Penggunaan Metode *K-Fold* untuk *Data Imbalance* pada Klasfikasi HWE dan QPQ dalam Kejahatan Tweet Pelecehan Seksual

Irfan Dwi Wijaya[1], Aji Gautama Putrada [2], Dita Oktaria[3]

[1,2,3] Universitas Telkom, Bandung
irfandwi@students.telkomuniversity.ac.id, ajigps@telkomuniversity.ac.id[2], ditaoktaria@telkomuniversity.ac.id[3]

Abstrak

***Sexual Harrasment* merupakan perilaku yang ditandai dengan pesan atau komentar yang melecehkan, ancaman atau hal-hal yang tidak senonoh, ajakan untuk melakukan aksi porno, atau seluruh perilaku melenceng yang dilakukan di media *online*. Permasalahan terhadap pelecehan seksual di media sosial menjadi hal penting yang harus dikaji. Analisis sentimen dapat digunakan sebagai solusi untuk mengindentifikasi media sosial mengenai pelecahan seksual. Tujuan tugas akhir ini yaitu untuk mengklasifikasikan data ujaran yang menjerumus pada pelecehan seksual berdasarkan kelas *quid pro quo* dan *hostile work environment* dari data *tweet #MeToo* yang telah dirangkum oleh *website* survey *theprofesorission.com* dengan menggunakan metode *Gaussian Naïve Bayes* dan KNN. Sistem dibangun melalui tahap *preprocessing*, *oversampling*. pembagian data latih dan data uji menggunakan *k-fold* dan evaluasi dengan membandingkan nilai akurasi. Dalam pengujian sistem, didapatkan nilai akurasi pada model klasifikasi *Naïve bayes* dengan menggunakan *K-Fold cross Validation* sebesar 90.7% dan model klasifikasi KNN dengan menggunakan *K-Fold cross validation* mencapai nilai 87.9%.**

Kata kunci : ***sexual harassment, gaussian naïve bayes*, analisis sentiment, *quid pro quo, hostile work environment.***

Abstract

Sexual Harrasment is behavior marked by harassing messages or comments, threats or obscene things, invitations to do pornographic actions, or all deviant behavior carried out in online media. The problem of sexual harassment on social media is an important thing that must be studied. Sentiment analysis can be used as a solution to identify social media regarding sexual harassment. The purpose of this final project is to classify speech data that leads to sexual harassment based on the quid pro quo class and hostile work environment from the #MeToo tweet data that has been summarized by the survey website theprofesorission.com using the Gaussian Naïve Bayes and KNN methods. The system is built through the preprocessing and oversampling stages. distribution of training data and test data using k-fold and evaluation by comparing the accuracy values. In system testing, the accuracy value of the Naïve Bayes classification model using K-Fold cross Validation is 90.7% and the KNN classification model using K-Fold cross validation reaches a value of 87.9%.

Keywords : ***sexual harassment, gaussian naïve bayes*, sentiment analysis, *quid pro quo, hostile work environment.***

1. Pendahuluan

Latar Belakang

Pelecehan seksual adalah tingkah laku seksual yang tidak diinginkan permintaan untuk melakukan tindakan seksual secara lisan atau fisik yang menjadikan seseorang menjadi tersinggung[21]. Beberapa contoh perilaku pelecehan seksual diantaranya seperti komentar tentang tubuh seseorang, komentar seksual, lelucon seksis, dan terus mengajak seseorang berkencan. Perilaku ini mungkin terjadi secara *online* ataupun secara personal dan langsung[14]. Hal ini merupakan suatu masalah karena dapat mempengaruhi kesejahteraan psikologis dan fisik korban. Korban pelecehan seksual mengalami malu, marah dan merasa terhina, sehingga korban yang sadar menjadi korban pelecehan seksual akan melaporkannya[1]. Kurangnya informasi, edukasi, dan pengetahuan tentang pelecehan seksual dan tentang jenis pelecehan seksual membuat korban bingung dan tertekan dalam menangani pelecehan seksual.

