanya cenderung menurun dengan adanya praproses. Model klasifikasi terbaik didapatkan dengan representasi fitur trigram, pembobotan biner, dan algoritma C4.5.

Tabel 3. Akurasi berbagai model klasifikasi tweet

Fitur	Praproses	Pembobotan	Akurasi Model Klasifikasi (%)		
			NB	**C4.5**	**IBk**
kata	-	biner	73.21%	84.18%	77.64%
kata	stopword	biner	68.18%	73.03%	69.70%
kata	Case-folding	biner	72.91%	83.70%	75.52%
kata	-	TF.IDF	73.21%	84.18%	77.64%
kata	Minimum	biner	73.27%	84.18%	77.70%

	frequency =3				
2-gram	-	biner	70.18%	85.03%	79.70%
3-gram	-	biner	71.52%	***85.09%***	80.18%
2-gram	Case-folding, minimum frequency =3	biner	70.97%	84.91%	77.94%
3-gram	Case-folding, minimum frequency =3	biner	70.97%	84.79%	77.76%

Walaupun kinerja model klasifikasi cukup baik, yaitu dalam rentang 68.18%-85.09%, belum ada penanganan jumlah data per kategori yang tidak seimbang (*imbalanced dataset*). Kategori relevan merupakan target utama dalam eksperimen ini karena hanya tweet berkategori relevan yang akan diproses pada tahap ekstraksi. Gambar 4 menunjukkan hanya 42 tweet berhasil diklasifikasikan dengan benar berkategori relevan (kategori b=1) dari 139 tweet yang seharusnya berkategori relevan.

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 496  20 107 |   a = 0
  50  42  47 |   b = 1
  20   2 866 |   c = 2
```

Gambar 4. *Confusion matrix* dari model terbaik (C4.5 3-gram)

5.2 Klasifikasi Token untuk Ekstraksi

Sama seperti eksperimen klasifikasi tweet, eksperimen ini menggunakan skema 66% *percentage split*, tetapi dengan menggunakan korpus ekstraksi. Tabel 4 menunjukkan jumlah token untuk setiap kategori pada data pembelajaran dan pengujian.

Tabel 4. Jumlah tweet pada korpus ekstraksi pada eksperimen

Kategori	#token data pembelajaran	#token data pengujian	Total token
B-IP	81	46	127
I-IP	72	36	108
B-PP	34	23	57
I-PP	24	12	36
B-SP	224	116	340
I-SP	7	1	8
B-NP	2	2	4
I-NP	0	0	0
B-PM	6	2	8
I-PM	0	1	1
B-CS	8	15	23
I-CS	36	41	77
B-CI	32	16	48
I-CI	178	91	269
B-CE	2	2	4
I-CE	10	10	20
B-UL	10	2	12
I-UL	5	0	5
B-PR	13	20	33
I-PR	9	4	13
Other (O)	1562	700	2262
Total token	**2315**	**1140**	**3455**

Eksperimen klasifikasi token juga dilakukan dengan Weka dengan algoritma pembelajaran NB, IBk, dan C4.5. Untuk mendapatkan fitur sesuai Tabel 1, dikembangkan model ekstraksi fitur. Contoh tweet pada Gambar 2 menghasilkan 20 token, sedangkan vektor fitur dari setiap token ditunjukkan oleh Gambar 5 dengan format:

```
<currentWord>,<currentTag>,<Bef1Word>, Bef1Tag>,
<Bef1Class>,<Bef2Word>,<Bef2Tag>,<Bef2Class>,
<isLink>,<isNumber>,<isMention>,<isTag>,<isPrice
>,<isDate>,<Class>
```

