Identifikasi Kalimat Penting untuk Ekstraksi Informasi 5W1H pada Teks Berita dengan *Sequence Labeling*

Najib Darmawan – 14113001[1]
Program Studi Teknik Informatika
Jurusan Sains
Institut Teknologi Sumatera, Jl. Terusan Ryacudu, Desa Way Hui, Kecamatan Jati Agung, Lampung Selatan 35365
[1]najib.darmawan@gmail.com

Ekstraksi Informasi 5W1H mengidentifikasi siapa melakukan apa, kapan, dimana, mengapa, dan bagaimana. Melabeli teks berita Bahasa Indonesia dengan menggunakan notasi BIO (*Begin In Others)* memberikan nilai *Fmeasure* tertinggi sebesar 0,666. Akan diteliti apakah dengan melakukan pemilihan kalimat penting dari suatu artikel dapat meningkatkan kinerja *sequence labeling* dengan notasi BIO dalam mengekstraksi informasi 5W1H, serta untuk menemukan kombinasi fitur terbaik pada tahapan identifikasi kalimat penting. Fitur terbaik untuk mengklasifikasikan kalimat penting yaitu *term weight* dan *sentence location*. Hasil ekstraksi informasi terbaik didapatkan dengan menggunakan *Naïve Bayes* dan menyeleksi fitur *is time pattern,* memberikan nilai *Fmeasure* sebesar 0.860. Identifikasi kalimat penting berhasil meningkatkan nilai *Fmeasure* ekstraksi informasi 5W1H yang dilakukan dengan *sequence labeling* sebesar 28,378%.

Kata Kunci — Ekstraksi Informasi 5W1H, Identifikasi Kalimat Penting, *Sequence Labeling,* Notasi BIO, Naïve Bayes.

I. Latar Belakang

Sebanyak 132,7 juta dari 256,2 juta penduduk Indonesia merupakan pengguna internet (Anonim, 2016). Pengguna internet memiliki alasan yang berbeda ketika mengakses internet, 96,4% diantaranya mengakses internet untuk mencari berita (Anonim, 2016). Hal ini menunjukkan terjadinya perubahan cara mengakses berita dari media cetak ke media digital. Pembaca berita biasanya mencari berita untuk mengetahui kondisi terkini dari lingkungannya, dalam hal ini situs berita *online* dapat menjadi solusi untuk memenuhi kebutuhan tersebut.

Adanya agregator berita dapat membantu pembaca mengumpulkan berita dari berbagai sumber dan menyajikannya dalam satu tempat. Agregator berita akan menyajikan daftar topik berita dan kumpulan berita untuk setiap topik sehingga pembaca dapat dengan mudah menemukan berita terkait suatu topik (Khodra, 2014). Namun, agregator berita saja tidak cukup untuk menghemat waktu dan meringankan upaya pembaca dalam mendapatkan informasi yang lengkap. Pembaca masih harus membaca kumpulan berita dalam suatu topik. Hal ini dapat menyebabkan pembaca mendapatkan informasi yang sama dalam beberapa berita berberda dengan satu topik.

Khodra (2014) telah mengembangkan Sistem Cerdas Aggregator Berita Bahasa Indonesia dengan 4 komponen utama yaitu 5W1H *extractor*, *classifier, clusterer,* dan *summarizer*. Namun, pada sistem tersebut ekstraksi informasi hanya dilakukan pada entitas yang ada pada teks berita sehingga dapat menghasilkan informasi 5W1H yang tidak lengkap. Pada penelitian lainnya, Khodra melakukan ekstraksi informasi 5W1H menggunakan *sequence labeling* yang diimplementasikan dengan notasi *Begin In Other* (BIO) dan menghasilkan nilai *Fmeasure* tertinggi sebesar 0,666 (Khodra, 2015). Wang (2012) melakukan penelitian untuk ekstraksi informasi 5W1H dari teks berita dengan pendekatan yang berbeda. Pendekatan milik Wang dibagi menjadi beberapa tahap, salah satu tahapannya yaitu identifikasi kalimat penting. Identifikasi kalimat penting menghasilkan akurasi tertinggi di atas 95% (Wang, 2012). Namun pendekatan Wang masih memiliki akurasi yang kurang baik saat menentukan elemen 5W1H.