Tagar *#MeToo* pada *twitter* merupakan kampanye yang ditujukan untuk berbagi cerita tentang pelecehan seksual dan merupakan tindakan solidaritas antar korban yang disebarkan guna menumbuhkan kesadaran secara luas[15]. Dengan banyaknya orang yang membagikan pengalaman mereka tentang pelecehan seksual, penting untuk kita sebagai akademisi menggunakan pendekatan ilmiah untuk menunjukkan kontribusi yang berarti. Situs *web theprofessorisin.com* melakukan filtering terhadap data pelecehan seksual dari platform twitter dengan tagar #MeToo. Oleh karena itu, penulis memanfaatkan data tersebut untuk melakukan klasifikasi berdasarkan kelas *Hostile Work Environment* dan *Quid pro Quo* pada kasus yang dialami oleh korban. *Quid pro Quo* dan *Hostile Work Environment* merupakan pelabelan yang digunakan untuk menggambarkan dua jenis pelecehan seksual yang

relevan dengan pertanyaan fundamental pada persidangan [3]. Dari peraturan yang dirilis, *Title VII of the Civil Rights Act* of 1964 [20], pelecehan seksual dibagi menjadi dua, yaitu *Quid pro Quo* dan *Hostile Work Environment*. Oleh karena itu, penulis mencoba meneliti hal tersebut dengan membangun suatu model yang dapat mengklasifikasikan pesan pelecehan seksual dengan memanfaatkan metode sentimen analisis.

Penulis akan melakukan klasifikasi dengan menggunakan metode *Gaussian Naïve Bayes* dan K-NN untuk membandingkan nilai akurasi dengan memanfaatkan *K-fold cross validation* sebagai acuan implementasi. *Naïve Bayes Classifier* sering digunakan sebagai dasar dalam klasifikasi teks karena cepat dan mudah diimplementasikan serta memiliki tingkat akurasi yang tinggi[16] menurut Harrington[17] pada bukunya, K-NN merupakan salah satu algoritma paling populer dalam machine learning hal ini karena prosesnya mudah dan sederhana.

Topik dan Batasannya

Rumusan masalah dari penelitian ini adalah bagaimana pengklasifikasian *sexual harassment* yang terjadi menggunakan *K-fold Cross Validation* dengan metode *Gaussian Naïve Bayes* dan KNN sebagai pembanding nilai akurasi yang akan dihasilkan. Adapun batasan masalah pada penelitian ini adalah *dataset* yang digunakan merupakan data yang diambil pada *website theprofessorisin.com* dan berbahasa Inggris dengan total dataset 2105. Pelabelan dilakukan dalam dua kelas yaitu *Hostile Work Environment* dan *Quid pro Quo*.

Tujuan

Tujuan penelitian Tugas Akhir ini adalah untuk melakukan pengklasifikasian pengalaman pelecehan seksual yang terjadi menggunakan *K-fold Cross Validation* dengan model klasifikasi *Gaussian Naive Bayes* dan K-NN untuk mengetahui tingkat akurasinya.

Organisasi Tulisan

Pada bagian selanjutnya akan membahas perihal studi terkait pada penelitian sebelumnya yang berkaitan dengan penelitian ini, sistem yang akan dibangun, dan hasil pengujian juga analisa terhadap sistem yang telah dibangun.

2. Studi Terkait

Sexual Harrashment merupakan tindakan yang sering terjadi di masyarakat dan melukai korbannya. Seseorang yang pernah mengalami perlakuan seperti pelecehan seksual oleh orang-orang di sekitarnya akan memberikan trauma psikologis dan berdampak negatif pada pembentukan kepribadiannya. Pelecehan seksual dapat berupa konten seksual, membuat lelucon yang mengarah pada seksualitas dan penghinaan terhadap bagian tubuh seseorang, dan melakukan kontak fisik berupa sentuhan atau sejenisnya.[4].