```
a. @bukabuku,mention,<start>,null,N/A,<start>,null,N/
   A,false,false,true,false,false,false,shopping_plac
   e_begin
b. @mizanmediautama,mention,@bukabuku,shopping_place_
   begin,mention,<start>,null,N/A,false,false,true,fa
   lse,false,false,shopping_place_begin
c. the,N/A,@mizanmediautama,shopping_place_begin,ment
   ion,@bukabuku,shopping_place_begin,mention,false,f
   alse,false,false,false,false,interest_product_begi
   n
d. kite,N/A,the,interest_product_begin,N/A,@mizanmedi
   autama,shopping_place_begin,mention,false,false,fa
   lse,false,false,false,interest_product
e. runner,N/A,kite,interest_product,N/A,the,interest_
   product_begin,N/A,false,false,false,false,false,fa
   lse,interest_product
f. republish,N/A,runner,interest_product,N/A,kite,int
   erest_product,N/A,false,false,false,false,false,fa
   lse,interest_product
g. 2008,number,republish,interest_product,N/A,runner,
   interest_product,N/A,false,true,false,false,false,
   false,interest_product
h. katanya,N/A,2008,interest_product,number,republish
   ,interest_product,N/A,false,false,false,false,fals
   e,false,other
i. ada,verb,katanya,other,N/A,2008,interest_product,n
   umber,false,false,false,false,false,false,other
j. yang,relpronoun,ada,other,verb,katanya,other,N/A,f
   alse,false,false,false,false,false,false,other
k. versi,noun,yang,other,relpronoun,ada,other,verb,fa
   lse,false,false,false,false,false,false,other
```

