Klasifikasi *Multi-Label* Pada Topik Berita Berbahasa Indonesia Menggunakan *Multinomial Naïve Bayes*

I Made Riartha Prawira[1], Adiwijaya[2], Mohamad Syahrul Mubarok[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]riartha@student.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id,
[3]msyahrulmubarok@telkomuniversity.ac.id

Abstrak

Berita merupakan informasi yang dialirkan dari berbagai sumber mengenai kejadian factual yang dapat mempengaruhi lingkungan sekitar. Klasifikasi label topik biasanya dilakukan dalam pengelompokan artikel berita berdasarkan topiknya. Variabel ciri artikel merupakan penentu dalam klasifikasi label. Namun apabila suatu ciri yang menjadi ciri dari satu label artikel merupakan ciri dari label artikel lainnya maka artikel tersebut memiliki lebih dari satu topik atau disebut topik *multi-label.* Penelitian ini melakukan pembangunan pemodelan suatu klasifikasi teks berita dengan menggunakan metode *multinomial naïve bayes* untuk melakukan klasifikasi *multi-label* dengan metode *hamming loss* sebagai pengukuran performa model klasifikasi tersebut. Hasil *hamming loss* yang dihasilkan dari penelitian ini sebesar 0,18. Berdasarkan hasil penelitian, metode *multinomial naïve bayes* ini mampu untuk menyelesaikan permasalahan klasifikasi teks pada kasus *multi-label*.

Kata kunci : Berita, Klasifikasi, *Multi-Label,* MNB, *Hamming Loss*

Abstract

News is a distributed information from any resources contains factual events that affect the environment. Label topic classification generally used for specifying label topic of news article and gather it according to the same topic. Article feature variable is a definite factor in label classification. However, when a feature from one article feature label become another article feature label then that article may contain more than one topic or generally called multi-label topic. This research is about build a text classification model with multinomial naïve bayes method for classifying a multi-label classification problem with hamming loss method as a performance classification model measurement. Result of hamming loss that we obtain from this research is 0,18. Based on result from this research, this multinomial naïve bayes method can solve text classification problem on multi-label case.

Keywords : News, Classification, Multi-Label, MNB, Hamming Loss

1. Pendahuluan

Latar Belakang

Berita pada umumnya merupakan sebuah informasi yang dialirkan dari banyak sumber, definisinya pun sangat banyak namun dapat ditarik beberapa kesimpulan mengenai makna arti berita. Sering sekali ditemukan judul artikel berita dengan satu topik saja namun di dalam artikel tersebut bisa saja mengandung satu atau lebih topik berita.

Dapat memungkinkan bahwa terjadi dimana suatu artikel berita mengandung ciri lebih dari satu topik.

Banyaknya *learning tasks* yang dilakukan dengan *machine learning* mendapatkan asumsi klasifikasi yang kurang tepat dikarenakan suatu ciri topik ternyata merupakan ciri topik lainnya dan menyebabkan ketidakpastian dalam menentukan suatu kelas [1]. Apakah isi dari artikel itu mengandung satu topik saja atau yang disebut sebagai *single-label*, atau lebih dari satu topik di dalam artikelnya atau yang disebut sebagai *multi-label*. Hal ini perlu diperhatikan untuk mendapatkan hasil pengklasifikasian yang lebih akurat sesuai dengan isi dari artikel tersebut.

Pada kasus *Multi-label* yang dihadapi pada tugas akhir ini, penulis mengadopsi algoritma *Multinomial Naïve Bayes* (MNB) untuk melakukan klasifikasi *multi-label* pada topik berita berbahasa Indonesia pada data secara keseluruhan [2]. Berbeda dengan algoritma lainnya seperti *Fuzzy C-Means* (FCM), *Gustafon-Kessel* (GK), *K-Means* dan *Relief Method* yang merupakan klasifikasi model geografik dengan menggunakan jarak untuk melakukan klasifikasi [3] [4], Naïve bayes klasifikasi teks dikenal sebagai metode perhitungan probabilitas dengan performansi yang baik dengan kalkulasi yang tidak rumit, berdasarkan penelitian [5] naïve bayes mampu menghasilkan *f*1-*measure* sebesar 88,13% untuk menyelesaikan permasalahan klasifikasi teks.