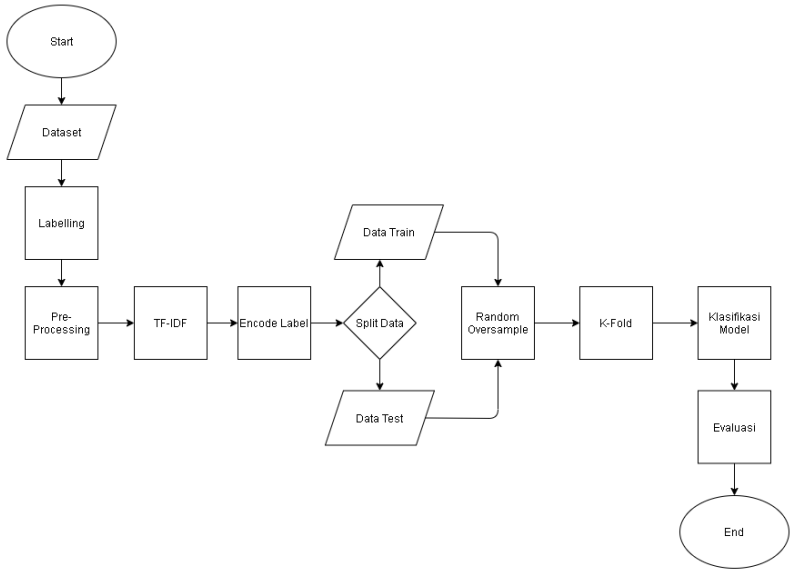
K-fold cross validation merupakan teknik validasi untuk menilai bagaimana hasil analisis statiskit mengeneralisasi kumpulan data. teknik ini digunakan sebagai prediksi model dimana mengakomodasi perkiraan arasi dari sebuah model ketika dijalankan[16]. Salah satu teknik dari validasi silang adalah *k-fold cross validation*, yang mana memecah data menjadi k bagian set data dengan ukuran yang sama. Penggunaan *k-fold cross validation* untuk menghilangkan bias pada data[16].

Pada penelitian sebelumnya, [8] tentang perbandingan kinerja metode *Naïve Bayes* dan K-NN untuk klasifikasi artikel Bahasa indonesia Hasil yang didapatkan menunjukkan metode *Naive Bayes* memiliki kinerja yang lebih baik dengan tingkat akurasi 70%, sedangkan metode K-NN memiliki tingkat akurasi yang cukup rendah yaitu 40%. Lalu pada penelitian yang dilakukan oleh Widaningsih[2] yaitu perbandingan metode data mining untuk prediksi nilai dan waktu kelulusan, Naïve bayes memiliki nilai akurasi tertinggi yaitu sebesar 76.79% dan untuk KNN sebesar 68.05%. Dalam penelitian ini penulis berfokus pada analisis algoritma klasifikasi yang akan digunakan adalah *Gaussian Naïve Bayes* yang bertujuan untuk mengklasifikasikan *tweets* pelecehan seksual dengan tagar *#MeToo* berdasarkan jenisnya yang ditujukan dengan besarnya nilai akurasi, presisi, recall, dan f1-score.

3. Sistem yang Dibangun

3.1 *Flowchart* Diagram Sistem

Gambar 3.1 menjelaskan bagaimana alur pembangunan sistem dalam penelitian ini.



Gambar 3.1 *Flowchart* Diagram Sistem

3.2 Dataset

Dataset yang digunakan pada penelitian Tugas Akhir ini adalah *dataset* yang bersumber dari *website theprofessorisin.com* dan tersedia pada *link https://github.com/amir-karami/Workplace_Sexual_Harassment*. *dataset* diilustrasikan dalam lampiran 1.

3.3 Pelabelan *Dataset*

Pelabelan dilakukan secara manual dengan cara mencocokkan kata kedalam 2 kelas yaitu: *Quid Pro Quo* (qpq) dan *Hostile Work Environment* (hwe).

3.4 Pre-Processing

Metode *pre-processing* merupakan peran yang sangat penting dalam teknik dan aplikasi *text mining*. Ini adalah langkah pertama dalam proses penambangan teks[6].

3.4.1 Case Folding

Merupakan metode konversi seluruh karakter dalam seluruh dokumen ke dalam suatu bobot atau dalam hal ini merubah seluruh kalimat menjadi *lower case*.Implementasi dari proses *Case Folding* digambarkan dalam lampiran 2.

3.4.2 Data Cleaning

Text cleaning adalah proses yang digunakan untuk menghilangkan huruf yang tidak diperlukan seberti simbol,angka, dan *url/link* [10]. Implementasi dari proses *data cleaning* digambarkan dalamlampiran 3.

3.4.3 Tokenization

merupakan metode yang digunakan untuk memecah kalimat menjadi kata yang terpisah dikenal dengan nama *term* atau token[7]. Implementasi dari proses *Tokenization* digambarkan dalam lampiran 4.

