KLASIFIKASI MULTI-LABEL PADA TOPIK AYAT QURAN TERJEMAHAN BAHASA INGGRIS MENGGUNAKAN TREE AUGMENTED NAÏVE BAYES (TAN)

A MULTI-LABEL CLASSIFICATION ON TOPICS OF QURANIC VERSES IN ENGLISH TRANSLATION USING TREE AUGMENTED NAÏVE BAYES (TAN)

Al Mira Khonsa Izzaty[1], Mohamad Syahrul Mubarok[2], Adiwijaya[3]

[1,2,3]Prodi S1 Informatika, Fakultas Informatika, Universitas Telkom Bandung

[1]**almirakhonsa@gmail.com,** [2]**msyahrulmubarok@gmail.com**, [3]**adiwijaya@telkomuniversity.ac.id**

Abstrak

Al-Qur'an merupakan salah satu mukjizat yang diturunkan untuk dijadikan pedoman hidup bagi umat Muslim. Setiap umat Muslim wajib memahami serta mengamalkan ajaran yang dianjurkan Al-Qur'an. Ayat Al-Qur'an memiliki bahasan atatu topik yang dikaji, satu ayat dapat membahas satu topik atau lebih, pada kasus ini ayat Al-Qur'an termasuk dalam multi-label. Untuk memudahkan umat Muslim dalam memahami Al-Qur'an perlu dibangun sistem klasifikasi ayat Al-Qur'an. Berdasarkan penelitian sebelumnya, teorema Bayes dianggap *common* dalam menyelesaikan kasus klasifikasi, sehingga pada penelitian ini digunakan pendekatan probabilistik untuk membangun klasifikasi multi-label dengan Tree Augmented Naïve Bayes (TAN). Dalam pembangunan struktur TAN digunakan seleksi ciri dengan *Mutual Information* yang menghitung kebergantungan antar variabel input. Pada akhir pengujian nilai performa dari sistem dihitung dengan menggunakan *hamming loss* yang menghitung nilai error pada hasil klasifikasi multi-label. *Hasil* pengujian terbaik diperoleh ketika menggunakan threshold MI 3 yaitu dengan nilai *hamming loss* sebesar 0.1121, sedangkan nilai *hamming loss* terendah diperoleh ketika pembangunan struktur tidak menggunakan MI yaitu dengan nilai *hamming loss* 0.1208.

Kata kunci: Ayat Al-Qur'an, Klasifikasi Multi-label, Tree Augmented Naïve Bayes, Mutual Information

Abstract

The Qur'an is one of the revealed miracles to be used as a living guide for Muslims. Every Muslim must understand and practice the teachings suggested by the Qur'an. The verses of the Qur'an have a subject or a topic that is examined, one verse can discuss one or more topics, in this case the verses of the Qur'an included in multi-label. To facilitate Muslims in understanding the Qur'an, it is necessary to build a system of classification of verses of the Qur'an. Based on previous research, Bayes's theorem is considered common in solving classification cases, so in this study a probabilistic approach is used to construct a multi-label classification with Tree Augmented Naïve Bayes (TAN). In the development of TAN structure used characteristic selection with Mutual Information which calculates the dependence between input variables. At the end of the test the performance value of the system is calculated by using hamming loss which calculates the error value on the multi-label classification result. The best test results obtained when using the threshold of MI 3 is the value of hamming loss of 0.1121, while the lowest hamming loss value obtained when the construction of the structure does not use the MI is the value of hamming loss 0.1208

Keyword: Al-Qur'an Verses, Multi-label Classification, Tree Augmented Naïve Bayes, Mutual Information

1. Pendahuluan

1.1 Latar Belakang

Al-Qur'an adalah bentuk keajaiban yang abadi karena mencirikan kesempurnaan linguistik, benar, dan memvalidasi penemuan ilmiah terkini [1]. Umat Islam wajib mengimani, menjalankan perintah, serta menjauhi larangan Allah dalam Al-Qur'an. Al-Qur'an memiliki 114 surah, dan 6236 ayat [2]. Setiap ayat Al-Qur'an memiliki bahasan topik masing-masing, untuk satu ayat Al-Qr'an dapat membahas satu topik atau bahkan lebih dari satu topik. Sehingga pada kasus ini ayat Al-Qur'an termasuk dalam kasus multi-label. Untuk memudahkan umat Muslim dalam mempelajari topik ayat Al-Qur'an perlu dilakukan klasifikasi Ayat-Qur'an.