Nilai *Fmeasure* sebesar 0,666 yang dihasilkan oleh Khodra (2015) menunjukkan adanya potensi metode *sequence labeling* dengan notasi BIO dapat menangani karakteristik teks berita Bahasa Indonesia. Hanya saja pada penelitian tersebut tidak dilakukan *filtering* kalimat sehingga kalimat-kalimat yang tidak mengandung informasi 5W1H masih diproses. Dalam hal ini tahapan identifikasi kalimat penting yang melakukan pemilihan kalimat penting suatu teks berita dapat menjadi solusi. Oleh karena itu, kontribusi dari penelitian ini adalah membangung korpus kalimat penting dan mengkombinasikan kedua metode tersebut..

II. Tinjauan Pustaka

Terdapat dua pendekatan umum yang digunakan dalam melakukan ekstraksi informasi yaitu pendekatan berbasis aturan dan pendekatan pembelajaran statistik. Kebanyakan penelitian yang dilakukan sekarang menggunakan pendekatan pembelajaran statistik.

Sebagian besar dari pendekatan tersebut memperlakukan ekstraksi informasi sebagai *sequence labeling problem*.

Pada *sequence labeling,* suatu dokumen dianggap sebagai urutan token. Urutan label setiap token menunjukkan *property* dari token tersebut. Berbeda dengan pendekatan berbasis aturan*, sequence labeling* memungkinkan untuk mendeskripsikan ketergantungan antar target informasi. Ketergantungan ini dapat dimanfaatkan untuk meningkatkan akurasi saat ekstraksi (Tang dkk., 2007). Dengan menganggap ekstraksi informasi sebagai *sequence labeling,* kumpulan label perlu didefinisikan terlebih dahulu berdasarkan tujuan ekstraksi. Sebagai contoh, pada ekstraksi informasi 5W1H dari penelitian yang dilakukan Khodra (2015) label dengan notasi BIO didefinisikan. Notasi tersebut terdiri dari 13 label yaitu Beg_who, In_who, Beg_what, In_what, Beg_when, In_when, Beg_where, In_where, Beg_why, In_why, Beg_how, In_how, dan Other. Suatu dokumen dianggap sebagai unit observasi yang dapat berupa kata, kalimat, ataupun unit lain. Kemudian *sequence labeling* akan mencari label *y* yang memenuhi kondisi probabilitas *p(y|x)* terhadap unit observasi.

Berdasarkan linguistik dan karakteristik struktural dari gaya bahasa pada teks berita, Wang (2012) mengajukan metode yang menggunakan fitur *surface* dan semantik dalam memilih kalimat yang paling informatif. Fitur-fitur tersebut adalah *term weight*, *sentence location*, *sentence length*, *number of name entities*, *title words overlaps rate*, dan *informativeness of news headline.* Kombinasi linier dari bobot fitur-fitur tersebut digunakan untuk menentukan bobot dari setiap kalimat. Bobot yang digunakan untuk fitur *title words overlap rate* merupakan nilai dari fitur *informativeness of news headline*. Menurut Wang hal ini akan memperkuat kemiripan antara kalimat dan judul berita untuk judul yang informatif. Sedangkan untuk judul yang tidak informatif hal ini akan menghindari efek negatif dari fitur itu sendiri. Setiap kalimat diurutkan berdasarkan nilai bobot dan dipilih sejumlah N kalimat dengan bobot tertinggi. Pada penelitian Wang hasil terbaik didapatkan ketika nilai N=3. Menurut Wang tahapan ini menghasilkan kumpulan kalimat yang lebih singkat dari peringkasan dan memiliki informasi yang lebih banyak daripada judul berita. Namun, tahapan identifikasi kalimat penting memiliki kelemahan yaitu akurasi akan menurun pada teks berita yang panjang.