3.4.4 Remove Stopword

merupakan metode yang bertujuan untuk menghapus kata yang tidak bermanfaat atau tidak memiliki pengaruh dalam proses[5]. Implementasi dari proses *Remove Stop Word* digambarkan dalam lampiran 5.

3.4.5 Stemming

merupakan metode untuk mendapatkan kata dasar dari kata yang telah mendapatkan imbuhan atau keterangan lainnya[5]. Implementasi dari proses *Stemming* digambarkan dalam lampiran 6.

3.5 TF-IDF

Term frequence (TF) dan *Inverse document Frequency (IDF)* adalah pembobotan yang paling sering digunakan . Metode *TF-IDF* merupakan cara untuk mencari bobot suatu kata (*term*) pada sebuah dokumen[18], yaitu dengan menghitung frekuensi kemunculan kata (*TF*) dan melakukan perhitungan invers terhadap frekuensi dokumen yang mengandung kata tersebut (*IDF*) [9]. Berikut ilustrasi TF-IDF pada sampel data yang digambarkan pada tabel 3.1.

Word1 = raped by customer. (Hostile Work Environment)
Word2 = asked to get touch for recommendation letter (Quid pro Quo)
D1 = *Hostile Work Environment*
D2 = *Quid Pro Quo*

Tabel 3.1 Ilustrasi Implementasi *TF-IDF*

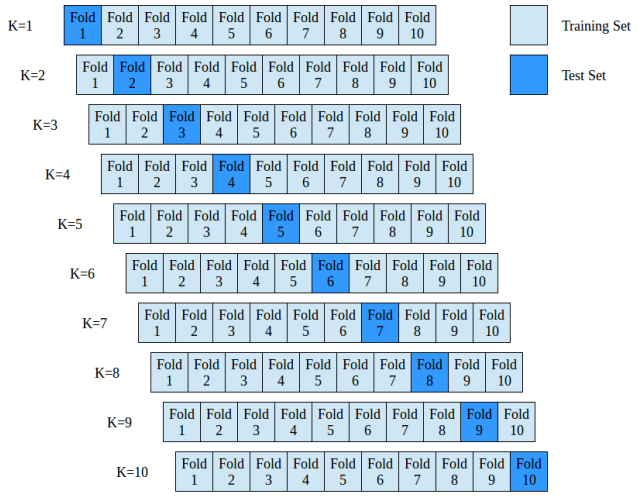
Word	***TF***		***DF***	***IDF***	***TF-IDF***	
	D1	**D2**			***D1***	***D2***
Raped	1	0	1	log(2/1)= 0.301	0.301	0
By	1	0	1	log(2/1)= 0.301	0.301	0
Customer	1	0	1	log(2/1)= 0.301	0.301	0
Asked	0	1	1	log(2/1)= 0.301	0	0.301
To	0	1	1	log(2/1)= 0.301	0	0.301
Get	0	1	1	log(2/1)= 0.301	0	0.301
Touch	0	1	1	log(2/1)= 0.301	0	0.301
For	0	1	1	log(2/1)= 0.301	0	0.301
recomendation	0	1	1	log(2/1)= 0.301	0	0.301
Letter	0	1	1	log(2/1)= 0.301	0	0.301

3.6 Random *Oversampling*

Teknik ini digunakan karena data yang didapat memiliki data *imbalance* yang dapat dilihat pada lampiran 8. Dari lampiran 8 dapat dikalkulasikan ke dalam persentase sehingga bobot dari pelabelan untuk hwe dan qpq (88.43 : 11.57) maka dilakukan teknik *oversampling.* Masalah *dataset* tidak seimbang adalah jenis khusus dari masalah klasifikasi di mana prioritas kelas sangat tidak sama dan tidak seimbang[11].

3.7 K-Fold Cross Validation

Pada teknik ini, data akan dibagi sesuai dengan banyaknya lipatan *(fold)* yang ditentukan. Cara kerja *K-Fold Cross Validation* dengan sepuluh lipatan (*10-fold cross validation*) diilustrasikan dalam Gambar 3.4.