Berdasarkan penelitian yang dilakukan oleh Al-Kabi, dkk.[1] dan Mubarok dkk.[3, 4, 5, 6, 7] diketahui bahwa model klasifikasi menggunakan teorema Bayes memiliki performa cukup tinggi untuk klasifikasi teks. Hal ini terbukti dari nilai-nilai performa yang selalu diatas 80%. Salah satu varian dari teorema Bayes yaitu Tree Augmented Naïve Bayes (TAN). TAN memodelkan kebergantungan langsung yang ada pada variabel-variabel input (kata). Hal ini sesuai dengan sifat alamiah kata dalam suatu teks dimana ada kata-kata tertentu yang kemunculannya mempengaruhi atau dipengaruhi oleh kata-kata lain. Pada penelitian ini TAN digunakan untuk membangun classifier yang lebih presisi untuk kasus multi-label terhadap ayat Al-Qur'an.

Dalam membangun struktur TAN dibutuhkan seleksi ciri yang dapat menyeleksi variabel input (kata) sehingga dapat memperoleh struktur TAN yang lebih presisi. Mutual Information (MI) merupakan salah satu teknik seleksi ciri

yang mengukur kebergantungan antar variabel. MI memberikan informasi untuk variabel input lainnya. Pada penelitian ini MI digunakan utnuk memodelkan menyediakan informasi untuk variabel input lainnya. Pada penelitian ini MI digunakan utnuk memodelkan kebergantungan antar kata yang digunakan untuk membangun struktur TAN.

Penggunaan dataset dengan menggunakan bahasa Inggris dikarenakan bahasa Inggris merupakan bahasa Internasional. Sehingga diharapkan penelitian ini dapat berguna dan dapat dimanfaatkan oleh umat Muslim diseluruh dunia.

1.2 Rumusan Masalah

Rumusan masalah dari penelitian ini adalah bagaimana merancang sistem klasifikasi multi-label pada topik ayat Al-Qur'an terjemahan bahasa Inggris, bagaimana pengaruh penggunaan MI, serta bagaimana performa dari sistem klasifikasi multi-label yang telah dibangun.

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah membangun sistem klasifikasi multi-label dengan menggunakan TAN, menganalisis pengaruh penggunaan MI dalam pemabngunan struktur TAN, dan menganalisis performa dari TAN dalam mengklasifikasi ayat Al-Qur'an.

1.4 Manfaat Penelitian

Manfaat dari dilakukannya penelitian ini diharapkan hasil penelitian dapat membantu penelitian-penelitian selanjutny, serta hasil penelitian diharapkan dapat membantu umat Muslim di seluruh dunia dalam mempelajari Al-Qur'an.

2. Dasar Teori

2.1 Klasifikasi Ayat-Qur'an

Quranic classification merupakan salah satu pemanfaatan teks klasifikasi yang menggunakan ayat Al-Qur'an sebagai dataset. Beberapa penelitian terkait klasifikasi teks pernah dilakukan Mubarok, dkk. [8, 9, 10] dan Adiwijaya [11, 12, 13]. Penelitian terkait klasifikasi ayat Al-Qur'an pernah dilakukan sebelumnya oleh Mohammed N. Al-Kabi, dkk. [1]. Data yang digunakan merupakan Arabic text yaitu Modern Standard Arabic (MSA). Tujuan penelitian ini adalah mengklasifikasi ayat Al-Qur'an kedalam empat belas kelas yang telah ditentukan, serta menemukan metode klasifikasi yang terbaik. Metode yang diteliti antara lain Decision Tree, K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), dan Naïve Bayes (NB). Ada tiga kali percobaan yang dilakukan, percobaan pertama menggunakan 15.000 ayat, percobaan kedua 1.445, dan percobaan ketiga dengan jumlah ayat 5.121. Dari hasil penelitian pertama dilakukan tiga perbandingan metode yaitu Naïve Bayes, KNN, dan Rocchio. Dari beberapa hasil percobaan yang dilakukan Naïve Bayes dapat mengklasifikasi ayat Al-Qur'an hasil yang lebih akurat dibanding metode lainnya [1].