III. KORPUS

Korpus yang akan digunakan pada tugas akhir ini merupakan korpus 5W1H yang digunakan Khodra (2015) dalam penelitiannya. Korpus tersebut dapat digunakan untuk mengevaluasi hasil ekstraksi informasi 5W1H sistem. Contoh teks berita yang ada pada korpus 5W1H tersebut dapat dilihat pada Tabel 1. Pada korpus tersebut teks berita diambil dari 4 sumber yaitu Tempo, Antara News, Republik, dan Metrotvnews. Distribusi banyak kalimat suatu berita pada korpus tersebut dapat dilihat pada Tabel 2. Korpus yang sudah ada ini tidak dapat sepenuhnya langsung digunakan. Perlu ditambahkan anotasi baru sehingga tahapan identifikasi kalimat penting dapat dievaluasi.

Rakyat Venezuela Mulai Pilih Pengganti Chavez	Information 5W1H
REPUBLIKA.CO.ID, KARAKAS -- Transisi kekuasaan akan dimulai di Venezuela. Tidak kurang dari 19 juta pemilih, akan menentukan masa depan politik negara tersebut saat Ahad (14/4) waktu setempat. Beberapa hasil polling masih menempatkan, pejabat presiden sementara, Nicolas Maduro unggul dari pesaingnya, Hanrique Capriles. Tapi, prediksi tersebut dikhawatirkan berbalik. Reuters mengatakan Maduro, menurut polling, unggul dua dijit meninggalkan bekas Gubernur Negara Bagian Miranda tersebut. Namun, kesenjangan popularitas itu kian menyempit menjelang pemilihan. Dikatakan jarak popularitas hanya terpaut tujuh persen. Caprilles mengatakan, aksi manipulasi politik Maduro telah memecah belah masyarakat Venezuela. Dirinya yakin akan mengambil peran presiden, dan akan menjungkalkan rezim sosialisme di negara itu. Dalam kampanye, dia mengatakan, pesaingnya itu culas memanfaatkan mendiang Presiden Hugo Chavez untuk kepentingan pribadinya. Padahal, kata dia, mendiang Chavez adalah milik semua masyarakat Venezuela. Tempat pemungutan suara akan di buka Ahad (14/4) pukul 06:30 waktu setempat, atau Senin (15/4) waktu Indonesia. TPS akan dibuka selama 10 jam, dengan voting secara elektronik. Pemenangnya akan dilantik 19 April mendatang, dan akan menentukan pemerintahan sampai 2019 mendatang.	What: menentukan masa depan politik negara tersebut Who: 19 juta pemilih Where: Venezuela When: Ahad (14/4) Why: Pemenangnya akan dilantik 19 April mendatang, dan akan menentukan pemerintahan sampai 2019 mendatang . How: TPS akan dibuka selama 10 jam, dengan voting secara elektronik.

Tabel 1. Contoh Teks Berita Pada Korpus

Wang (2012) menggunakan dua definisi berikut untuk menentukan kalimat penting:

Elemen 5W : merupakan waktu, lokasi, subjek, predikat, dan objek yang mendeskripsikan *when, where, who, what,* dan *whom* dari suatu *event*. Elemen 5W dilambangkan dengan <T,L,S,P,O> yang mana S, P, dan O merupakan elemen utama sedangkan T dan L merupakan elemen pendukung.

Kalimat penting: suatu kalimat merupakan kalimat penting jika dan hanya jika elemen S, P, dan O dan setidaknya salah satu elemen T atau L dapat diekstrak dari kalimat tersebut.

