Entropy Naskah Bahasa Sunda Dan Bahasa Jawa Untuk Kompresi Teks Menggunakan Algoritma *Binary Huffman Code*

Syura Fauzan[1], Muhammad Saepulloh[2], Nanang Ismail[3], Eki Ahmad Zaki Hamidi[4]
[1,2,3,4]Teknik Elektro UIN Sunan Gunung Djati Bandung,
Jl. A. H. Nasution 105 Bandung 40614 INDONESIA Telp. 022-7800525 Fax. 022-7803936.
fauzansyura@gmail.com[1], saepul327@gmail.com[2], nanang.is@uinsgd.ac.id[3],
ekiahmadzaki@uinsgd.ac.id[4]

Abstrak – *Kompresi adalah suatu teknik untuk memperkecil jumlah ukuran data dari data aslinya dengan tujuan agar lebih efektif dan lebih kecil dalam penyimpanan serta efisien dan lebih cepat dalam proses pentransmisian data. Paper ini membahas studi Entropy Bahasa Sunda dan Bahasa Jawa untuk kompresi teks. Tujuan yang ingin dicapai dari penelitian ini adalah dapat mengetahui Entropy bahasa Sunda sehingga dapat menjadi dasar untuk kompresi teks. Dalam makalah ini, digunakan algoritma Binary Huffman Code untuk menganalisis nilai entropy Bahasa Sunda dan Jawa. Pengkodean dengan metode Huffman Code dibangun dari panjang variabel kode-kode yang disusun dari bit-bit. Simbol yang memiliki nilai probabilitas lebih tinggi akan memperoleh kode-kode paling pendek, sedangkan simbol yang memiliki nilai probabilitas lebih rendah akan memperoleh kode-kode paling panjang. Hasil analisis menunjukan bahwa Entropy Bahasa Sunda sebesar 4.186 bits per simbol, sedangkan Entropy bahasa Jawa sebesar 4.101 bits per simbol.*

Kata kunci*: Kompresi Teks, Entropy, Bahasa Sunda, Bahasa Jawa, Binary Huffman Code*

1. Pendahuluan

Bahasa adalah sebuah sarana komunikasi yang efektif. Menurut Menezes et al (1997) semua bahasa natural redundant dalam hal struktur bahasanya. Didalam bahasa terdapat simbol yang kemunculannya lebih sering dibandingkan dengan simbol atau pasangan simbol yang lain. Sering atau tidaknya suatu simbol atau pasangan simbol itu muncul dapat diketahui dengan melihat distribusi frekuensinya [1].

Setiap sistem komunikasi mengikuti proses transmisi pesan yang umum dari satu titik ke titik lainnya. Sifat statistik tersembunyi dari proses komunikasi pertama-tama diakui oleh ahli matematika hebat *Claude Elwood Shannon*. Salah satu ciri terpenting teori Shannon adalah konsep *entropy*, yang ia tunjukkan setara dengan kekurangan konten informasi dalam sebuah pesan. Dengan demikian banyak kalimat bisa disingkat secara singkat tanpa kehilangan artinya [2].

Entropy adalah parameter statistik yang mengukur berapa banyak informasi yang dihasilkan rata-rata untuk setiap huruf dari sebuah teks dalam bahasa tersebut. Setiap bahasa biasanya memiliki beberapa fitur penting yang tersembunyi secara statistik dan redundansi tertentu. Fitur-fitur ini dapat dimanfaatkan untuk membentuk alat kompresi teks yang sesuai untuk penggunaan sumber daya yang optimal [3].

