ISSN : 2598-294X (print)
ISSN : 2598-2958 (online)

Perbandingan Algoritma *Conditional Random Field* dan *Hidden Markov* Model pada *Pos Tagging* Bahasa Indonesia

Singgih Briandoko[1], Atika Ratna Dewi [2], Muhammad Akbar Setiawan[3]
[1,2,3]STMIK Widya Utama
[1]briandokosinggih@swu.ac.id,[2]atikaratna@swu.ac.id, [3]akbar@swu.ac.id

Abstract— *Twitter is now an alternative source of real time information for the public. Technological developments cover all aspects of life, one of which is the field of language. Natural Language Processing (NLP) devices developed to support those needs are POS Tagger. This research use 10 tweets and HMM algorithm get 62,7% accuracy level while Conditional Random Field algorithm get 71%. This shows that CRF is better for performing POS tagging in Indonesian on Twitter. HMM and CRF can handle tagging of words that are not in the corpus but the results are not very good.*

Keywords— Twitter, POS Tagging, Conditional Random Field, Hidden Markov Model

1. Pendahuluan

Peningkatan jumlah pengguna internet di Indonesia berkembang semakin pesat. Laporan dari Tetra Pak Index 2017 mencatatkan ada sekitar 132 juta pengguna internet di Indonesia [1]. Hampir setengahnya adalah pengguna media sosial, atau berkisar di angka 40%. Munculnya media sosial seperti Twitter, Facebook, Yahoo, Google, Youtube, Instagram, dan Path mendorong adanya informasi tekstual yang besar dan mendorong pengguna mendapatkan informasi yang akurat dalam penyajian data.

Pengguna Twitter di Indonesia terus meningkat. Indonesia berada pada peringkat 5 pengguna Twitter terbesar di dunia. Berdasarkan data PT Bakrie Telecom, Twitter mempunyai sekitar 19,5 juta pengguna di Indonesia dari total 500 juta pengguna global [2]. Twitter users merupakan konsumen di Indonesia. Biasanya adalah yang tidak memiliki blog atau tidak pernah mengupload video ke Youtube tetapi sering memperbarui status di Twitter dan Facebook.

Twitter saat ini merupakan salah satu layanan mikrobloging terpopuler. Jumlah cuitan orang Indonesia selama Januari hingga Desember 2016 mencapai 4,1 miliar tweet [3]. Data tweet ini dapat menyatakan persepsi publik baik pendidikan, ekonomi, perilaku sosial, fenomena alam, perdagangan yang terjadi di seluruh dunia. Banyak pengguna twitter menuliskan atau mencurahkan apa yang dipikirkan dan rasakan tentang suatu kejadian, atau topik tertentu (tweet). Tweet tersebut dapat digunakan sebagai sumber data untuk menilai sentimen pada twitter. Contoh penggunaan praktisnya yaitu untuk mendapatkan informasi lalu lintas, peristiwa penting, event, dan lain-lain. Sehubungan dengan hal tersebut, studi yang meneliti pengolahan data Twitter menjadi kegiatan yang sangat menarik. Selanjutnya, twitter tidak dipakai hanya untuk mencari informasi saja, akan tetapi berkembang untuk beberapa hal, misalnya untuk mengekstraksi pengetahuan dan juga memprediksi keadaan di masa yang akan datang [4].

Banyak aspek dalam kehidupan sehari-hari yang dikicaukan pada Twitter tidak terkecuali bidang pendidikan. Ujian Nasional yang dilaksanakan menggunakan komputer banyak mendapat respon dari warganet. Penyelenggaraan UNBK dilakukan pertama kali tahun 2014 secara terbatas dan online [5]. Hasil ujiannya cukup memuaskan. Hal ini mendorong dalam peningkatan kemampuan siswa terhadap TIK (Teknologi Informasi dan Komunikasi). Namun, tidak sedikitnya kendala membuat banyak pro kontra pada masyarakat. Hal ini dapat dilihat dari banyaknya cuitan menggunakan tagar UNBK.

Tag adalah label (tag) yang diberikan pada tiap kata yang menyatakan kelas kata tersebut. Dengan tag, dapat diketahui kelas dari suatu kata, termasuk juga sifat-sifat apa yang cenderung melekat pada kata dalam kelas tersebut. Informasi ini dapat berguna dalam menentukan makna suatu kata, menentukan aturan tata bahasa suatu kalimat atau frasa. Tag juga berperan besar dalam mengatasi keambiguan makna kata ataupun kalimat.