```
l. goldenya,N/A,versi,other,noun,yang,other,relpronou
   n,false,false,false,false,false,false,other
m. juga,adverb,goldenya,other,N/A,versi,other,noun,fa
   lse,false,false,false,false,false,other
n. katanya,N/A,juga,other,adverb,goldenya,other,N/A,f
   alse,false,false,false,false,false,other
o. ada,verb,katanya,other,N/A,juga,other,adverb,false
   ,false,false,false,false,false,other
p. harganya,N/A,ada,other,verb,katanya,other,N/A,fals
   e,false,false,false,false,false,other
q. 20000,number,harganya,other,N/A,ada,other,verb,fal
   se,true,false,false,false,false,product_price_begi
   n
r. di,preposition,20000,product_price_begin,number,ha
   rganya,other,N/A,false,false,false,false,false,fal
   se,other
s. webset,N/A,di,other,preposition,20000,product_pric
   e_begin,number,false,false,false,false,false,false
   ,other
t. mizan,N/A,webset,other,N/A,di,other,preposition,fa
   lse,false,false,false,false,false,other
```

Gambar 5. Hasil ekstraksi fitur setiap token tweet Gambar 2

Pada Tabel 5, ditunjukkan hasil eksperimen untuk setiap set fitur dan algoritma pembelajaran. Akurasi terbaik sebesar 81.49% didapatkan dengan algoritma IBk dan set 7 fitur terbaik berdasarkan *gain ratio* yang disediakan Weka. Fitur terbaik yang didapatkan adalah currentWord+ Bef1Class+Bef2Class+ IsLink+ isNumber+ isMention+ isPrice.

Tabel 5. Akurasi berbagai model klasifikasi token

Set Fitur	Akurasi model klasifikasi (%)		
	NB	**C4.5**	**IBk**
All features in **Error! Reference source not found.**	77.81%	77.11%	75.70%
currentWord+ currentTag+ IsLink+ isNumber+isMention+isTag +isPrice+isDate	62.98%	66.84%	66.14%
currentWord+Bef1Word+Bef2Word	77.81%	77.11%	75.70%
currentWord+currentTag+Bef1Word+ Bef1Tag+Bef2Word+ Bef2Tag	64.56%	66.84%	62.11%
currentWord+ Bef1Word+ Bef1Class+Bef2Word+ Bef2Class	78.16%	77.11%	79.91%
currentWord+ Bef1Class+Bef2Class+ IsLink+ isNumber+ isMention+ isPrice	78.68%	77.11%	***81.49%***

6. KESIMPULAN

Pada makalah ini, telah dijelaskan aplikasi SaFE-F yang mengekstrak informasi transaksi *online* di Indonesia dari konten tweet. Terdapat empat tahapan pada SaFE-F yaitu: *Search*, Filter, Ekstraksi, dan *Filling*. Karena sistem ini menggunakan pendekatan klasifikasi untuk tahap filter dan ekstraksi, beberapa eksperimen dilakukan untuk mendapatkan model klasifikasi terbaik. Untuk tahap filter, didapatkan model terbaik dengan akurasi 85.09% dengan representasi fitur trigram, pembobotan biner, dan algoritma C4.5. Untuk tahap ekstraksi, didapatkan model terbaik dengan akurasi 81.49% dengan menggunakan algoritma IBk dan set 7 fitur terbaik berdasarkan *gain ratio* yaitu currentWord+ Bef1Class+Bef2Class+ IsLink+ isNumber+ isMention+ isPrice.

Untuk penelitian selanjutnya, perlu dilakukan penanganan imbalanced dataset terutama untuk model klasifikasi filter. Eksplorasi lebih lanjut untuk set fitur lain ataupun algoritma pembelajaran lainnya juga dapat dilakukan. Selain itu, eksperimen dalam penelitian ini dapat diterapkan juga untuk data media sosial lainnya seperti facebook ataupun kaskus.

7. REFERENSI

Aliandu, P. Analisis Sentimen Tweet Berbahasa Indonesia Di twitter. Tesis S2 Ilmu Komputer UGM. 2012

Boorstin, Julia. Twitter Sees a Surge in Retailer Activity. CNBC. [*Online*] 20 November 2012. [accessed: 31 March 2013.] http://www.cnbc.com/id/49901550.

Cunningham, H. et al. Text Processing with GATE (Version 6). s.l. : University of Sheffield Department of Computer Science, 2011.

DailySocial. eCommerce in Indonesia. www.DailySocial.net. [*Online*] August 2012. [accessed: 30 March 2013.] www.DailySocial.net.

Endarnoto, S., Pradipta, S., A.S, N., & Purnama, J. Traffic Condition Information Extraction & Visualizations from Social Media Twitter for Android Mobile Application. ICEEI (pp. 1-4). IEEE. 2011

Hasby, M., Khodra, M.L. Optimal Path Finding based on Traffic Information Extraction from Twitter. Prosiding International Conference on ICT for Smart Society 2013. Jakarta. 2013.

Hasby, M. Ekstraksi Informasi Dari Twitter Untuk Pencarian Jalur Optimal. Tugas Akhir S1 Teknik Informatika ITB. 2013.

Indra, Nikolaus. *Sistem Pemberi Tahu Kemacetan Lalu Lintas di Kota Bandung Berbasis Media Sosial*. Laporan tugas akhir, Institut Teknologi Bandung, Bandung: Program Studi Teknik Informatika.2012.

Jiang, J. Information Extraction from Text, in Mining Text Data. Springer. 2012.

Manning, C. Information Extraction and Named Entity Recognition.California: Stanford University. 2012.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H. The WEKA Data Mining Software: An Update. 2009. SIGKDD Explorations. Vol. 11 (Issue 1).

Mitchell, T. Machine Learning. New York : McGraw-Hill, 1997

Romelta, E. Opinion Mining Di Twitter Untuk Customer Feedback Smartphone Dengan Pembelajaran Mesin. Jurnal Sarjana ITB bidang Teknik Elektro dan Informatika. Vol. 1, No.2, 2012.

Sunni, I., Widyantoro, D.H. Analisis Sentimen dan Ekstraksi Topik Penentu Sentimen pada Opini Terhadap Tokoh Publik. Jurnal Sarjana ITB bidang Teknik Elektro dan Informatika. Vol. 1, No.2, 2012.

Twitter reaches half a billion accounts, More than 140 millions in the U.S.: Geolocation analysis of Twitter accounts and tweets. Semiocast. [*Online*].

Rodiyansyah, S.F. Klasifikasi Posting twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive. Tesis S2 Ilmu Komputer UGM. 2012

Winatmoko, Y.A., Khodra, M.L. Automatic Summarization of Tweets in Providing Indonesian Trending Topic Explanation. Prosiding International Conference on Electrical Engineering and Informatics 2013. Malaysia. 2013