2.2 *Tree Augmented Naïve Bayes*

Tree Augmented Naïve Bayes (TAN) merupakan salah satu bagian dari Probabilistic Graphical Model (PGM). PGM adalah kerangka yang digunakan untuk memodelkan sistem yang melibatkan feature dengan banyak ketidaktentuan. Pada model TAN, setiap variabel dependent terhadap kelas dan dengan variable lain dari *feature set* [14].

TAN merupakan *improve* dari *naïve bayes* dan dikembangkan sebagai salah satu pembangun struktur bayesian network [15]. TAN meningkatkan interaksi tiap variabel, hal ini dapat meningkatkan keakuratan prediksinya [15]. Penelitian terkait penggunaan TAN pernah dilakukan oleh Padmanaban [14]. Dilakukan klasifikasi Wanita Pima India yang menderita penyakit diabetes. Data yang digunakan merupakan data yang diperoleh dari KEEL repositori. Dataset terdiri dari data diskrit dan data kontinu. Pada penelitian ini dilakukan perbandingan performa antara TAN dengan Naïve Bayes. Untuk classifier TAN digunakan seleksi ciri mutual information untuk mereduksi dimensi data. Hasil penelitian menyimpulkan bahwa model TAN memberikan performa yang lebih baik untuk banyak kasus dibanding naïve bayes. Perhitungan yang digunakan TAN untuk mengklasifikasi adalah perhitungan posterior probability. Perhitungan posterior pada TAN dapat dilihat pada persamaan (2.1). Dibutuhkan nilai likelihood, prior probability terhadap class dan prior probability terhadap parent antar atribut. Berikut persamaan yang digunakan:

$$P(C|X_1,\dots,Xn) \propto P(C)\ \cdot P(X_{root}|C)\prod_i P(X_i|C,X_{parent}) \tag{2.1}$$

dimana P(C) adalah prior probability untuk suatu kelas, $P(X_{root}|C)$ adalah conditional probability dari variabel root (parent), dan $P(X_i|C,X_{parent})$ adalah conditional probability dari semua variabel

2.3 *Mutual Information*

Mutual Information (MI) adalah ukuran kebergantungan antara dua variabel. Mutual Information digunakan untuk menghitung seberapa banyak sebuah variabel menyediakan informasi untuk variabel lain [13]. MI merupakan teknik yang dapat dilakukan untuk seleksi fitur, penggunaan mutual information sangat berpengaruh terhadap proses reduksi dimensi. Salah satu keuntungan penggunaan MI sebagai seleksi fitur adalah kemampuannya untuk mendeteksi hubungan nonlinier antar variabel, hal ini memungkinkan pengambilan relevansi dan redundansi fitur secara bersamaan [16]. Perhitungan mutual information antara dua random variabel X dan Y [17] dapat dilihat pada persamaan (2.2).

$$I_p = \log\left(\frac{P(x,y)}{P(x)P(y)}\right) \tag{2.2}$$

dimana P(x,y) adalah joint probability function dari x dan y, P(x) adalah marginal probability dari x, P(y) adalah marginal probability dari y.