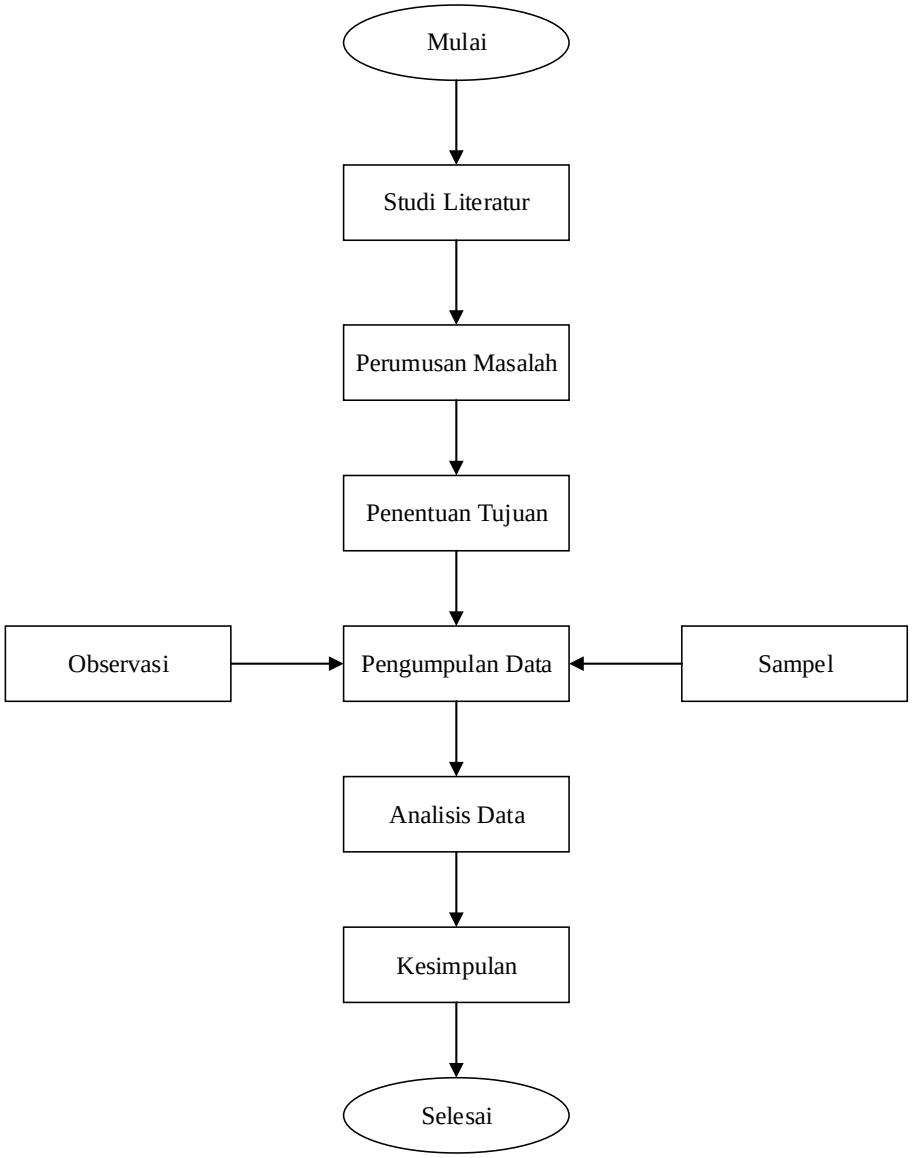
Kompresi diperlukan untuk penyimpanan informasi yang efektif dan untuk transmisi yang lancar melalui saluran. Setiap algoritma kompresi mencoba untuk mewakili input pesan dalam bentuk baru dengan jumlah bit yang lebih sedikit dengan memanfaatkan distribusi probabilitas. Ada dua jenis kompresi data, yaitu kompresi data sama dengan aslinya (*lossless data*

compression) dan kompresi data tidak sama dengan aslinya (*lossy data compression*). Salah satu teknik kompresi dari *lossless data compression* adalah *Huffman Code* [4].

Dalam paper ini, kompresi teks yang diusulkan mengikuti teknik pengkodean *Huffman Code*. *Huffman Code* adalah kode prefiks optimal (*shortest expected length*) untuk alfabet yang diberikan dimana setiap simbol dalam alfabet telah memiliki probabilitas kemunculan masing-masing [3]. *Huffman Code* dibangun dari panjang variabel kode-kode yang disusun dari bit-bit. Simbol yang memiliki nilai probabilitas lebih tinggi, maka akan memperoleh *codeword* yang lebih pendek. Sedangkan, simbol yang memiliki nilai probabilitas lebih rendah, maka akan memperoleh *codeword* yang lebih panjang. Dua simbol yang memiliki nilai probabilitas terendah akan memperoleh panjang *codeword* yang sama.

