Pembangunan *Languange Modeling* dengan *Probability Based Learning*

Raden Rizky F.P – 1301154211
ICM-39-GAB
School of Computing, Telkom University

I. INTRODUCTION

Bahasa adalah kemampuan yang dimiliki manusia untuk berkomunikasi dengan manusia lainnya menggunakan tanda, misalnya kata dan gerakan [1]. Bahasa terbagi menjadi 2 jenis, yaitu bahasa tertulis dan bahasa lisan. Dari segi konteksnya, bahasa lisan adalah suatu bahasa yang direpresentasikan dalam bentuk suara yang dapat didengar. Sedangkan bahasa tertulis adalah suatu bahasa yang direpresentasikan dalam suatu teks yang dapat dibaca oleh manusia.

Seiring dengan berkembangnya zaman, penggunaan bahasa berupa teks semakin banyak digunakan. Hal itu dikarenakan teks tidak hanya dapat sebagai penghubung antara manusia untuk saling berkomunikasi, akan tetapi teks juga dapat menjadi suatu penghubung antara manusia dan komputer untuk saling berkomunikasi juga. Ilmu yang membahas interaksi antara manusia dengan komputer dengan menggunakan suatu data berupa teks (bahasa) adalah *Natural Language Processing (NLP).*

Meskipun komputer sudah dapat memproses suatu teks dengan menggunakan metode yang ada pada NLP, akan tetapi hal tersebut belum memastikan model yang dibuat sudah berjalan dengan baik atau tidak. Itu disebabkan karena jumlah teks bahasa yang sangat beragam dan berubah-ubah apabila ada kata baru yang muncul. Sehingga sangatlah penting untuk membangun suatu kamus yang dapat menyimpan data teks tersebut.

Tujuan dari pembelajaran ini adalah untuk membangun suatu *Language Modeling* berbasis probabilitas antar setiap kata yang ada didalam kamus kata. Dengan menerapkan metode unigram dan bigram untuk pembuatan model probabilitas.

II. METHODOLOGY

A. Unigram

Unigram adalah suatu teknik untuk membuat suatu kamus bahasa berdasarkan tiap-tiap kata yang ada didalam suatu kalimat. Kata-kata tersebut kemudian direpresentasikan dalam bentuk probabilitas kemunculan ketika suatu data teks diproses. Untuk prosesnya unigram dapat dirumuskan sebagai berikut:

$$P(w_n) = \frac{count(w_n)}{total(w)} \qquad (1)$$

B. Bigram

Bigram adalah suatu teknik untuk membuat kamus bahasa berdasarkan keterkaitan kata awal dengan kata selanjutnya. Sama halnya dengan unigram, kata-kata tersebut direpresentasikan dalam bentuk probabilitas kemunculan. Untuk prosesnya bigram dapat dirumuskan sebagai berikut:

$$P(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count\ (w_{i-1})} \qquad (2)$$

C. Perplexity

Perplexity adalah suatu pengukuran yang digunakan untuk mengetahui seberapa bagus suatu model probabilitas ketika memprediksi sebuah sample [2]. Perplexity juga dapat digunakan untuk membandingkan beberapa model probabilitas. Semakin kecil nilai perplexitynya maka semakin bagus distribusi probabilitasnya ketika memprediksi sebuah sample.

Persamaan perplexity yang digunakan untuk model bigram yang dibuat adalah sebagai berikut:

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}} \qquad (3)$$

III. ANALYSIS

A. Word Prediction

Berdasarkan model yang dibuat dengan acuan penggunaan unigram dan bigram dapat disimpulkan bahwa suatu teks dapat dimodelkan dengan menggunakan dua buah metode tersebut. Dalam penerapan prediksi kata selanjutnya, bigram memiliki peran penting untuk menentukan probabilitas kata selanjutnya.

Hal itu dikarenakan, ketika melakukan prediksi ada sebuah proses pengecekan kata. Pengecekan kata tersebut memiliki tujuan untuk menentukan kata yang memiliki probabilitas yang sangat tinggi. Apabila hanya satu kata yang memiliki nilai probabilitas yang tinggi, maka kata tersebut menjadi hasil prediksi kata selanjutnya. Sedangkan apabila terdapat kata yang memiliki probabilitas yang sama maka kata yang menjadi hasil prediksi adalah kata yang pertama kali muncul.

Berdasarkan model yang dibuat hasil yang didapat ketika melakukan testing pada 10 kata adalah sebagai berikut:

Table 1 Hasil Prediksi

Testing Word	Next Word (Prediction)
abdulgani	sebagai
ac	milan
prabowo	subianto
seorang	wanita
kota	Malang
jokowi	Di
agama	mengeluarkan
kangkung	pakai
acara	ayo
akhir	Desember

Pemilihan kata yang dilakukan untuk proses prediksi diambil berdasarkan list kata yang berada di model unigram. Proses pengambilan kata tersebut diplih secara acak (manual) dan kata yang merupakan kata penghubung tidak dipilih meskipun kemunculannya paling banyak. Hal itu disebabkan karena kata penghubung seharusnya dihapus ketika melakukan proses *preprocess* dengan melakukan proses stopword pada kata tersebut karena tidak memiliki makna yang cukup berarti.

B. Testing Perplexity

Pada pengujian testing perplexity terhadap lima buah kalimat didapat hasil sebagai berikut:

Table 2 Hasil Perplexity

Kalimat	Perplexity
beli kangkung pakai uang mainan animah gegerkan pasar	4.174032625646324+0j
masyarakat malang akan sakit jika wali kota jadi terpidana	13.771800908397672-13.771800908397672j
kpk tahan wali kota nonaktif malang	3.610266933420735+0j
segudang khasiat daun salam bagi kesehatan	3.8191320426701147+0j
membangun start up tidak terbatas usia	5.9777197589822+0j

Dari data tabel diatas, nilai perplexity yang dihasilkan rata-rata bernilai kecil. Hal tersebut meyimpulkan bahwa model probabilitas yang dibuat dengan menggunakan bigram sudah baik dalam memprediksi suatu kata/kalimat. Hal itu didasarkan pada konsep aturan perplexity yang sudah dijelaskan sebelumnya. Semakin kecil nilai dari perplexity maka semakin bagus model yang dibuat.

Kalimat yang digunakan untuk melakukan testing perplexity adalah kalimat headline/title yang berasal dari dataset yang ada.

REFERENCES

[1] Wikipedia, “Wikipedia,” [Online]. Available: https://id.wikipedia.org/wiki/Bahasa. [Diakses 09 September 2018].

[2] Wikipedia, “Wikipedia,” [Online]. Available: https://en.wikipedia.org/wiki/Perplexity. [Diakses 10 September 2018].

[3] J. Domokos, “Text Conditioning and Statistical Language Modeling for Romanian Language,” IEEE, Rinabua, 2009.

[4] D. Jurafsky, “Language Modeling Slide,” Standford Univeristy.