Topik dan Batasannya

Berdasarkan latar belakang yang sudah dijelaskan, maka rumusan masalah yang dihadapi yaitu :

1. bagaimana membangun model klasifikasi *Multinomial Naïve Bayes* untuk mengklasifikasi topik berita yang memiliki satu atau lebih kategori topik dalam satu artikel pada artikel berbahasa Indonesia.
2. bagaimana performa dari *Multinomial Naïve Bayes Classifier* dalam pengklasifikasian topik dari artikel berita berbahasa Indonesia dalam kasus *Multi-label*.

Sementara itu, adapun batasan masalah untuk penelitian pada kasus *Multi-label* ini yaitu dataset artikel berita hanya memiliki maksimal tiga label.

Tujuan

Adapun tujuan dari tugas akhir ini berdasarkan rumusan masalah yang ada yaitu :

1. Membangun model klasifikasi *Multinomial Naïve Bayes* untuk mengklasifikasi topik berita berbahasa Indonesia yang memiliki satu atau lebih kategori topik dalam satu artikel dengan algoritma *Multinomial Naïve Bayes Classifier.*
2. Menganalisis hasil pada performa algoritma *Multinomial Naïve Bayes Classifier* serta meneliti apa pengaruh yang membedakan hasil performa dengan metode lainnya.

2. Studi Terkait

2.1 Penelitian Sebelumnya

Terdapat beberapa penelitian sebelumnya yang melakukan penelitian mengenai *multi-label classification* berbagai macam metode. Pada penelitian yang dilakukan oleh Tao Zhang, Jiansheng Wu dan Haifeng Hu pada papernya yang berjudul *Text Classification Based on a Novel Ensemble Multi-Label Learning Method* membuat model sistem klasifikasi teks dengan menggunakan metode *Ensemble Multi-Label k-Nearest Neighbour* dan melakukan pengujian dengan beberapa metode *Attribute Measure* seperti *Information Gain* (IG), CHI-*Square* (CHI) dan *Document Thresholding Frequency* (DF). Berdasarkan uji coba En-MLKNN pada data *reuters*-21578 dengan pengukuran *Hamming Loss* mendapatkan hasil 0,070, dimana pengukuran evaluasi tersebut semakin rendah semakin bagus [6].

Mohammed A. Shehab, Omar Badarneh, Mahmoud Al-Ayyoub dan Yaser Jararweh pada papernya yang berjudul *A Supervised Approach for Multi-Label Classification of Arabic News Articles* melakukan penelitian mengenai klasifikasi teks *multi-label* pada dataset berupa data teks Bahasa arab menggunakan metode *Decision Tree* (DT), *Random Forest* (RF) dan *k-Nearest Neighbor* dengan k = 5 (5NN). Pengujian dilakukan dengan perbandingan tiga metode tersebut dengan hasil evaluasi menggunakan *Hamming Loss* dengan 5NN mendapatkan hasil terendah sebesar 0.2 dimana pengukuran evaluasi tersebut semakin rendah semakin bagus [7].

Reynaldi Ananda Pane, Mohamad Syahrul Mubarok, Nanang Saiful Huda dan Adiwijaya pada papernya yang berjudul *A Multi-label Classification on Topic of Quranic Verses in English Translation using Multinomial Naïve Bayes* melakukan penelitian mengenai klasifikasi teks *multi-label* pada datateks berupa teks Bahasa inggris. Pengujian dilakukan dengan perbandingan hasil *Hamming Loss* pada sistem yang menggunakan *prior probability* dari data training dan *uniform prior probability* mendapatkan hasil sebesar 0,1247 [8].

Al Mira Khonsa Izzaty, Mohamad Syahrul Mubarok, Nanang Saiful Huda dan Adiwijaya pada papernya yang berjudul *A Multi-label Classification on Topic of Quranic Verses in English Translation using Tree Augmented Naïve Bayes* melakukan klasifikasi *multi-label* pada dataset berupa teks Bahasa inggris dengan menggunakan algoritma *Tree Augmented Naïve Bayes*. Pengujian juga dilakukan dengan perbandingan hasil *Hamming Loss* yang dihasilkan dengan pengujian *Mutual Information* dengan *threshold* yang berbeda dan struktur

TAN yang berbeda, mendapatkan hasil sebesar 0,1121 [9].