POS tagging merupakan proses otomatis menetapkan kategori leksikal untuk setiap tanda atau kata dalam kalimat sesuai dengan definisi serta konteksnya. POS tagging sangat berguna dalam mempengaruhi banyak aplikasi dari Natural Language Processing seperti parsing, ekstraksi informasi, disambiguasi makna dll. Tagging berarti menetapkan kelas gramatikal yaitu bagian yang tepat dari tag setiap kata dalam kalimat. Menetapkan tag untuk setiap kata dari teks dengan manual sangat memakan waktu [6]. Ada beberapa pendekatan yang dapat digunakan untuk tagging, yaitu pendekatan berdasarkan aturan (rule based), pendekatan probabilistik, dan pendekatan berbasis transformasi (transformation based).

Teknik probabilistik untuk tag dapat menggunakan Hidden Markov Model (HMM). Karena, proses tagging dapat mengklasifikasi suatu rangkaian atau urutan tag untuk setiap kata dalam suatu kalimat [7]. HMM adalah sebuah proses stokastik ganda dimana

ISSN : 2598-294X (print)
ISSN : 2598-2958 (online)

salah satu prosesnya tidak dapat diobservasi (hidden). HMM sering digunakan untuk memodelkan data sequential/temporal, dimana proses berjalan seiring waktu. HMM dapat menggabungkan dua atau lebih rantai Markov dengan hanya satu rantai yang terdiri dari state yang dapat diobservasi dan rantai lainnya membentuk state yang tidak dapat diobservasi (hidden), yang mempengaruhi hasil dari state yang dapat diobservasi dan memilih tag sequence terbaik untuk word sequence yang diamati.

Teknik probabilitas lainnya adalah Conditional Random Field. CRF merupakan framework untuk membangun model probabilistik untuk segmentasi dan pelabelan data yang berurutan. Kelebihan utama CRF adalah sifat bersyaratnya (conditional), sehingga mengurangi ketergantungan atau asumsi untuk memastikan inferensi mudah dikerjakan. Selain itu CRF juga menghindari masalah label bias [8]. Penelitian ini ditujukan untuk membandingkan CRF dan HMM dalam proses POS tagging Bahasa Indonesia pada Twitter. Peneliti menggunakan rule dalam mengetahui apakah sebuah merupakan opini atau bukan opini.

2. Metode Penelitian

2.1. Teknik Pengumpulan Data

Pengumpulan data dilakukan dengan mengumpulkan tweet pada taggar #UNBK satu per satu. Tweet dikumpulkan dari tanggal 1 April – 1 Agustus 2017, yang akan dipilih secara acak. Penelitian ini membutuhkan 2 macam data. Data pertama adalah data latih dan data kedua adalah data uji. Data ini diperoleh dari hasil studi yang telah diteliti oleh Universitas Indonesia (UI) sebagai wakil dari Indonesia dalam proyek Pan Localization (PANL10N) [20]. Kalimat – kalimat ditokenisasi ulang dengan memperhatkan ekspresi frase menggunakan kamus bahasa Indonesia kateglo. Korpus terdiri dari 10.000 kalimat dengan 256683 token dan 23 tagset [20]. Data yang kedua adalah data yang diambil dari twitter dengan 100 dokumen tweet sebagai data uji.Metode yang digunakan untuk mengumpulkan data dalam penelitian ini adalah metode wawancara, literatur dan dokumentasi.

Tabel 2.1. Data yang digunakan

	Data Latih	Data Uji
Jumlah Kalimat	10.000	100
Jumlah Token	256.683	1213

Tabel 2.1 menunjukan data yang akan digunakan sebagai data latih dan data uji. Data latih sebanyak 10.000 kalimat dengan jumlah token ada 256.683 sedangkan untuk data uji sebanyak 100 dokumen tweet dengan jumlah token sebanyak.