2. Metode Penelitian

Metode yang digunakan dalam penelitian ini dapat dilihat pada gambar 1.



Gambar 1. Metode Penelitian

ISBN: 978-602-512-810-3

1. Studi Literatur

Pada tahap kajian literatur dilakukan pengkajian terhadap bahasan dalam penelitian, yaitu tentang kompresi data dengan teknik *Huffman Code*. Dalam tahapan ini, berbagai referensi dikumpulkan sebagai dasar penelitian dan penunjang pelaksanaan penelitian sekaligus sebagai tinjauan dalam pemecahan masalah yang akan ditemui nantinya. Bahan-bahan studi literatur ini berasal dari hasil penelitian-penelitian yang telah dilakukan maupun jurnal-jurnal ilmiah dan buku.

2. Perumusan Masalah dan Penentuan Tujuan

Perumusan masalah berhubungan erat dengan beberapa persoalan yang muncul serta perencanaan dan konsep untuk menyelesaikan masalah tersebut baik secara teoritis maupun analisis. Dalam tahapan ini, masalah dirumuskan dalam pokok permasalahan. Tujuan ditentukan sesuai dengan permasalahan yang ada untuk mencari solusi yang tepat.

3. Pengumpulan Data

Pengumpulan data meliputi pengambilan sampel dan observasi. Sampel penelitian dalam penelitian ini diambil dari artikel bahasa sunda. Semua simbol yang muncul dalam artikel tersebut dihitung frekuensi, probabilitas dan diberikan *codeword*-nya menggunakan algoritma *Huffman Code*. Sampel yang telah diambil dan permasalahan yang ada diobservasi didalam sebuah ruangan atau laboratorium yang disandarkan sesuai dengan studi literatur.

4. Analisis Data dan Kesimpulan

Hasil yang telah didapatkan dari penelitian lalu dianalisis untuk mendapatkan kesimpulan yang tepat. Kemudian hasil tersebut disajikan dalam bentuk laporan penelitian.

Selain menggunakan metode yang telah dipaparkan diatas, penelitian ini dilakukan menggunakan algoritma *Huffman Code* dalam pemberian codeword untuk masing-masing simbol. Algoritma *Huffman Code* dapat dilihat pada gambar 2.

1. Buatlah suatu barisan terurut yang terdiri dari data yang berisi simbol serta probabilitas simbol dimulai dari probabilitas simbol tertinggi menuju probabilitas simbol terendah.
2. Ambil 2 data dengan probabilitas yang paling rendah dan jumlahkan probabilitasnya.
3. Buatlah satu *node* baru dengan cabang 2 data tersebut.
4. *Node* tersebut disisipkan ke dalam barisan dengan posisi yang sesuai agar barisan tetap dalam keadaan yang terurut.
5. Jika didalam barisan terdapat data dengan probabilitas yang sama dengan probabiitas hasil penjumlahan 2 data tadi, maka *node* baru tersebut disisipkan pada posisi yang lebih tinggi dari data yang sama tadi. Hal ini berguna untuk mendapatkan kode yang lebih efektif dan efisien.
6. Ulangi langkah 2, 3, 4 dan 5 sampai semua simbol diproses dan didapatkan bentuk pohon biner yang utuh dengan jumlah probabilitas 1.
7. Pembentukan kode untuk setiap simbol dilakukan dengan memberikan nilai 0 untuk cabang atas dan memberikan nilai 1 untuk cabang bawah.