Berdasarkan definisi di atas tidak ada kalimat pada teks berita di Tabel 1 yang memenuhi definisi kalimat penting Wang. Oleh karena itu definisi tersebut tidak akan digunakan. Pada penelitian ini didefinisikan suatu kalimat merupakan kalimat penting jika kalimat tersebut menggandung elemen 5W1H. Selain itu, jika informasi 5W1H muncul lebih dari 1 kali pada suatu kalimat dalam teks berita, maka akan dipilih kalimat dengan jumlah elemen 5W1H yang lebih banyak. Namun jika jumlah elemen 5W1H yang muncul sama, maka akan dipilih kalimat yang terletak lebih awal.

Jumlah kalimat penting dalam suatu teks berita dapat berbeda. Berdasarkan hal tersebut identifikasi kalimat penting pada penelitian ini tidak dibatasi dengan suatu *threshold* tertentu seperti pada penelitian Wang (2012). Dengan demikian model yang dibuat nantinya dapat menghindari kesalahan dalam memilih banyaknya kalimat penting. Setelah dilakukan anotasi terhadap 90 teks berita, didapatkan jumlah kalimat penting sebanyak 280 dan kalimat tidak penting sebanyak 857. Kalimat-kalimat penting tersebut kemudian dianotasi dengan notasi BIO menghasilkan jumlah *instance* untuk setiap label seperti pada Tabel 3.

Tabel 2. Distribusi Banyak Kalimat Pada Korpus

Banyak Kalimat	Banyak Teks Berita	Banyak Kalimat	Banyak Teks Berita
4	1	14	11
5	1	15	5
6	4	16	5
7	5	17	2
8	3	18	3
9	6	20	2
10	9	21	3
11	6	22	2
12	10	28	1
13	11		

Tabel 3. Jumlah *Instance* Label 5W1H

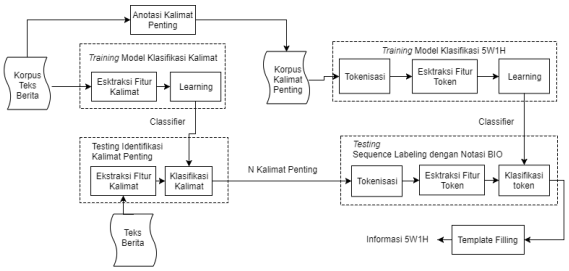
Label	Jumlah	Label	Jumlah
Beg_who	75	Beg_where	74
In_who	364	In_where	127
Beg_what	89	Beg_why	75
In_what	988	In_why	804
Beg_when	83	Beg_how	74
In_when	224	In_how	796
Others	2777		

IV. ANALISIS SOLUSI

Penerapan *sequence labelling* dengan menggunakan notasi BIO yang dilakukan Khodra berhasil mencapai nilai *Fmeasure* sebesar 0.666. Hasil ini menunjukkan adanya potensi bahwa metode yang digunakan mampu menangani karakteristik teks berita Bahasa Indonesia. Penggunaan metode pembelajaran AdaboostM1 yang mampu menangani dataset tidak seimbang tidak memberikan pengaruh yang signifikan. Penggunaan metode pembelajaran tersebut tidak mampu melebihi hasil yang didapatkan dengan metode C4.5. Seleksi fitur yang dilakukan juga belum dapat meningkatkan nilai *Fmeasure*. Salah satu tahapan yang belum digunakan untuk meningkatkan hasil ekstraksi yaitu *filtering*.

Pada penelitian Khodra (2015) keseluruhan kalimat pada teks berita diproses. Padahal dari 1137 kalimat yang terdapat dari korpus tersebut, 630 kalimat diantaranya merupakan kalimat yang semua tokennya adalah *others*. Dengan kata lain kalimat-kalimat tersebut merupakan kalimat yang tidak penting.

Kalimat penting kemudian akan dilabeli dengan *sequence labelling* berdasarkan grup fitur yang digunakan pada penelitian Khodra (2015). Jumlah token yang harus dilabeli menjadi lebih sedikit, sehingga kemungkinan salah melabeli suatu token dapat berkurang. Selain itu teks yang harus dilabeli merupakan kalimat penting yang merepresentasikan *event* utama, sehingga kemungkinan besar informasi 5W1H ada pada beberapa kalimat penting tersebut. Terlebih lagi jumlah token *others* akan berkurang secara signifikan sehingga korpus akan lebih seimbang. Dengan demikian kemungkinan suatu token merupakan informasi 5W1H meningkat. Cara kerja sistem yang dibuat dapat dilihat pada Gambar 1.