Gambar 3.4 Ilustrasi 10-*fold cross validation*

3.8 Model Klasifikasi *Naïve Bayes*

Model algoritma *Naive Bayes Classifier* memiliki tingkat kesalahan yang sangat minimum dan dikenal dengan perhitungannya yang sederhana, cepat, dan sangat akurat [12] dan penggunaan N*aïve Bayes* akan lebih baik jika data train banyak. *Naïve Bayes* membangun model probabilistik dari *term*

documents matrix data labeled[13]. Teorema dikombinasikan dengan *Naive* yaitu mengasumsikan kondisi antar atribut saling bebas.[10]. Berikut teorema dari *Naïve bayes*:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

x = Data dengan kelas yang belum diketahui
c = Hipotesa data X merupakan suatu kelas spesifik
P(c|x) = Probabilitas hipotesis H berdasarkan kondisi X (*posterior probability*)
P(c) = Probabilitas hipotesis H (*prior probability*)

3.9 Model Klasifikasi KNN

K-NN merupakan metode yang memperhatikan jarak terdekat antara data latih dan objek yang akan diklasifikasi[19], sehingga model ini sering disebut juga lazy learning[19]. Formula *Euclidean Distance* d(x,y)[19] antara x dan y adalah:

$$\mathrm{d}(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{i=1}^{n} a_i(x) \quad - \quad a_i(a))^2} \quad (2)$$

3.10 Evaluasi

Pada tahap evaluasi, penulis menggunakan metode *confusion matrix* untuk menghitung nilai *accuracy*, *precision*, *recall*, dan *F1 score*. Formula yang digunakan sebagai berikut:

Formula *accuracy*:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{Total\ TP}{Total\ Dataset} \quad (3)$$

Formula *Precision:*:

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{Total\ Prediction} \quad (4)$$

Formula *Recall:*:

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{Total\ Actual} \quad (5)$$

Formula *F1-Score:*:

$$F1 - Score_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (6)$$

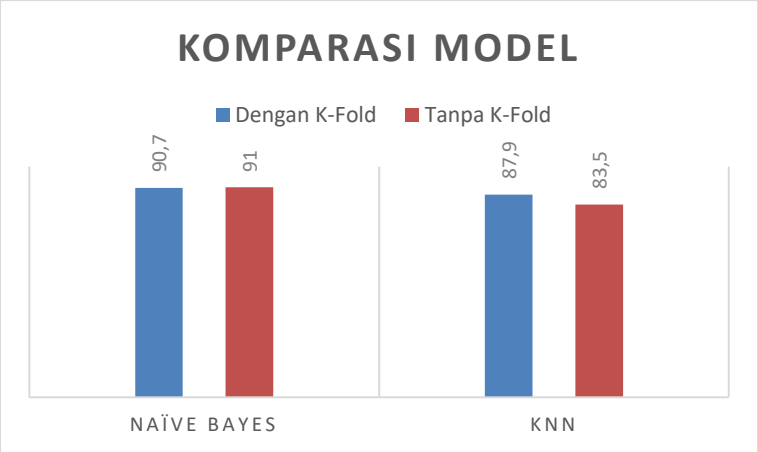
Cara kerja *confusion matrix* diilustrasikan pada lampiran 7.

4. Evaluasi

4.1 Hasil Pengujian

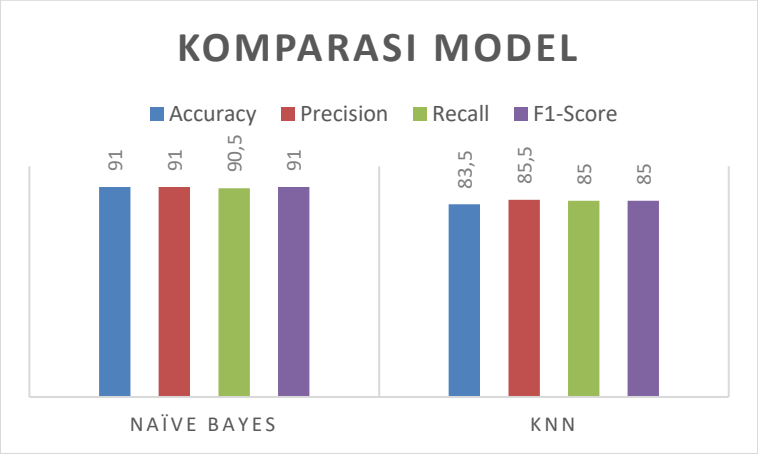
4.1.1 Akurasi

Ujicoba sistem pada penelitian ini dilakukan sesuai yang telah dipaparkan pada bab 3. Hasil akurasi dari uji coba ini akan dikelompokkan dalam bentuk tabel.yang dapat dilihat pada Gambar 4.1.