Gambar 2. Algoritma *Huffman Code*

(Sumber: Y. E. Hutasoit, "Huffman Coding untuk Kompresi Data Teks Berbahasa Indonesia," 2001)

ISBN: 978-602-512-810-3

3. Hasil dan Analisis

a. *Source Compression*

Source Compression yang digunakan dalam penelitian ini diambil dari sebuah sampel artikel bahasa sunda dan bahasa jawa. Total simbol untuk bahasa sunda sebanyak 1505 simbol, sedangkan bahasa jawa sebanyak 1554 simbol.

b. *Entropy*

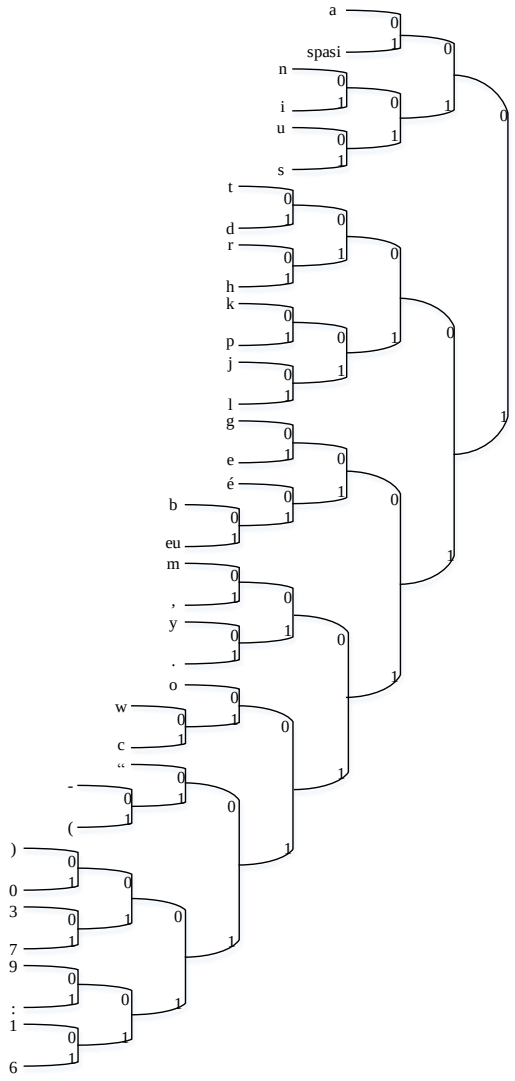
Dari banyaknya frekuensi (*F*) masing-masing simbol yang muncul dan probabilitas (*P*) tiap karakter akan didapatkan *entropy* (*H*) seperti yang diperlihatkan pada tabel 1.

Tabel 1. Bahasa dan *Entropy*-nya

Bahasa	**Entropy (*H*)**	**Expected Code Length (*L*)**
Sunda	4.168	4.242
Jawa	4.101	4.122

c. Bahasa Sunda

Codeword (*C*) dan *length codeword* (*LC*) akan didapatkan dari *Huffman Tree* yang diperlihatkan pada gambar 3.



Gambar 3. *Binary Huffman Tree Code* Bahasa Sunda

ISBN: 978-602-512-810-3

Aturan main dalam penggunaan *Tree Huffman*, mengacu pada algoritma *Binary Huffman Code* yang telah dijelaskan pada gambar 2. Dua simbol dengan probabilitas terendah dijumlahkan sehingga membentuk node baru. Simbol "**6**" (tanpa petik, simbol yang muncul dalam artikel) dijumlahkan dengan simbol "**1**". Simbol **:** dijumlahkan dengan simbol "**9**". Keduanya dijumlahkan sehingga diperoleh node baru. Simbol "**7**" dijumlahkan dengan simbol "**3**". Simbol "**0**" dijumlahkan dengan simbol "**)**". Keduanya dijumlahkan sehingga diperoleh node baru. Simbol "**(**" dijumlahkan dengan simbol "**-**" dan dijumlahkan pula dengan simbol **"** sehingga diperoleh node baru. Ulangi hal yang sama untuk simbol yang lainnya. Penjumlahan dilakukan untuk dua simbol atau node yang memiliki nilai probabilitas sama atau nilai probabilitas yang atas lebih tinggi daripada nilai probabilitas yang bawah. Node-node tersebut dijumlahkan dengan node-node lainnya sampai diperoleh *Tree Huffman* yang utuh dengan jumlah probabilitas 1. Pembentukan *codeword* untuk setiap simbol dilakukan dengan memberikan nilai 0 untuk cabang atas dan memberikan nilai 1 untuk cabang bawah. Pemberian *codeword* untuk setiap simbol dilakukan dengan mengambil urutan bit dari akar menuju daun.