Gambar 1. Cara Kerja Sistem

V. EKSPERIMEN

Eksperimen dilakukan untuk mendapatkan nilai *Fmeasure* terbaik yang relevan terhadap kelas-kelas *beginning* dan *in*. 3 Algortima pembelajaran yang digunakan untuk membangun model klasifikasi yaitu AdaBoostM1, C4.5, dan Naïve Bayes. Kakas yang digunakan adalah Weka versi 3.8.1. Filter yang digunakan untuk menangani atribut bertipe *string* adalah *string to* nominal. Untuk melakukan ekstraksi fitur, pada penelitian ini digunakan *library* INA NLP.

Pada eksperimen pemodelan identifikasi kalimat penting, hasil tertinggi didapatkan dengan algoritma

Naïve Bayes. Evaluasi dilakukan dengan menggunakan 10 *folds cross-validation*. Selanjutnya hasil tersebut diseleksi fitur dengan beberapa kombinasi fitur sebagai berikut:

1. *Sentence location, title words overlap rate, term weight,* dan *sentence length*
2. *Sentence location, title words overlap rate, term weight,* dan *number of named entity*
3. *Sentence location, title words overlap rate,* dan *term weight*
4. *Sentence location* dan *title words overlap rate*
5. *Sentence location* dan *term weight*
6. *Sentence location*

Tabel 5. Hasil Eksperimen Seleksi Fitur

	Kombinasi Fitur					
	1	2	3	4	5	6
Fmeasure	0,860	0,860	0,860	0,856	0,860	0,856
Recall	0,862	0,862	0,862	0,857	0,862	0,867
Precision	0,859	0,859	0,859	0,855	0,859	0,855

Hasil pada Tabel 5 menunjukkan bahwa fitur *number of named entity* dan *sentence length* tidak memberikan pengaruh terhadap klasifikasi kalimat penting. Hal ini ditunjukkan pada kombinasi fitur 1, 2, dan 3. Tidak ada perubahan hasil yang didapatkan ketika salah satu atau kedua fitur tersebut diseleksi. Berbeda dengan fitur *term weight*, fitur ini memberikan pengaruh terhadap hasil ekstraksi. Hal tersebut ditunjukkan pada kombinasi fitur 4 dan 6. Nilai *Fmeasure, recall,* dan *precision* menurun ketika fitur *term weight* diseleksi. Fitur lainnya yang tidak memberikan pengaruh terhadap hasil ekstraksi yaitu fitur *title words overlap rate*. Hal tersebut ditunjukkan pada kombinasi fitur 5. Hasil yang didapatkan merupakan hasil tertinggi meskipun fitur ini tidak digunakan. Kombinasi fitur yang dipilih untuk membangun model klasifikasi adalah kombinasi 5, karena dengan kombinasi ini hasil tertinggi didapatkan dengan hanya 2 fitur. Dengan demikian proses ekstraksi fitur akan menjadi lebih cepat.

Eksperimen pemodelan ekstraksi informasi 5W1H dilakukan dengan skema mencari ukuran *window* terbaik untuk kemudian diseleksi fitur. Ukuran *window* terbaik yaitu 2, menghasilkan nilai *Fmeasure* tertinggi menggunakan algoritma C4.5 sebesar 0,855. Namun hasil yang lebih relevan didapatkan dengan algoritma Naïve Bayes dengan *Fmeasure* sebesar 0,831. Hal tersebut dikarenakan nilai *Fmeasure* kelas *beginning* dari algoritma Naïve Bayes lebih baik. Perbandingan nilai *recall* dapat dilihat pada Tabel 6.