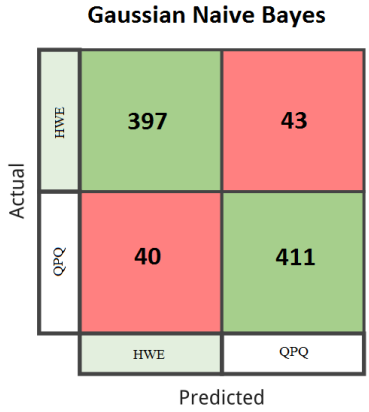


Gambar 4.1 Hasil Komparasi *Score*

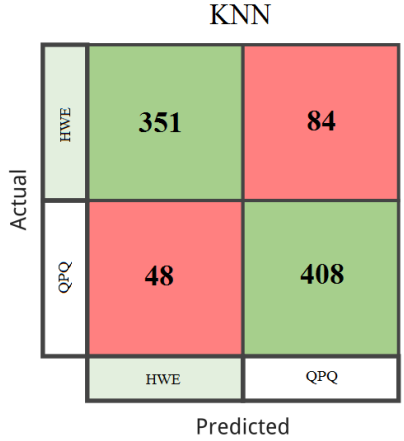
Dengan nilai Akurasi pada Gambar 4.1, akurasi dari *Gaussian Naïve Bayes* dengan K-Fold sebesar 90,7% dan tanpa K-Fold sebesar 91%. Untuk akurasi KNN dengan K-Fold sebesar 87.9 % dan tanpa K-Fold sebesar 83.5% untuk hasil *Precision*, *Recall*, *F1-Score*, dan *confussion matrix* dari model klasifikasi dapat dilihat pada Gambar 4.2, Gambar 4.3 dan Gambar 4.4



Gambar 4.2 Hasil Komparasi *akurasi, presisi, recall,* dan *F1-score*



Gambar 4.3 Hasil *Confussion Matrix Naïve Bayes*

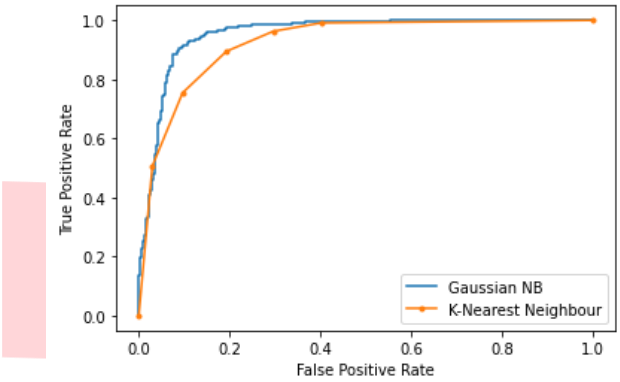


Gambar 4.4 Hasil *Confussion Matrix KNN*

Tabel 4.1 Komparasi nilai AUROC

Model	AUROC *Score*
Naïve Bayes	0.954935
KNN	0.923682

Dapat dilihat pada Gambar 4.1 nilai presisi dari *Naïve Bayes* sebesar 91% dan untuk KNN sebesar 85.5%. Nilai *recall* yang dihasilkan untuk *Naïve Bayes* sebesar 90.5% dan untuk KNN sebesar 85% . Nilai *F-1 score* yang dihasilkan dari *Naïve Bayes* sebesar 91% dan KNN sebesar 85%. Dengan nilai Akurasi dan *Confussion matrix* yang telah diuji coba pada penelitian ini maka dihasilkan *score AUROC* dari *Naïve Bayes* adalah 0.954935 dan untuk *score* AUROC dari KNN adalah 0.923682 yang dapat dilihat pada Tabel 4.1. Maka dihasilkan *ROC Curve* yang dapat dilihat pada Gambar 4.5 yang merupakan persebaran *confussion matrix* dari masing-masing model klasifikasi.