Expected code length (*L*) pada tabel 2 didapatkan dari jumlah perkalian antara probabilitas (*P*) tiap karakter dan *length codeword* (*LC*) nya yang telah diperoleh dari *Binary Huffman Tree Code* pada gambar 3.

Tabel 2. *Expected Code Length Binary Huffman Code* Bahasa Sunda

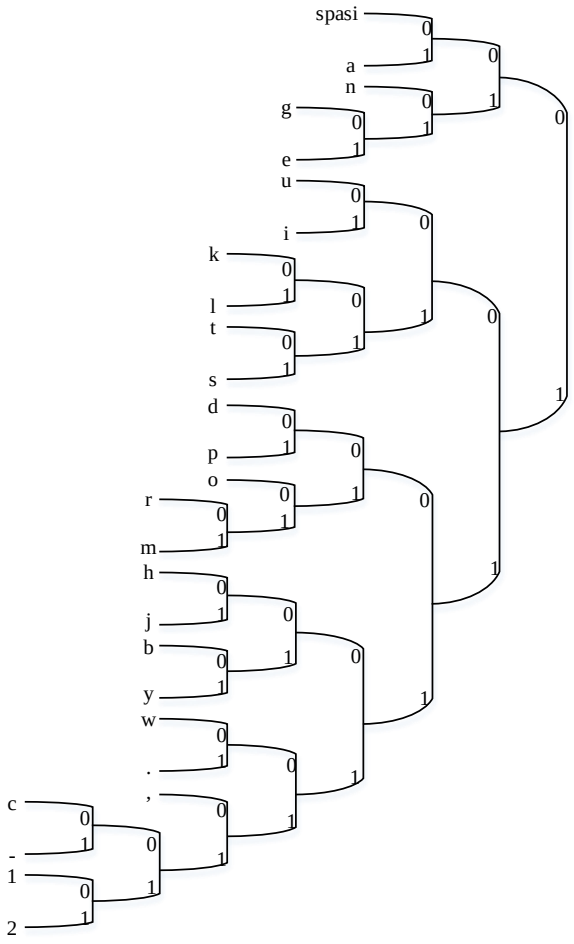
Simbol	**Probabilitas**	***Length Code***	***Codeword***	***Expected Code Length***
a	0.192	3	000	0.576
spasi	0.155	3	001	0.465
n	0.084	4	0100	0.336
i	0.056	4	0101	0.224
u	0.047	4	0110	0.188
s	0.041	4	0111	0.164
t	0.038	5	10000	0.190
d	0.036	5	10001	0.180
r	0.036	5	10010	0.180
h	0.034	5	10011	0.170
k	0.033	5	10100	0.165
p	0.027	5	10101	0.135
j	0.025	5	10110	0.125
l	0.025	5	10111	0.125
g	0.019	5	11000	0.095
e	0.019	5	11001	0.095
é	0.017	5	11010	0.085
b	0.017	6	110110	0.102
eu	0.015	6	110111	0.090
m	0.015	6	111000	0.090
,	0.011	6	111001	0.066
y	0.011	6	111010	0.066
.	0.010	6	111011	0.060
o	0.009	6	111100	0.054

ISBN: 978-602-512-810-3

Simbol	**Probabilitas**	***Length Code***	***Codeword***	***Expected Code Length***
w	0.007	7	1111010	0.049
c	0.005	7	1111011	0.035
"	0.004	7	1111100	0.028
-	0.003	8	11111010	0.024
(	0.001	8	11111011	0.008
)	0.001	9	111111000	0.009
0	0.001	9	111111001	0.009
3	0.001	9	111111010	0.009
7	0.001	9	111111011	0.009
9	0.001	9	111111100	0.009
:	0.001	9	111111101	0.009
1	0.001	9	111111110	0.009
6	0.001	9	111111111	0.009
Total	1	-	-	4.242

d. Bahasa Jawa

Codeword (*C*) dan *length codeword* (*LC*) akan didapatkan dari *Huffman Tree* yang diperliahtkan pada gambar 4.