Dapat dilihat pada Tabel 6.1 dan 6.2 metode Naive Bayes memiliki nilai *recall* yang lebih baik ketika mengklasifikasikan token ke kelas *beginning*. Metode C4.5 lebih baik ketika mengklasifikasikan token ke kelas *in,* sehingga secara rata-rata nilai *recall* dan *Fmeasure* metode C4.5 lebih baik dari metode Naive Bayes. Pada penelitian ini lebih mempertimbangkan nilai *Fmeasure* ataupun *recall* pada kelas *beginning*. Hal tersebut dikarenakan kelas *beginning* merupakan kelas minoritas. Dengan lebih diutamakannya kelas *beginning,* diharapkan setiap elemen 5W1H dapat diekstrak.

Selanjutnya, hasil dari metode Naive Bayes dengan ukuran *window=2* akan diseleksi fitur. Hasil tertinggi didapatkan dengan menyeleksi fitur NE *tag* dan *isTimePattern* yaitu sebesar 0,832. Nilai *recall* dari seleksi NE *tag* lebih tinggi, namun hasil seleksi *isTimePattern* lebih relevan. Perbandingan nilai *recall* keduanya dapat dilihat pada Tabel 7.

Tabel 6.1. Perbandingan Nilai *Recall* C4.5 dan Naïve Bayes

	beg who	**in who**	**beg what**	**in what**	**beg when**	**In when**	**beg where**
C4.5	0,693	0,896	0,360	0,974	0,675	0,804	0,095
Naive Bayes	0,760	0,882	0,449	0,949	0,855	0,848	0,581

Tabel 6.2. Perbandingan Nilai *Recall* C4.5 dan Naïve Bayes

	beg where	**in where**	**beg why**	**in why**	**beg how**	**in how**	**others**
C4.5	0,095	0,913	0,000	0,898	0,000	0,974	0,886
Naive Bayes	0,581	0,890	0,120	0,881	0,095	0,937	0,800

Tabel 7.1. Perbandingan Nilai *Recall* Seleksi Fitur *isTime Pattern* dan NE *tag*

Kombinasi Fitur	**beg who**	**in who**	**beg what**	**in what**	**beg when**	**In when**	**beg where**
isTimePattern	0,760	0,879	0,449	0,951	0,843	0,839	0,581
NE *tag*	0,760	0,904	0,393	0,952	0,675	0,848	0,473

Tabel 7.2. Perbandingan Nilai *Recall* Seleksi Fitur *isTime Pattern* dan NE *tag*

Kombinasi Fitur	**beg where**	**in where**	**beg why**	**in why**	**beg how**	**in how**	**others**
isTimePattern	0,581	0,890	0,107	0,887	0,095	0,941	0,801
NE *tag*	0,473	0,890	0,120	0,886	0,081	0,947	0,806

Pada penelitian Khodra (2015) digunakan dua algoritma pembelajaran yaitu AdaBoostM1 dan C4.5. Hasil terbaik didapatkan dengan menggunakan algoritma C4.5, ukuran *window*=2, dan tidak dilakukan seleksi fitur.

Perbandingan dengan hasil penelitian Khodra (2015) dapat dilihat pada Tabel 8.

Persentase kenaikan dihitung relatif berdasarkan hasil penelitian Khodra (2015). Metode identifikasi kalimat penting berhasil meningkatkan ekstraksi informasi yang dilakukan dengan BIO *sequence labeling*. Hal ini dikarenakan jumlah *instance* setiap kelas yang dihasilkan setelah dilakukannya tahapan identifikasi kalimat penting menjadi relatif lebih seimbang. Algoritma C4.5 akan menghasilkan nilai *Fmeasure* lebih baik jika jumlah *instance* dari setiap kelas pada korpus yang digunakan untuk membuat model klasifikasi seimbang. Metode identifikasi kalimat penting yang diterapkan berhasil mengurangi 16165 *instance others*.