2.2 *Multinomial Naïve Bayes* (MNB)

Naïve Bayes merupakan metode pengambilan keputusan dengan perhitungan probabilitas. Naïve bayes sendiri untuk klasifikasi teks dikenal sebagai metode dengan performansi yang baik dengan kalkulasi yang tidak rumit [5].

$$\boldsymbol{P(c_k|X) \propto P(c_k) \cdot P(X|c_k)} \quad \textbf{(1)}$$

Pada persamaan (1) $\boldsymbol{P(c_k|X)}$ *posterior probability* untuk dokumen $\boldsymbol{X}$ terhadap kelas $\boldsymbol{c_k}$, $\boldsymbol{P(c_k)}$ merupakan *prior probability* kemunculan kelas $\mathbf{C_k}$ pada seluruh **C** tiap dokumen $\boldsymbol{X}$**.** dan $\boldsymbol{P(X|c_k)}$ *likelihood probability* dari kemunculan term fitur $\boldsymbol{X}$ pada kelas $\mathbf{C_k}$.

Multinomial Naïve Bayes (MNB) merupakan serapan dari teorema *Bayes* yang denominatornya dianggap konstan. Metode *Naïve Bayes* ini memiliki ciri utama yaitu asumsi yang kuat (*naïve)* terhadap ketidak tergantungan antar variabel [10]. MNB menerapkan fungsi sebagai berikut :

$$\boldsymbol{\hat{P}(c|d) \propto \arg max_{c \in C} \hat{P}(c) \prod_{1<k<n_d} \hat{P}(t_k|c)} \quad \textbf{(2)}$$

Pada persamaan (2) $\boldsymbol{\hat{P}(c|d)}$ adalah *posterior probability* untuk dokumen d terhadap kelas c, $\boldsymbol{\hat{P}(c)}$ adalah *prior probability* kemunculan dokumen d terhadap kelas c, $\boldsymbol{\hat{P}(t_k|c)}$ adalah *likelihood probability* kemunculan *term* $\boldsymbol{t_k}$ pada dokumen kelas c, dan $\boldsymbol{n_d}$ merupakan jumlah dokumen yang tersedia.

Prior probability digunakan untuk mencari peluang suatu dokumen tertentu dari seluruh jumlah dokumen yang tersedia. Perhitungan *prior probability* yaitu sebagai berikut :

$$\boldsymbol{\hat{P}(c) = \frac{N_c}{N}} \quad \textbf{(3)}$$

Pada persamaan (3) $\boldsymbol{N_c}$ merupakan jumlah dokumen pada kelas c dan **N** adalah total seluruh jumlah dokumen yang ada.

Permasalahan yang muncul adalah apabila suatu kata t muncul di dokumen satu tapi tidak muncul di dokumen lainnya, sedangkan perhitungan $\boldsymbol{\hat{P}(t_k|c)}$ akan dilakukan merata, maka akan terdapat peluang pada suatu kata t yaitu *zero probability*. Untuk menghindari masalah tersebut digunakan *add-one* atau *laplace smoothing* dengan perhitungan sebagai berikut :

$$\boldsymbol{\hat{P}(t_k|c) = \frac{T_{ct} + 1}{\sum_{t' \in V}(T_{ct'} + 1)} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B'}} \quad \textbf{(4)}$$

Dimana $\mathbf{T_{ct}}$ adalah jumlah kemunculan kata t dalam dokumen *train* pada kelas c yang diobservasi, $\boldsymbol{\sum_{t' \in V}(T_{ct'})}$ merupakan total kata keseluruhan yang terdapat yang terdapat pada suatu kelas c dan B' merupakan jumlah kata unik pilihan pada semua kelas yang menjadi dokumen *train*.

langkah terakhir yaitu untuk mencari *posteriori probability*. Namun untuk menentukan kelas apa yang cocok untuk dokumen inputan tersebut maka dilakukan *Maximum A Posteriori* (MAP) yang merupakan nilai peluang tertinggi pada tiap kelas dengan perhitungan sebagai berikut :

$$\boldsymbol{C_{MAP} = arg\ max_{c \in C} \hat{P}(c|d) = \arg max_{c \in C} \hat{P}(c) \prod_{1<k<n_d} \hat{P}(t_k|c)} \quad \textbf{(5)}$$