2.2. Preprocessing

Tahapan pada text preprocessing yang dilakukan adalah

a. Stopword Removal

Proses menghilangkan kata umum (common words) yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna.

b. Case Folding

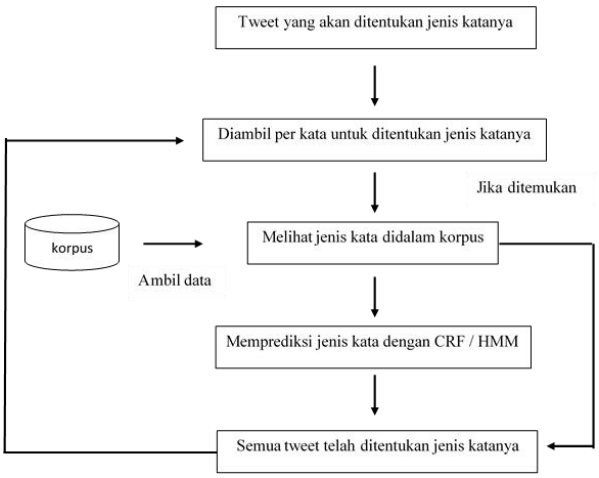
Case folding merupakan proses penyamaan huruf dalam sebuah dokumen, dilakukan untuk mempermudah pencarian. Tidak semua teks konsisten dalam penggunaan huruf capital. Maka dari itu peran case folding sangat penting dalam mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar. Case folding mengubah semua huruf dalam dokumen menjadi huruf kecil, hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap delimeter.

c. Tokenization

Proses pemecahan kalimat yang ada dalam sebuah dokumen menjadi kata kata. Setiap kata yang telah dipisah disebut token/term. Kata-kata ini digunakan untuk proses pencarian. Misalnya, teks “saya tidak setuju” akan dipecah menjadi “saya”,”tidak”,”setuju”.

2.3. Proses POS-Tagging

Pada penelitian ini menggunakan algoritma Conditional Random Field dalam memprediksi kelas kata.



Gambar 2.1. Kerangka Pemikiran

3. Hasil dan Pembahasan

Pengumpulan data yang dilakukan dengan mengambil tweet dengan menggunakan taggar #unbk satu per satu dari tanggal 1 April - 1 Agustus 2016,

ISSN : 2598-294X (print)
ISSN : 2598-2958 (online)

yang akan dipilih secara acak. Data tweet tersebut belum dilakukan preprocesing sehingga tweet masih terdapat kata-kata atau tanda baca yang tidak perlu.

Tabel 3.1. Sampel tweet yang digunakan

Username	Tweet
@aides_aigipty	katakan tidak untuk #UNBK
@seloniregina04	Semoga lulus semua #UNBK
@iqbaal_harits	Untuk apa ujian #UNBK
@imronFR	Kepala terasa pusing #unbk
@patasdotid	Ujian nasional berbasis kenangan #UNBK
@suaradesa	Memang hidup penuh ujian #UNBK
@Indahsulis17	Matematika susah parah #UNBK
@jpnncom	2 hari lagi semangat #UNBK
@WAHYURR_9	Sing penting yakin #UNBK
@ribqi_metrix	Senang bisa bekerja sama #UNBK

Selanjutnya dari 100 tweet yang akan diproses, peneliti mengambil 10 sampel tweet seperti yang ditunjukan pada tabel 4.1. proses berikutnya adalah preprosesing agar tweet bersih dari kata-kata yang tidak perlu dan siap untuk dicari kelas katanya.

3.1. Conditional Random Field

Conditional Random Field berfungsi untuk mendeteksi tag yang sesuai pada kalimat atau frasa. CRF menggunakan HMM dan juga feature function dari Maximum Entropy.

Tabel 3.2 kelas kata yang dicari

Katakan	**Tidak**	**Untuk**
VB	NEG	SC/IN?

Tabel 3.2 menunjukkan sebuah kalimat berisi 3 kata dan 4 tag. Kalimat diatas diambil dari salah satu data testing. Kata "untuk" memiliki 2 tag. Maka akan dicari susunan kelas kata / tag. Kata "untuk" muncul sebagai "SC" dan "IN" berdasarkan sampel korpus. Maka akan dicari urutan tag terbaik atau kelas kata yang sesuai dengan kalimat tersebut.