Gambar 4. *Binary Huffman Tree Code* Bahasa Jawa

Aturan main *Huffman Tree* untuk pemberian *Codeword* pada masing-masing simbol sama dengan *Binary Huffman Tree Code* bahasa sunda dengan mengacu pada algoritma *Huffman Code* pada gambar 2.

Expected code length (*L*) pada tabel 3 didapatkan dari jumlah perkalian antara probabilitas (*P*) tiap karakter dan *length codeword* (*LC*) nya yang telah diperoleh dari *Binary Huffman Tree Code* pada gambar 4.

Tabel 3. *Expected Code Length Binary Huffman Code* Bahasa Jawa

Simbol	**Probabilitas**	***Length Code***	***Codeword***	***Expected Code Length***
spasi	0.131	3	000	0.393
a	0.127	3	001	0.381
n	0.125	3	010	0.375
g	0.071	4	0110	0.284
e	0.069	4	0111	0.276
u	0.058	4	1000	0.232
i	0.055	4	1001	0.220
k	0.047	5	10100	0.235
l	0.038	5	10101	0.190
t	0.035	5	10110	0.175
s	0.034	5	10111	0.170
d	0.028	5	11000	0.140
p	0.027	5	11001	0.135
o	0.026	5	11010	0.130
r	0.020	6	110110	0.120
m	0.016	6	110111	0.096
h	0.013	6	111000	0.078
j	0.013	6	111001	0.078
b	0.013	6	111010	0.078
y	0.013	6	111011	0.078
w	0.012	6	111100	0.072
.	0.012	6	111101	0.072
,	0.011	6	111110	0.066
c	0.002	8	11111100	0.016
-	0.002	8	11111101	0.016
1	0.001	8	11111110	0.008
2	0.001	8	11111111	0.008
Total	1	-	-	4.122

e. Analisis

Dari hasil penelitian dapat dilihat bahwa *entropy* bahasa jawa lebih rendah dibandingkan dengan *entropy* bahasa sunda. Hal tersebut tejadi karena simbol yang muncul pada bahasa jawa lebih sedikit dibandingakan dengan simbol yang muncul pada bahasa sunda. Sehingga, prediktabilitas dari masing-masing simbol akan semakin besar pada saat proses komunikasi. *Entropy* yang lebih rendah akan menghasilkan kompresi yang lebih baik. Karena *entropy*

ISBN: 978-602-512-810-3

merupakan batas minimum sebuah pesan dapat dikompres atau dengan kata lain mengungkapkan sejauh mana pesan dapat dikompres.

4. Kesimpulan

Kompresi diperlukan untuk penyimpanan informasi yang efektif dan transmisi yang lancar dan efisien melalui saluran. *Entropy* mengungkapkan sejauh mana sebuah pesan dapat dikompres. Dari hasil penelitian yang telah dilakukan, *entropy* Bahasa Sunda dengan total simbol pada naskah sebanyak 1505 simbol sebesar 4.168 bits per simbol, sedangkan *entropy* Bahasa Jawa dengan total simbol pada naskah sebanyak 1554 simbol, sebesar 4.101 bits per simbol. Hasil penelitian ini diharapkan bermanfaat besar untuk teknik kompresi dimasa yang akan datang.

Ucapan Terimakasih

Kami mengucapakan terimakasih kepada Dr. Eng Khoirul Anwar ST., M. Eng selaku Direktur *Advenced Wirelles and Technology* (AdWiTech) yang banyak memberikan bimbingan pada penelitian ini.

Daftar Pustaka

[1] Y. E. Hutasoit. *Huffman Coding untuk Kompresi Data Teks Berbahasa Indonesia*. 2001.
[2] C. E. Shannon, *A Mathematical Theory of Communication*. Bell Syst. Techn. J., 1948.
[3] M. Kuruvila and D. P. Gopinath. *Entropy of Malayalam Language and Text Compression Using Huffman Coding*. *2014 1st Int. Conf. Comput. Syst. Commun. ICCSC 2014*; no. December, pp. 150–155, 2003.
[4] K. Sayood, *Introduction to Data Compression*. 2012.

ISBN: 978-602-512-810-3