Hasil dari *posteriori probability* akan mengalami *floating point underflow* dimana angka yang dihasilkan terlalu kecil dikarenakan *denominator* nya dihilangkan dan untuk menghindari *floating point underflow* dapat dilakukan maksimalisasi sebagai berikut :

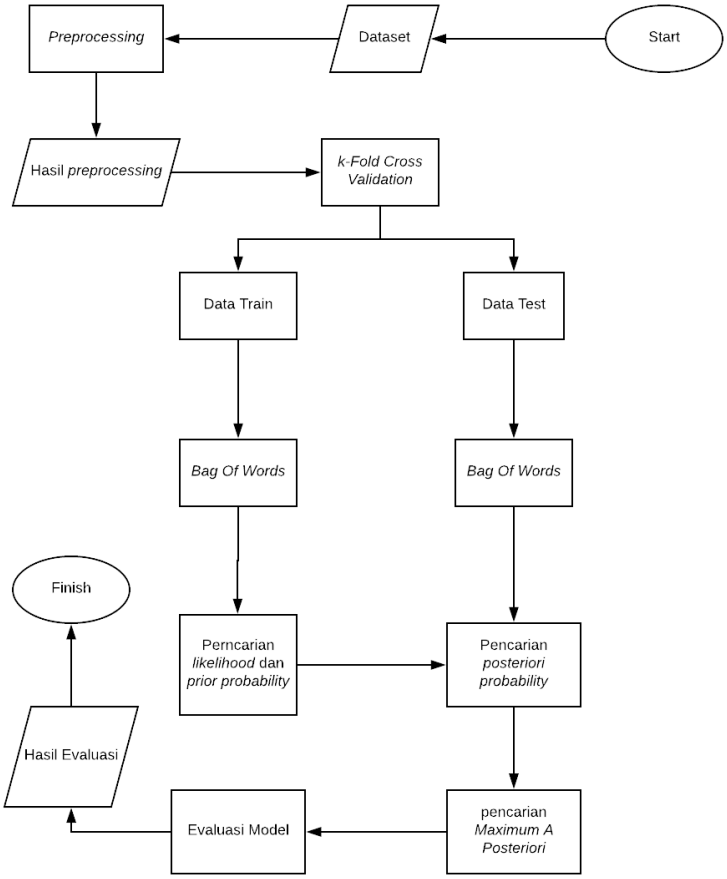
$$\boldsymbol{C_{map} = arg\ max_{c \in C} [log \hat{P}(c) + \sum_{1<k<n_d} log \hat{P}(t_k|c)}} \quad \textbf{(6)}$$

Apabila nilai masing-masing kelas sudah tidak terlalu kecil maka mesin dapat mengambil nilai *posterior* tertinggi.

3. Sistem yang Dibangun

3.1 Rancangan Sistem

Alur sistem yang akan dibangun dapat dilihat melalui *flowchart* berikut :



Gambar 1. Flow Chart Sistem

Sistem yang dibangun pada penelitian ini adalah suatu model sistem yang dapat mengklasifikasi *multi-label* topik dari suatu artikel berita berbahasa Indonesia dengan menggunakan *Multinomial Naïve Bayes*. Dataset yang digunakan melewati tahap *preprocessing* setelah itu mulai menjalankan *training* dan *testing* pada dataset. Partisi data latih dan data uji menggunakan *k-fold cross validation,* dimana metode ini selesai jika sudah melakukan kalkulasi nilai evaluasi [11]. Sistem ini menggunakan *k-fold* dengan k = 10. Setelah partisi data latih dan data uji dilakukan ekstraksi ciri dengan menggunakan *bag of words,*

Pada proses data latih, data dilakukan pencarian *prior probability* dan *likelihood probability* sebagai acuan pembelajaran yang akan digunakan pada proses pencarian *posteriori probability* oleh data uji. Karena kasus yang dihadapi adalah kasus *multi-label*, pencarian *prior probability* dan *likelihood probability* dilakukan dengan masing-masing mencari probabilitas kelas yes dan probabilitas kelas no. Hasil dari *learning* lalu disimpan ke dalam *database*.