Gambar 4.5 Hasil *ROC Curve*

5. Kesimpulan

Dari hasil pengujian dan analisis pada penelitian ini didapatkan kesimpulan sebagai berikut:

- Sistem yang dibangun dari model klasifikasi *Naïve Bayes* dengan *K-Fold cross validation* memiliki nilai akurasi sebesar 90.7%, untuk model klasifikasi *Naïve Bayes* tanpa *K-Fold cross validation* sebesar 91% dan untuk K-NN dengan *K-Fold cross validation* sebesar 87.9 %, dan untuk K-NN tanpa *K-Fold cross validation* sebesar 83.5%. Hal ini dipengaruhi oleh pemakaian pendekatan *Random Oversampling* dan juga data yang digunakan tergolong sedikit dimana *Naïve Bayes* sangat baik diterapkan. Dengan adanya pendekatan ini, tingkat keakuratan dari model klasifikasi *Naïve Bayes* semakin efektif.
- *Dataset* yang digunakan pada penelitian ini merupakan data yang tidak seimbang dimana perbandingan dari kelas *Hostile Work Environtment* dan *Quid pro Quo* sebesar 88:12. Jarak kelas sangat jauh dari data yang ideal.

Adapun saran yang penulis berikan untuk penelitian selanjutnya yaitu:

1. Gunakan data yang telah melalui *pre-processing*. Hal ini dapat membantu mempersingkat waktu pemprosesan data.
2. Gunakan data yang seimbang sehingga hasil performa dari sistem akan lebih baik.
3. Gunakan *Tuning* dan data yang lebih banyak untuk meningkatkan nilai akurasi dari sistem

Referensi :

[1] D. N. Simorangkir, M. S. Saraswati, E. Melissa, L. L. K. Perangin-angin, and S. Schumacher, “IN THE MEDIA INDUSTRY,” vol. 4, no. 3, pp. 332–340, 2020.

[2] S. Widaningsih, “Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, Knn Dan Svm,” *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019, doi: 10.36787/jti.v13i1.78.

[3] C. Girgis, “Sexual harassment,” *Burn. Women Physicians Prev. Treat. Manag.*, pp. 105–128, 2020, doi: 10.1007/978-3-030-44459-4_6.

[4] S. Xu, “Bayesian Naïve Bayes classifiers to text classification,” *J. Inf. Sci.*, vol. 44, no. 1, pp. 48–59, 2018, doi: 10.1177/0165551516677946.

[5] Z. Efendi and M. Mustakim, “Text Mining Classification sebagai Rekomendasi Dosen Pembimbing Tugas Akhir Program Studi Sistem Informasi,” *Semin. Nas. Teknol. Inf. Komun. dan Ind.*, vol. 0, no. 0, pp. 235–242, 2017, [Online]. Available: http://ejournal.uin-suska.ac.id/index.php/SNTIKI/article/view/3273.

[6] S. Vijayarani, M. J. Ilamathi, M. Nithya, A. Professor, and M. P. Research Scholar, “Preprocessing Techniques for Text Mining -An Overview,” vol. 5, no. 1, pp. 7–16.

[7] M. Wongkar and A. Angdresey, “Sentiment Analysis Using Naive Bayes Algorithm Of The Data

Crawler: Twitter," *Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019*, pp. 1–5, 2019, doi: 10.1109/ICIC47613.2019.8985884.

[8] M. A. Maricar and Dian Pramana, "Perbandingan Akurasi Naïve Bayes dan K-Nearest Neighbor pada Klasifikasi untuk Meramalkan Status Pekerjaan Alumni ITB STIKOM Bali," *J. Sist. dan Inform.*, vol. 14, no. 1, pp. 16–22, 2019, doi: 10.30864/jsi.v14i1.233.

[9] D. A. Prabowo, M. Fhadli, M. A. Najib, H. A. Fauzi, and I. Cholissodin, "TF-IDF-Enhanced Genetic Algorithm Untuk Extractive Automatic Text Summarization," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, p. 208, 2016, doi: 10.25126/jtiik.201633217.

[10] K. Y. Raharja, H. Oktavianto, and R. Umilasari, "PERBANDINGAN KINERJA ALGORITMA GAUSSIAN NAIVE BAYES DAN K-NEAREST NEIGHBOR (KNN) UNTUK MENGKLASIFIKASI PENYAKIT HEPATITIS C VIRUS (HCV) C . Cara untuk melakukan diagnosa dini terhadap penyakit Hepatitis C adalah potensial dan berguna yang tersimpan da," pp. 1–12, 2021.