2.4 *Hamming Loss*

Hamming loss merupakan evaluasi yang dilakukan untuk multi-label learning [18]. Evaluasi ini dilakukan dengan menghitung error atau kesalahan prediksi dari hasil klasifikasi [18]. Semakin kecil nilai hamming loss maka model klasifikasi yang dibangun semakin bagus, begitu pula sebaliknya. Persamaan (2.3) merupakan persamaan yang digunakan untuk perhitungan hamming loss [18]:

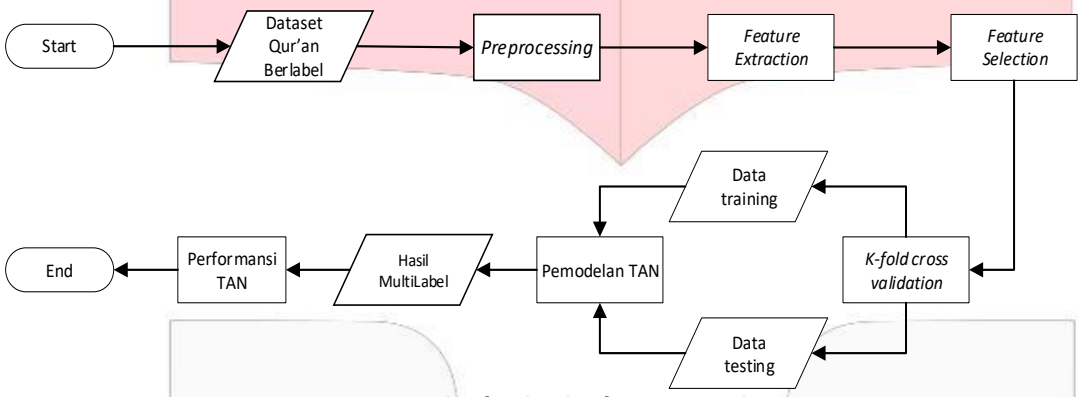
$$Hamming\ Loss = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L} \mathrm{I}\,(\hat{y}_j^{(i)} \neq y_j^{(i)}) \quad (2.3)$$

dimana N adalah Jumlah data, L adalah panjang output multi-label, $\hat{y}_j^{(i)}$ adalah target klasifikasi multi-label, dan $y_j^{(i)}$ adalah output klasifikasi multi-label.

3. Pengujian

3.1 Perancangan Sistem

Classification model yang dibangun pada Tugas Akhir ini bertujuan untuk mengklasifikasi ayat Al-Qur'an secara multi-label kedalam topik ayat telah ditetapkan. Gambaran umum sistem dapat dilihat pada Gambar (3.1).



Gambar 3.1 Gambaran Umum Sistem

Sebelum dataset digunakan pada tahap klasifikasi, perlu dilakukan *preprocessing* terhadap data. Tahap *preprocessing* merupakan kombinasi dari *case folding*, *tokenization*, *stopword removal*, dan *stemming*. Tahap ini dilakukan dengan tujuan menghilangkan *noisy* pada data.

Feature extraction dengan *bag of words* dilakukan setelah tahap *preprocessing*. Tahap ini menghitung frekuensi kemunculan masing-masing term untuk digunakan pada perhitungan *conditional probability*. Selanjutnya dilakukan *feature selection* dengan *mutual information* guna menemukan nilai ketergantungan antar variabel. Tingkat kebergantungan semakin tinggi jika sepasang variabel sering bertemu pada suatu dokumen secara bersamaan. Variabel yang saling bergantung penting untuk digunakan dalam pembangunan struktur TAN. Data yang telah di ektraksi dan di seleksi dibagi menjadi dua segmen yaitu *data training* dan *testing*. Pembagian data kedalam dua segmen menggunakan *k-fold cross validation.*