Pada proses data uji, data dilakukan dengan pencarian *posteriori probability* dari hasil pembelajaran pada data latih. Evaluasi sistem dilakukan dengan metode *Hamming Loss* setelah melakukan *running model*. Semakin kecil *Hamming Loss* maka performa dari model dapat dikatakan baik.

3.2 Penggunaan Dataset

Penelitian dilakukan dengan pengambilan dataset yang berasal dari artikel berita berbahasa Indonesia dengan prosedur pencariannya yaitu dengan menngunduh artikel berita berbahasa Indonesia dari *website* penyedia artikel berita berbahasa Indonesia. Referensi yang dipakai yaitu *jawapos.com*. Jumlah dataset yang digunakan yaitu sebanyak 177 artikel dan untuk topik-topik yang tersedia yaitu 13 label tersebar pada seluruh artikel. Untuk informasi label yang digunakan untuk dataset dapat dilihat pada tabel 1. Yang menjelaskan nama label dan id label.

Tabel 1. Label Dataset

Id Label	Nama Label
1.	Politk

2.	Hukum
3.	Ekonomi
4.	Sosial
5.	Budaya
6.	Teknologi
7.	Gaya Hidup
8.	Olahraga
9.	Entertainment
10.	Pendidikan
11.	Pertahanan dan Keamanan
12.	Kesehatan
13.	Lainnya (diluar 12 label)

3.3 Preprocessing

Proses *preprocessing* yang dilakukan pada penelitian ini ada empat tahap, antara lain *case folding, tokenization, stopword removal* dan *stemming*. Berikut adalah penjelasan dari tiap tahapnya.

a. Case Folding

Pada *case folding* inputan yang masuk berupa teks artikel secara utuh dan outputnya adalah teks dengan rata huruf kecil tanpa tanda baca. Hal ini dilakukan untuk menghindari ketidaksamaan *value* yang didapat apabila menggunakan huruf kapital dan huruf kecil pada kata yang sama.

b. Tokenization

Pada tokenisasi input yang masuk berupa teks hasil *case folding* dan output nya adalah teks berupa kalimat yang dipecah menjadi perkata. Hal ini dilakukan untuk memudahkan sistem melakukan pengecekan kemunculan suatu kata pada tiap artikel.

c. Stopword Removal

Pada *stopword removal* input yang masuk berupa teks hasil tokenisasi dan output nya adalah teks tanpa penulisan kata depan dan kata-kata sambung yang dianggap tidak berpengaruh. Hal ini dilakukan untuk mengefisienkan waktu eksekusi program dengan cara membuang *noise*

d. Stemming

Pada *stemming* input yang masuk berupa teks hasil *stopword removal* dan output nya adalah teks dengan kata dasar dan tanpa menggunakan imbuhan. Hal ini dilakukan untuk menghasilkan kata-kata tanpa imbuhan yang siap diklasifikasi.

3.4 Ekstraksi Ciri

Pada tahap ekstraksi ciri, tugas yang dilakukan merupakan klasifikasi teks dimana akan merubah kalimat-kalimat menjadi kata-kata yang dapat dimengerti komputer, atau dilakukan pemrosesan bahasa alami [12]. Representasi model *bag of words* ini dilakukan dengan menghitung jumlah kemunculan suatu kata pada suatu artikel yang merupakan bagian dari data *train* untuk masing-masing label yang sudah disediakan. Setiap kata unik dan frasa dapat menjadi fitur pada suatu dokumen, dimana menghasilkan jumlah fitur yang sangat banyak [13].

3.5 Evaluasi

Evaluasi sistem dilakukan dengan *hamming loss* dimana dilakukan pencarian *hamming loss* setelah melakukan *running* model. Semakin kecil hasil *hamming loss* yang didapat maka performa dari model dapat dikatakan baik.

Tabel 2 menjelaskan representasi data dari *hamming loss* pada banyak data sebanyak 6 data.