3.2.1. Probabilitas Transisi

Dari sampel corpus diatas bisa dilihat terdapat tag SC sebanyak 16 kali dan yang diikuti oleh VB sebanyak 1 kali. Sedangkan jumlah VB yang diikuti tag SC muncul sebanyak 7 kali.

$$P(VB|SC)=\frac{C(VB,SC)}{C(SC)}=\frac{1}{16}=0{,}063$$

$$P(SC|VB)=\frac{C(SC,VB)}{C(VB)}=\frac{7}{54}=0{,}13$$

3.2.2. Probabilitas Emisi

Probabilitas emisinya P(wi|ti), menunjukan kemungkinan diberikan pada tag (JJ), akan sangat berhubungan dengan sebuah kata (bobot). Akan dicari berapa nilai kemungkinanya untuk kata "untuk" yang diberi tag "SC / IN". kemungkinan maksimum dari probabilitas emisinya adalah:

$$P\ (untuk|SC) = \frac{C(SC,untuk)}{C(SC)} = \frac{C(4)}{C(14)} = 0.29$$

$$P\ (untuk|IN) = \frac{C(IN,untuk)}{C(IN)} = \frac{C(3)}{C(4)} = 0.75$$

Probabilitas emisi dari masing-masing kata ditunjukan pada tabel 3.3.

Tabel 3.3. Probabilitas Emisi

State	observation		
	katakan	**tidak**	**untuk**
VB	0,019	0	0
NEG	0	0,5	0
SC	0	0	0,286
IN	0	0	0,75

Setelah didapat nilai dari probabilitas transisi dan probabilitas emisi, maka langkah selanjutnya untuk menentukan urutan tag yang benar atau decoding menggunaan algoritma viterbi. Untuk pencarian urutan tag terbaik dilihat pada tabel 4.7.

3.2.3. Bobot feature probabilitas transisi dan emisi

Bobot diperoleh dari regresi logistik, dengan menghitung tag sebelum atau tag yang mengikuti tag yang dicari.

1. Pembuatan feature

Sebelum pembuatan fiture perlu dilihat kelas kata sebelum dan setelah kata yang akan kita cari kelas katanya. Setelah dilakukan maka kita buat fiturnya.

ISSN : 2598-294X (print)
ISSN : 2598-2958 (online)

$$f1 \begin{cases} 1 \text{ jika } wi=\text{untuk} \\ 0 \text{ jika selain itu} \end{cases} \quad C=SC$$

$$f2 \begin{cases} 1 \text{ jika } w-1=\text{tidak} \\ 0 \text{ jika selain itu} \end{cases} \quad C=NEG$$

$$f3 \begin{cases} 1 \text{ jika } w-2=\text{katakan} \\ 0 \text{ jika selain itu} \end{cases} \quad C=VB$$

$$f4 \begin{cases} 1 \text{ jika } wi=\text{untuk} \\ 0 \text{ jika selain itu} \end{cases} \quad C=IN$$

Menghitung bobot tiap feature transisi

Setelah fitur berhasil dibuat langkah selanjutnya adalah menghitung bobot setiap fitur dengan menggunakan regresi logistik. Tiap fitur mempunyai bobot yang sesuai. Sehingga bobot w1(c,x) akan menunjukan seberapa kuat kata "untuk" sebagai tag SC atau IN. Bobot w-1, w-2(c,x) akan menunjukan seberapa kuat tag sebelum"tidak" dan "katakan" untuk mengukur kata "untuk" menjadi SC atau IN.

Tabel 3.4 Hasil perhitungan bobot feature transisi

	VB	NEG	SC	IN
VB	1,3	0,2	1,3	1,4
NEG	0,75	0	0	0
SC	1,8	0	0	0
IN	0,9	0	0	0

3.3. HMM

Pada HMM, menghitung probabilitas dengan menghitung tag pada corpus. Pada asumsi yang pertama, probabilitas tag transisi dipengaruhi oleh tag sebelumnya. Berikutnya menghitung tag pada awal kalimat. Untuk memudahkan contoh perhitungan pada penelitian ini menggunakan sampel kalimat corpus.

Tabel 3.5 kelas kata yang dicari

Katakan	**Tidak**	**Untuk**
VB	NEG	SC/IN?

Pada tabel 3.5 diatas merupakan kalimat dengan susunan kelas kata yang akan dicari. Terdapat kata "untuk" yang akan dicari kelas katanya. Pada corpus tabel 3.2 tersebut kata "untuk" muncul sebagai "SC, dan IN". Maka akan dicari kelas kata yang sesuai atau cocok dengan kalimat tersebut.