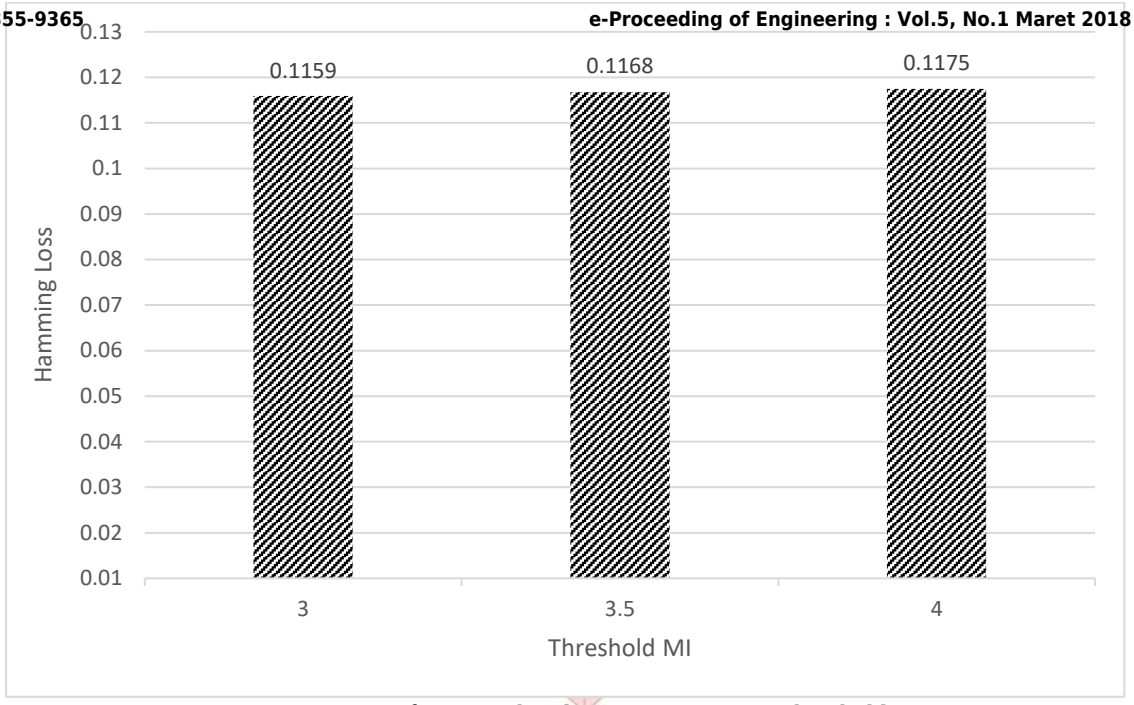
Tree Augmented Naïve Bayes dibangun untuk mengklasifikasi *dataset* Al-Qur'an yang telah melalui beberapa tahap. Klasifikasi TAN dilakukan dengan melakukan perhitungan *posterior*, *conditional probability variabel parent* terhadap kelas, dan *conditional probability variabel child* terhadap parent dan kelas. Dari model yang telah dibangun, ditentukan kelas untuk N data testing dengan menggunakan MAP. Setelah semua data memiliki kelasnya masing-masing selanjutnya dilakukan perhitungan performa menggunakan *Hamming Loss*.

3.3 Hasil Pengujian dan Analisis Sistem

1. Skenario 1 Penentuan variabel dependent pada pembangunan struktur TAN berdasarkan threshold MI

Pada skenario ini dilakukan pengujian untuk menentukan apakah nilai threshold yang digunakan pada pemilihan MI dapat mempengaruhi performa sistem. Ada tiga skema yang dilakukan pada skenario ini, skema 1 merupakan pembangunan struktur TAN dengan nilai threshold 3, skema 2 menggunakan nilai threshold 3.5, dan skema 3 menggunakan nilai threshold 4.

Nilai threshold berpengaruh pada pemilihan variabel yang digunakan untuk pembangun struktur TAN. Semakin rendah nilai threshold maka variabel yang digunakan semakin banyak, begitu pula sebaliknya semakin tinggi nilai threshold jumlah variabel yang digunakan semakin sedikit. Pemilihan nilai threshold dilakukan secara konstan dengan selisih sekitar 0.5 dari nilai maksimum MI.