Tabel 2. Representasi Data Hamming Loss

	Y	$\hat{Y}$
X_1	[1 0 0 1 0 0 0 0 0 0 0 0 0]	[1 1 0 0 0 0 0 0 0 0 0 0 0]
X_2	[1 1 0 1 0 0 0 0 0 0 0 0 0]	[0 1 1 1 0 0 0 0 0 0 0 0 0]
X_3	[0 0 1 1 0 0 0 0 0 0 0 0 0]	[0 0 1 1 0 0 0 0 0 0 0 0 0]
X_4	[0 1 1 1 0 0 0 0 0 0 0 0 0]	[1 1 1 1 0 0 0 0 0 0 0 0 0]
X_5	[1 0 0 0 0 0 0 0 0 0 0 0 0]	[1 0 0 0 0 0 0 0 0 0 0 0 0]

X_6	[0 1 1 1 0 0 0 0 0 0 0 0 0]	[1 0 0 1 0 0 0 0 0 0 0 0 0]

Dengan perhitungan *hamming loss* pada persamaan (7)

$$HL = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{N}[\hat{Y}_j \neq Y] \quad \textbf{(7)}$$

Dimana X adalah data, Y adalah representasi bit, N jumlah seluruh data dan L adalah banyaknya bit representasi data.

4. Evaluasi

4.1 Skenario Pengujian

Dalam menentukan parameter terbaik dari metode yang digunakan, penulis melakukan empat skenario pengujian terhadap model klasifikasi yang dibangun, yaitu sebagai berikut :

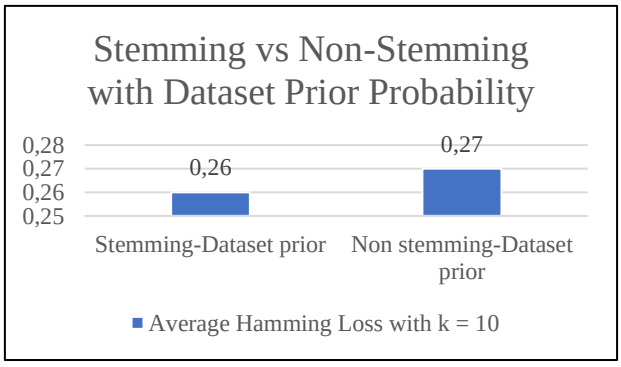
1. **Skenario 1 : Analisis penggunaan *Stemming* vs *non-Stemming* pada bagian *Preprocessing* dan menggunakan dataset *prior probability* pada bagian proses klasifikasi** pada skenario ini, data latih dan data uji akan menjalankan *preprocessing* dengan menggunakan *Stemming* dan tidak menggunakan *Stemming,* dan menggunakan nilai *prior probability* sesuai dengan dataset yang didapat.
2. **Skenario 2 : Analisis penggunaan *Stemming* vs non-*Stemming pada* bagian *Preprocessing* dan menggunakan uniform *prior probability* pada bagian proses klasifikasi** pada skenario ini, data latih dan data uji akan menjalankan *preprocessing* dengan menggunakan *Stemming* dan tidak menggunakan *Stemming,* dan menggunakan nilai uniform *prior probability* yaitu penyetaraan nilai seluruh prior probability (dengan nilai sebesar 0,5).
3. **Skenario 3 : Analisis penggunaan dataset *prior probability* vs *uniform prior probability* pada bagian proses klasifikasi terhadap model MNB** pada skenario ini, data latih dan data uji akan menjalankan *preprocessing* dengan menggunakan *Stemming*. Analisis dilakukan dengan menganalisa performa yang dihasilkan dari penggunaan dataset *prior probability* dan *uniform prior probability* pada bagian proses klasifikasi.

4.2 Analisis dan Hasil Pengujian

Berikut adalah Analisa dan hasil yang didapat dari empat skenario pengujian yang dilakukan :

4.2.1 Skenario 1 : Analisis penggunaan *Stemming* vs *non-Stemming* pada bagian *Preprocessing* dan menggunakan dataset *prior probability* pada bagian proses klasifikasi

Sesuai dengan skenario yang sudah dijelaskan pada poin 4.1 pada skenario, dimana akan menggunakan *k-fold cross validation* dengan k sebanyak 10 pada pengujian *stemming* dengan *non-stemming* pada dataset *prior probability*, berikut adalah hasil pada skenario 1.