3.3.1. Probabilitas Transisi

Dari sampel corpus diatas bisa dilihat terdapat tag SC sebanyak 16 kali dan yang diikuti oleh VB sebanyak 1 kali. Sedangkan jumlah VB yang diikuti tag SC muncul sebanyak 7 kali. Jadi untuk mencari probabilitas transisinya adalah:

$$P(VB|SC)=\frac{C(VB,SC)}{C(SC)}=\frac{1}{16}=0{,}063$$

$$P(SC|VB)=\frac{C(SC,VB)}{C(VB)}=\frac{7}{54}=0{,}13$$

3.3.2. Probabilitas Emisi

Probabilitas emisinya P(wi|ti), menunjukan kemungkinan diberikan pada tag (JJ), akan sangat berhubungan dengan sebuah kata (bobot). Akan dicari berapa nilai kemungkinanya untuk kata "untuk" yang diberi tag "SC / IN". kemungkinan maksimum dari probabilitas emisinya adalah:

$$P\ (\text{untuk}|SC)=\frac{C(SC,\text{untuk})}{C(SC)}=\frac{C(4)}{C(14)}=0.29$$

$$P\ (\text{untuk}|IN)=\frac{C(IN,\text{untuk})}{C(IN)}=\frac{C(3)}{C(4)}=0.75$$

Probabilitas emisi dari masing-masing kata ditunjukan pada tabel 3.6

Tabel 3.6 Probabilitas Emisi

State	observation		
	katakan	**tidak**	**untuk**
VB	0,019	0	0
NEG	0	0,5	0
SC	0	0	0,286
IN	0	0	0,75

4. Kesimpulan

Dari penelitian yang telah dilakukan menggunakan 10 tweet diperoleh algoritma HMM mendapatkan tingkat akurasi 62,7% sedangkan algoritma CRF yang mendapatkan 71%. Ini menunjukan bahwa CRF lebih baik untuk melakukan POS tagging bahasa Indonesia di twitter. HMM dan CRF dapat menangani pemberian tag pada kata yang tidak terdapat didalam corpus tapi tidak begitu baik. Ada beberapa kesalahan dalam pemberian tag pada katanya.

DAFTAR PUSTAKA

[1] Inet.detik.com. (2017, 27 September). 132 Juta Pengguna Internet Indonesia, 40% Penggila Medsos. Diperoleh 13

ISSN : 2598-294X (print)
ISSN : 2598-2958 (online)

Februari 2018, dari https://inet.detik.com/cyberlife/d-3659956/132-juta-pengguna-internet-indonesia-40-penggila-medsos/

[2] Kominfo.go.id. (2017, 11 Januari). Pengguna Internet di Indonesia 63 Juta Orang. Diperoleh 12 Februari 2018, dari https://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker/

[3] Tekno.liputan6.com. (2016, 6 Desember). Netizen Indonesia Cuitkan 4,1 Miliar Tweet Sepanjang 2016. Diperoleh 12 Februari 2018, dari http://tekno.liputan6.com/read/2671236/netizen-indonesia-cuitkan-41-miliar-tweet-sepanjang-2016/

[4] S. E. Yuda Munarko, Yufis Azhar, Maulina Balqis, "POS Tagger Tweet Bahasa Indonesia," Kinet. Inform. Univ. Muhammadiyah Malang, vol. 2, no. 1, pp. 2 – 3, 2016.

[5] unbk.kemdikbud.go.id. (2015, 5 Maret). Ujian Nasional Berbasis Komputer. Diperoleh 11 Desember 2017, dari https://unbk.kemdikbud.go.id/tentang#content/

[6] F. M. Hasan, N. Uzzaman, and M. Khan, "Comparison of different POS Tagging Techniques (-Gram , HMM and Brill ' s tagger) for Bangla," Corpus.

[7] D. Jurafsky and J. Martin, "Hidden Markov Models," Speech Lang. Process., no. Chapter 20, p. 21, 2017.

[8] F. Saefulloh, "Part of Speech Tagger untuk Bahasa Indonesia Menggunakan Conditional Random Field (CRF). Universitas Komputer Indonesia," 2017.

ISSN : 2598-294X (print)
ISSN : 2598-2958 (online)