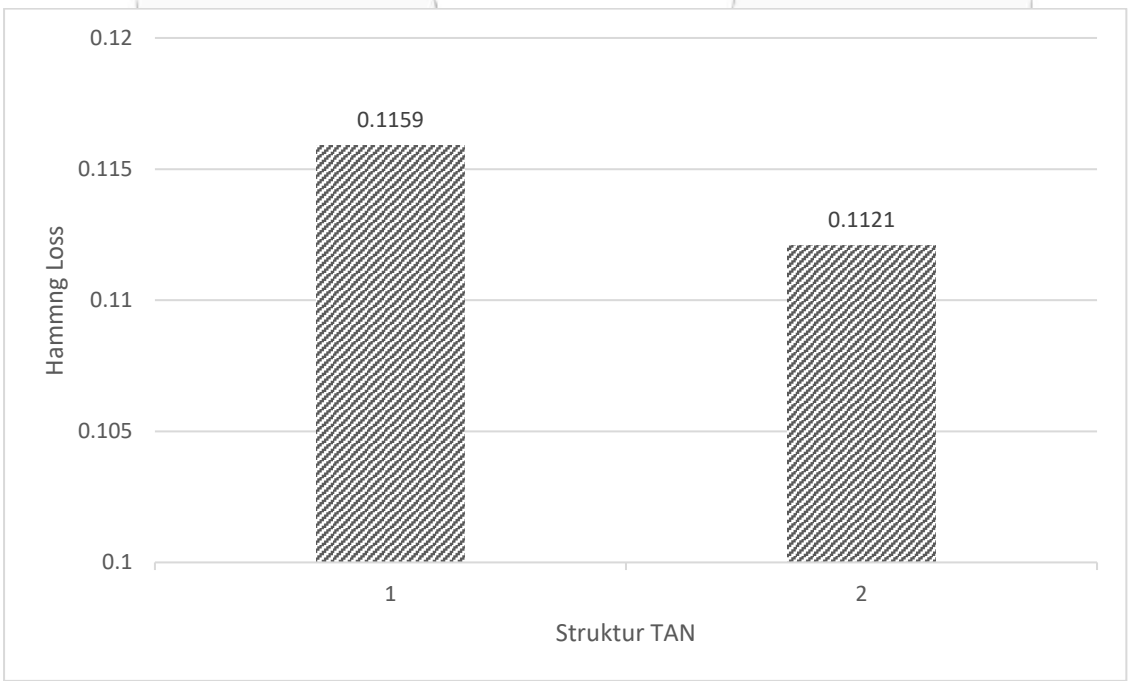


Gambar 3.2 Performa Perbandingan Penggunaan Threshold MI

Berdasarkan hasil pengujian Gambar 3.2 dihasilkan nilai *hamming loss* terendah dengan nilai 0.1159 yang menggunakan threshold MI 3 dan jumlah variabel terpilih adalah 300. Sedangkan penggunaan threshold dengan nilai 4 yang menggunakan variabel sebanyak 78 pasang menghasilkan *nilai hamming loss* tertinggi yaitu sebesar 0.1175. Semakin tinggi nilai threshold, maka nilai *hamming loss* cenderung meningkat. Pemilihan nilai threshold yang terlalu besar menyebabkan semakin sedikitnya variabel yang terpilih. Pemilihan variabel yang sedikit menyebabkan pembangunan struktur TAN hanya menggunakan sedikit fitur yang saling bergantung, hal ini dapat menyebabkan pembangunan struktur TAN yang kurang maksimal dandapat mempengaruhi performa sistem.

2. Skenario 2 Pertukaran variabel *parent* dengan *child* pada struktur TAN

Pada pengujian ini digunakan struktur TAN dengan performa terbaik hasil skenario 1 skema 1. Dilakukan pengujian terhadap pertukaran *variabel parent* dengan *variabel child* untuk mengetahui struktur yang lebih baik. Dilakukan pembalikan *edge* antar variabel sehingga semua variabel yang terpilih sebagai *child* dijadikan *parent* untuk setiap pasangannya. Gambar 4.2 merupakan perbandingan performa antara dua struktur, struktur 1 merupakan hasil dari skenario 1 dengan skema 1, sedangkan struktur 2 merupakan hasil dari skenario 1 dengan variabel yang telah dibalik edge antar variabelnya.