Tabel 8. Perbandingan Hasil Penelitian

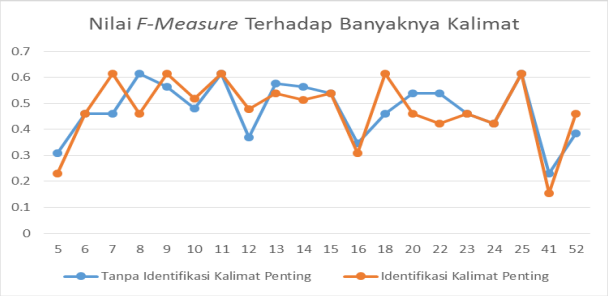
	Khodra (2015)	**Darmawan (2017)**	**Selisih**	**Persentase Kenaikan**
Precision	0,740	0,844	0,104	14,054%
Recall	0,646	0,869	0,223	34,520%
Fmeasure	0,666	0,855	0,189	28,378%

Pada penelitian Khodra (2015) seleksi fitur yang dilakukan tidak berhasil meningkatkan nilai *Fmeasure*. Sedangkan pada tugas akhir ini menyeleksi fitur *isTimePattern* dan NE *tag* berhasil meningkatkan nilai *Fmeasure*. Selisih nilai *Fmeasure* tanpa seleksi fitur dan dengan seleksi fitur dapat dilihat pada Tabel 9. Nilai negatif menunjukkan hasil seleksi fitur menurunkan niali *Fmeasure,* sedangkan nilai positf menunjukkan hasil seleksi fitur meningkatkan nilai *Fmeasure*.

Tabel 9. Selisih Nilai Fmeasure Sebelum dan Setelah Seleksi Fitur

Set Fitur	**Khodra (2015)**	**Darmawan (2017)**
Semua fitur (window=1)	-0,005	-0,005
Semua fitur – leksikal	-0,003	-0,001
Semua fitur – POS *tag*	-0,127	-0,008
Semua Fitur – NE *tag*	-0,074	+0,001
Semua fitur – *left prediction*	-0,596	-0,393
Semua fitur - *isTitleToken*	-0,003	-0,002
Semua Fitur – *isTimePattern*	-0,036	+0,001

Pada penelitian Wang (2012), tahapan identifikasi kalimat penting memiliki kelemahan yaitu akurasi akan menurun pada teks berita yang panjang. Untuk melihat apakah hal tersebut mempengaruhi hasil ekstraksi informasi 5W1H, dilakukan perbandingan nilai *Fmeasure* hasil ekstraksi yang dapat dilihat pada Gambar 2. Data yang digunakan adalah 40 teks berita baru dengan sumber yang sama dengan teks berita yang ada pada korpus. 4 sumber tersebut yaitu Tempo, Republika, Metrotvnews, dan Antara News. Pada setiap sumber diambil 10 teks berita. Distribusi banyaknya kalimat pada 40 teks berita tersebut dapat dilihat pada Tabel 10. Nilai horizontal menunjukkan banyak kalimat pada teks berita sedangkan nilai vertikal menunjukkan *Fmeasure*. Untuk teks berita yang memiliki banyak kalimat yang sama, nilai *Fmeasure* yang digunakan adalah rata-rata dari nilai yang dihasilkan. Pengaruh yang diberikan oleh banyaknya kalimat pada teks berita ternyata tidak signifikan, nilai *Fmeasure* yang dihasilkan tidak selalu menurun. Hal ini dikarenakan fitur yang lebih mempengaruhi hasil ekstraksi kalimat penting yaitu fitur *sentence location*.