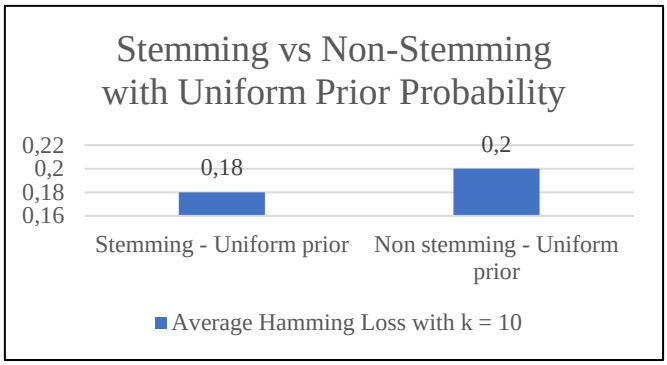


Gambar 2. Grafik *Average Hamming Loss Stemming vs Non Stemming* Dengan *Dataset Prior Probability* MNB

Berdasarkan grafik pada Gambar 2. Hasil rata-rata *hamming loss* dari penggunaan *stemming* menghasilkan sebesar 0,26. Hasil ini lebih kecil dari hasil rata-rata *hamming loss* dari tanpa penggunaan *stemming* yang menghasilkan sebesar 0,27. Maka hasil rata-rata *hamming loss* dari penggunaan *stemming* lebih baik dikarenakan hasilnya lebih kecil.

4.2.2 Skenario 2 : Analisis penggunaan *Stemming* vs *non-Stemming* pada bagian *Preprocessing* dan menggunakan uniform *prior probability* pada bagian proses klasifikasi

Pada skenario kedua, menggunakan *k-fold cross validation* dengan k sebanyak 10 pada pengujian *stemming* dengan *non-stemming* pada *uniform prior probability*. Berikut adalah hasil yang didapat pada skenario 2.

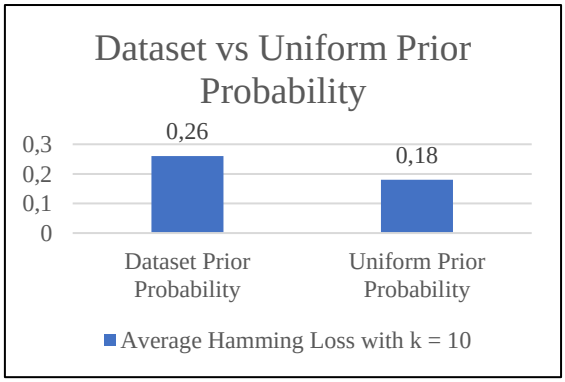


Gambar 3. Grafik *Average Hamming Loss Stemming vs Non-Stemming* Dengan *Uniform Prior Probability* MNB

Berdasarkan grafik pada Gambar 3. Hasil rata-rata *hamming loss* dari penggunaan *stemming* menghasilkan sebesar 0,18. Hasil ini lebih kecil dari hasil rata-rata *hamming loss* dari tanpa penggunaan *stemming* yang menghasilkan sebesar 0,2. Maka hasil rata-rata *hamming loss* dari penggunaan *stemming* lebih baik dikarenakan hasilnya lebih kecil.

4.2.3 Skenario 3 : Analisis penggunaan dataset *prior probability* vs *uniform prior probability* pada bagian proses klasifikasi terhadap model MNB

Pada skenario ketiga, menggunakan *k-fold cross validation* dengan k sebanyak 10 pada pengujian pemakaian dataset *prior probability* dan *uniform prior probability*. Berikut adalah hasil yang didapat pada skenario 3.



Gambar 4. Grafik *Average Hamming Loss Dataset Prior Probability* vs *Uniform Prior Probability* Pada Model MNB

Berdasarkan grafik pada Gambar 4. Hasil rata-rata *hamming loss* dari penggunaan dataset *prior probability* menghasilkan sebesar 0,26. Hasil ini lebih besar dari hasil rata-rata *hamming loss* dari penggunaan *uniform prior probability* yang menghasilkan sebesar 0,18. Maka hasil rata-rata *hamming loss* dari penggunaan *uniform prior probability* lebih baik dikarenakan hasilnya lebih kecil.

5. Kesimpulan

5.1 Kesimpulan

Dari hasil analisis seluruh pengujian yang dilakukan dalam penelitian tugas akhir ini, dapat ditarik kesimpulan yaitu Model klasifikasi *multi-label* topik berita berbahasa Indonesia dapat dibangun dengan menggunakan metode *multinomial naïve bayes* dan menghasilkan *hamming loss* paling rendah yaitu 0.18.