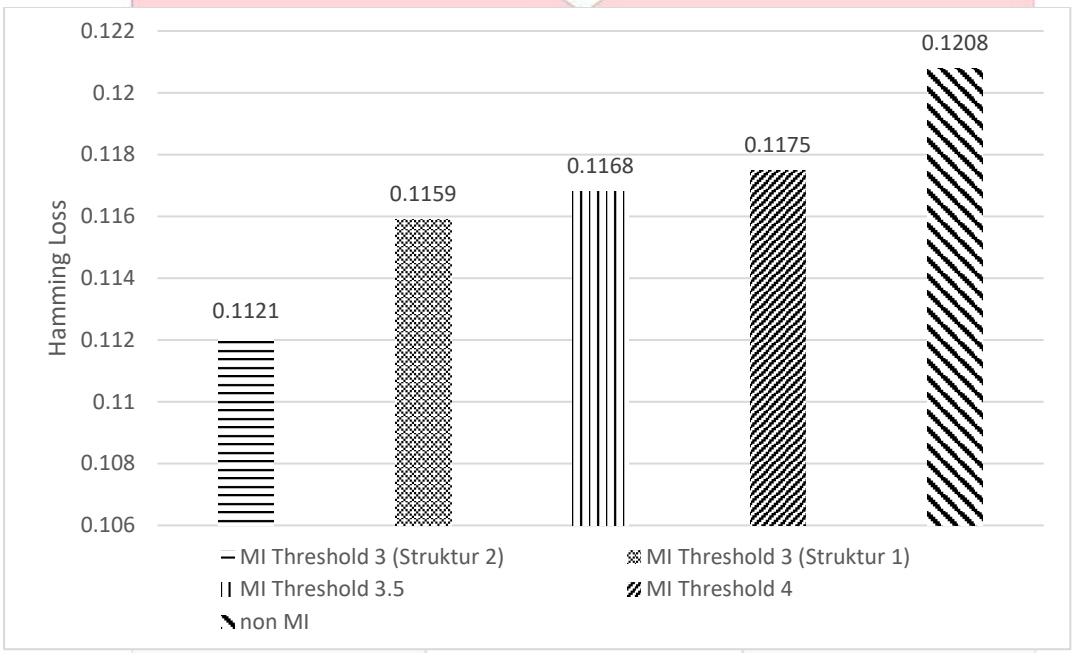


Gambar 3.3 Perbandingan Perubahan Struktur TAN 1 dengan TAN 2

Berdasarkan hasil pada Gambar 3.3 diketahui bahwa TAN yang tidak dibuat dengan [illegible] memberikan performa dengan nilai 0.1159, sedangkan ketika dibalik *edge*–nya memberikan performa sebesar 0.1121. Hal ini dikarenakan terjadinya perubahan pada nilai hasil perhitungan likelihood. Terjadinya perbedaan nilai perhitungan likelihood dapat menyebabkan perbedaan penentuan kelas oleh classifier.

3. Skenario 3 Penentuan struktur TAN terbaik dengan menggunakan *mutual information* dan tanpa menggunakan *mutual information*

Pada skenario ini dilakukan perbandingan antara beberapa struktur TAN yang dibangun dengan menggunakan MI dan struktur TAN yang dibangun tidak menggunakan MI. Dari pembangunan struktur TAN yang telah dilakukan dengan tanpa MI diperoleh hasil performa 5 struktur seperti pada Gambar 3.4. Hasil performa tertinggi sebesar 0.1159 diperoleh ketika menggunakan threshold 3 sedangkan hasil terendah adalah 0.1208 dengan struktur yang dibangun tanpa MI.



Gambar 3.4 Perbandingan Struktur dengan dan Tanpa Mutual Information

Berdasarkan hasil pengujian terlihat bahwa pengaruh penggunaan MI cukup signifikan. Struktur TAN yang dibangun dengan MI memiliki performa lebih baik dibandingan struktur TAN yang tidak menggunakan MI. Hal ini dikarenakan pembangunan struktur TAN menjadi kurang maksimal jika pemilihan variabel dilakukan dengan trial and error yaitu tanpa menggunakan MI. Sedangkan jika menggunakan MI kebergantungan langsung antar variabel dapat diukur dan digunakan untuk membangun struktur TAN, sehingga struktur TAN yang dihasilkan lebih merepresentasikan relasi antar variabel dan meningkatkan performa klasifikasi menjadi lebih baik.