Gambar 2. Perbandingan Nilai *Fmeasure* terhadap Banyak Kalimat

Tabel 10. Distribusi Banyak Kalimat

Banyak Kalimat	Banyak Teks Berita	Banyak Kalimat	Banyak Teks Berita
5	1	15	3
6	1	16	2
7	1	18	1
8	1	20	2
9	3	22	2
10	4	23	1
11	1	24	4
12	5	25	1
13	2	42	1
14	3	52	1

Pengujian dengan teks berita baru dilakukan untuk melihat apakah model yang dibuat akan menghasilkan performa yang sama baiknya seperti pada saat *training*. Namun ternyata model yang dibuat tidak mampu mengekstrak semua elemen 5W1H. Beberapa faktor penyebabnya adalah karena tidak semua kalimat penting berhasil dipilih dan filter *string to nominal* yang mengabaikan nilai grup fitur leksikal dari kata-kata yang tidak ada pada korpus. Diabaikannya nilai grup fitur leksikal akan meningkatkan probabilitas suatu kata diklasifikasikan kedalam label yang salah. Selain itu, pemilihan kalimat penting sangat bergantung pada fitur *sentence location*. Jika kalimat yang seharusnya merupakan kalimat penting terletak pada bagian tengah atau akhir teks dan tidak memiliki nilai fitur *term weight* yang cukup tinggi, maka kemungkinan besar kalimat akan diklasifikasikan menjadi kalimat tidak penting.

VI. KESIMPULAN DAN SARAN

Identifikasi kalimat penting meningkatkan nilai *Fmeasure* model ekstraksi informasi 5W1H pada seleksi fitur *isTimePattern* dan NE *tag* sebesar 0,001. Selain itu, identifikasi kalimat penting meningkatkan nilai *Fmeasure* dari model ekstraksi informasi 5W1H sebesar 28,378%.

Identifikasi kalimat penting meningkatkan nilai *Fmeasure* model ekstraksi informasi 5W1H pada seleksi fitur *isTimePattern* dan NE *tag* sebesar 0,001. Selain itu, identifikasi kalimat penting meningkatkan nilai *Fmeasure* dari model ekstraksi informasi 5W1H sebesar 28,378%.

Untuk penelitian selanjutnya, eksplorasi fitur pada tahapan identifikasi kalimat penting yang sesuai dengan karakteristik teks berita Bahasa Indonesia dapat dilakukan untuk meningkatkan nilai *Fmeasure* model klasifikasi kalimat penting. Selain itu, penggunaan filter berbeda ataupun penggunaan algoritma pembelajaran yang dapat menangani atribut dengan tipe *string* dapat digunakan untuk meningkatkan nilai *Fmeasure* model ekstraksi informasi 5W1H.

REFERENCES

Anonim. (2016). Penetrasi dan Perilaku Pengguna Internet Indonesia. Jakarta: APJII.

Chinchor, Nancy & Marsh, Elaine. (1998). *MUC-7 Information Extraction Task Definition (Version 5.1)*. Fairfax, Virginia.

Clark, Alexander dkk. (2010). *The Handbook of Computational Linguistics and Natural Language Processing.* Singapore: Wiley Blackwell.

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

Manning D. Christopher dkk (2008). *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.

Marcus P. M. (1993). *Building a Large Annotated Corpus of English: The Penn Tree Bank.* Cambridge, MA: MIT Press.

Masayu L. Khodra. (2014). Pembangunan Sistem Cerdas Agregator Berita Bahasa Indonesia. ITB, Bandung, Indonesia.

Masayu L. Khodra. (2015). *Event Extraction on Indonesian News Article Using Multiclass Categorization*. ITB, Bandung, Indonesia.

Paul D. Larson. (2007). *Information on Overload: Too Much of a Good Thing.* 92nd Annual International Supply Management Conference, Las Vegas.

Piskorski, J & Yangarber R. *Information Extraction: Past, Present, and Future*.

Tang dkk. (2007). *Information Extraction: Methodologies and Applications.* Beijing, China.

Wei Wang. (2010). *Chinese News Event 5W1H Elements Extraction using Semantik Role Labeling.* Peking University, Beijing, China.

Wei Wang. (2012). *Chinese News Event 5W1H Semantiks Element Extraction for Event Ontology Population*. Peking Universsity, Beijing, China.

Wijanto, M.C., Purwarianti, A. (2013). *Indonesian Dialogue Management System Using Information State For Independent Closed Domain.* ITB, Bandung, Indonesia.