Selain itu, model MNB dengan menggunakan metode *stemming* pada bagian *preprocessing* mendapatkan hasil yang lebih baik daripada tanpa menggunakan metode *stemming* pada bagian preprocessing, dapat dilihat dari hasil *hamming loss* yang didapatkan pada penggunaan dataset *prior probability* dan *uniform prior probability*

yaitu sebesar 0.26 dan 0.18.

Model MNB dengan menggunakan *uniform prior probability* juga mendapatkan hasil yang lebih baik daripada menggunakan dataset *prior probability*, dapat dilihat dari hasil *hamming loss* yang didapatkan pada penggunaan *uniform prior probability* yaitu sebesar 0.18.

5.2 Saran

Dalam penelitian tugas akhir ini, terdapat saran yang dapat disampaikan untuk pengembangan penelitian berikutnya, yaitu jumlah dataset yang digunakan untuk melakukan *learning* lebih banyak sehingga proses *learning* pada label data dapat lebih merata. Saran lainnya yaitu umlah antar label pada seluruh dokumen dibuat dengan jumlah yang sama.

Daftar Pustaka

[1] M.-L. Zhang and Z.-H. Zhou, "A Review on Multi-Label Learning Algorithms," *IEEE,* vol. 26, no. 6, pp. 1819-1837, 2007.

[2] J. Liangxiao, W. Shasha, Z. Lungan and L. Chaoqun, "Structure Extended Multinomial Naive Bayes," *Information Sciences,* vol. 329, pp. 346-356, 2016.

[3] A. F. B. Firmansyah and S. Pramana, "Ensemble Based Gustafson Kessel Fuzzy Clustering," *Journal of Data Science and Its Applications (JDSA),* vol. 1(1), pp. 1-9, 2018.

[4] H. Aydenta and A. , "A clustering approach for feature selection in microarray data classification using random forest," *Journal of Information Processing System,* vol. 14(5), 2018.

[5] M. S. Mubarok, Adiwijaya and M. D. Aldhi, "Aspect-based sentiment analysis to review products using Naïve Bayes," *In AIP Conference Proceedings,* vol. 1867, p. 020060, 2017.

[6] T. Zhang, J. Wu and H. Hu, "Text Classification Based on a Novel Ensemble Multi-Label Learning Method," *2nd International Conference on Systems and Informatics,* pp. 964-968, 2014.

[7] M. A. Shehab, O. Badarneh, M. Al-Ayyoub and Y. Jararweh, "A Supervised Approach for Multi-Label Classification of Arabic News Articles," *7th International Conference on Computer Science and Information Technology (CSIT),* 2016.

[8] R. A. Pane, M. S. Mubarok, N. S. Huda and A. , " A Multi-lable Classification on Topics of Quranic Verses in English Translation using Multinomial Naive Bayes," *In 2018 6th International Conference on Information and Communication Technology (ICoICT). IEEE,* 2018.

[9] A. M. K. Izzaty, M. S. Mubarok, N. S. Huda and A. , "A Multi-label Classification on Topics of Quranic Verses in English Translation Using Tree Augmented Naïve Bayes," in *In 2018 6th International Conference on Information and Communication Technology (ICoICT). IEEE.*, 2018.

[10] R. A. Azis, M. S. Mubarok and A. Adiwijaya, "Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naive Bayes," in *In Indonesia Symposium on Computing (IndoSC)*, 2016.

[11] Adiwijaya, M. N. Aulia, M. S. Mubarok, U. N. W. and F. Nhita, "A Comparative Study of MFCC-KNN and LPC-KNN for Hijaiyyah Letters Pronounciation Classification System," *In Information and Communication Technology (ICoIC7), 2017 5th International Conference, IEEE,* pp. 1-5, 2017.

[12] A. F. Said, E. Jasin and A. Kusumaningrum, "Classification of hadith into positive suggestion, negative suggestion, and information," *In Journal of Physics: Conference Series,* vol. 971, p. 012046, 2018.

[13] A. I. Pratiwi and Adiwijaya, "Information Gain Based Feature Selection and Classification for Sentiment Analysis," in *Applied Computational Intelligence and Soft Computing*, 2018.