4. Kesimpulan

Berdasarkan pengujian dan analisis yang telah dilakukan dapat disimpulkan bahwa penggunaan nilai *threshold* MI mempengaruhi penggunaan jumlah *variabel* yang digunakan untuk membangun struktur TAN. Semakin rendah nilai threshold maka jumlah variabel dependent yang digunakan semakin banyak. Semakin rendah nilai threshold MI dapat menurunkan nilai loss. Penggunaan MI juga terbukti dapat mempengaruhi performa dari *classification model* yang dibangun, hal ini disebabkan karena pemilihan kedekatan antar variabel dihitung dengan pasti, sedangkan pemilihan variabel tanpa MI melakukan pemilihan variabel dengan cara *trial and error*. Performa terbaik yang diperoleh MI pada penelitian ini adalah 0.1121 sedangkan performa tanpa MI adalah 0.1208.

Daftar Pustaka

[1] Al-Kabi, M. N., Ata, B. M. A., Wahsheh, H. A., & Alsmadi, I. M. (2013, December). A topical classification of Quranic Arabic text. In Proceedings of the 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences (pp. 252-257).

[2] M.H. Shakir. Al-Qur'an English Translation.

[3] Aziz, R. A., Mubarok, M. S., & Adiwijaya, A. (2016, September). Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naive Bayes. In Indonesia Symposium on Computing (IndoSC) 2016.

[4] Mubarok, M. S., Adiwijaya, & Aldhi, M. D. (2017, August). Aspect-based sentiment analysis to review products using Naïve Bayes. In AIP Conference Proceedings (Vol. 1867, No. 1, p. 020060). AIP Publishing.

[5] Syahputra, M. H., Bijaksana, M. A., & Mubarok, M. S. (2016). Kategorisasi Topik Tugas Akhir Berdasarkan Abstrak Dan Makassar Dengan Metode Multinomial Naïve Bayes Classifier. eProceedings of Engineering, 3(2).
[6] Saputra, A., Adiwijaya, A., & Mubarok, M. (2017). Klasifikasi Sentimen Pada Level Aspek Terhadap Ulasan Produk Berbahasa Inggris Menggunakan Bayesian Network (case Study: Data Ulasan Produk Amazon). eProceedings of Engineering, 4(3).
[7] Julianto, B., Adiwijaya, A., & Mubarok, M. (2017). Identifikasi Parafrasa Bahasa Indonesia Menggunakan Naive Bayes. eProceedings of Engineering, 4(3).
[8] Sitompul, D., Adiwijaya, A., & Mubarok, M. (2017). Analisis Sentimen Level Kalimat Pada Ulasan Produk Menggunakan Bayesian Networks. eProceedings of Engineering, 4(3).
[9] Putri, L., Mubarok, M., & Adiwijaya, A. (2017). Klasifikasi Sentimen Pada Ulasan Buku Berbahasa Inggris Menggunakan Information Gain Dan Naïve Bayes. eProceedings of Engineering, 4(3).
[10] Mubarok, M. S., & Asriadie, M. S. Klasifikasi Emosi Pada Twitter Menggunakan Bayesian Network.
[11] Pratiwi, A.I., & Adiwijaya, A. (2018). On The Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis, Applied Computational Intelligence and Soft Computing 2018. Hindawi, (1-9)
[12] Adiwijaya. (2014). Aplikasi Matriks dan Rung Vektor. Yogyakarta: Graha Ilmu.
[13] Adiwijaya. (2016). Matematika Diskrit dan Aplikasinya. Bandung: Alfabeta.
[14] Padmanaban, H. (2014). Comparative analysis of Naive Bayes and tree augmented naïve Bayes models.
[15] Awalia, P. (2011). Penerapan Algortima Tree Augmented Naive Bayesian pada Penentuan Peubah Penting
[16] Doquire, G., & Verleysen, M. (2013). Mutual information-based feature selection for multilabel classification. Neurocomputing, 122, 148-155.
[17] Ziveria, M. (2014). Perancangan Aplikasi Pencarian Kata Yang Berkaitan Secara Semantik Menggunakan Teori Mutual Information. Sesindo 2014, 2014.
[18] Read, J. (2015, September 20). Multi-label Classification. Helsinki, Finland: Aalto Univeristy.