[11] A. Liu, J. Ghosh, and C. Martin, "Generative oversampling for mining imalanced datasets," *Int. Conf. Data Min.*, pp. 25–28, 2007.

[12] Y. F. Safri, R. Arifudin, and M. A. Muslim, "K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor," *Sci. J. Informatics*, vol. 5, no. 1, p. 18, 2018, doi: 10.15294/sji.v5i1.12057.

[13] A. P. Wijaya and H. A. Santoso, "Naive Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government Naïve Bayes Classification on Document Classification to Identify E-Government Content," *J. Appl. Intell. Syst.*, vol. 1, no. 1, pp. 48–55, 2016.

[14] W. Perkins and J. Warner, "Sexual Violence Response and Prevention: Studies of Campus Policies and Practices," *J. Sch. Violence*, vol. 16, no. 3, pp. 237–242, 2017, doi: 10.1080/15388220.2017.1318569.

[15] E. C. Kenneth and N. Heffernan, "Cybercrime detection in communications: An Experimental case of cyber Sexual Harassment Accuracy detection on Twitter Using Supervised Learning Classifiers MSc Internship Cybersecurity," 2020, [Online]. Available: http://norma.ncirl.ie/id/eprint/4499.

[16] F. Tempola, M. Muhammad, and A. Khairan, "Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 577, 2018, doi: 10.25126/jtiik.201855983.

[17] Harrington, Peter. Machine learning in action. Simon and Schuster, 2012.

[18] K. Budiman, N. Zaatsiyah, U. Niswah, F. Muhanna, and N. Faizi, "Analysis of Sexual Harassment Tweet Sentiment on Twitter in Indonesia using Naïve Bayes Method through National Institute of Standard and Technology Digital Forensic Acquisition Approach," *J. Adv. Inf. Syst. Technol.*, vol. 2, no. 2, pp. 21–30, 2020, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/jaist.

[19] M. Ramadhani and D. H. Murti, "Klasifikasi Ikan Menggunakan Oriented Fast and Rotated Brief (Orb) Dan K-Nearest Neighbor (Knn)," *JUTI J. Ilm. Teknol. Inf.*, vol. 16, no. 2, p. 115, 2018, doi: 10.12962/j24068535.v16i2.a711.

[20] *Title VII of the Civil Rights Act* of 1964

[21] B. Fileborn, "Justice 2.0: Street harassment victims' use of social media and online activism as sites of informal justice," *Br. J. Criminol.*, vol. 57, no. 6, pp. 1482–1501, 2017, doi: 10.1093/bjc/azw093.

Lampiran 1.

No	*Gender*	*Publication Year*	*Type*	*Experience*
1	*Male*	*2017*	*hwe*	*There were rumors (unsubstantiated, just gossip) that my advisor had an affair with one of his female graduate students before I was a student here. An older grad student who supposedly knew of the affair told me*
2	*Female*	*2018*	*Hwe*	*When I was in grad school a male faculty member "joked" to a group of three female PhD students (myself included) who had just mentioned how stressed we were about comps*
3	*Male*	*2018*	*Hwe*	*A senior colleague made overt sexual comments to me, including describing himself naked and having sex*

Lampiran 2.

Kalimat	Case Folding
There were rumors (unsubstantiated, just gossip) that my advisor had an affair with one of his female graduate students before I was a student here. An older grad student who supposedly knew of the affair told me	*there were rumors (unsubstantiated, just gossip) that my advisor had an affair with one of his female graduate students before i was a student here. an older grad student who supposedly knew of the affair told me*

Lampiran 3.

Kalimat	*Data Cleaning*
There were rumors (unsubstantiated, just gossip) that my advisor had an affair with one of his female graduate students before I was a student here. An older grad student who supposedly knew of the affair told me	*there were rumors unsubstantiated just gossip that my advisor had an affair with one of his female graduate students before i was a student here an older grad student who supposedly knew of the affair told me*

Lampiran 4.

Kalimat	*Tokenization*
there were rumors unsubstantiated just gossip that my advisor had an affair with one of his female graduate students before i was a student here an older grad student who supposedly knew of the affair told me	*"there" "were" "rumors" "unsubstantiated" "just" "gossip" "that" "my" "advisor" "had" "an" "affair" "with" "one" "of" "his" "female" "graduate" "students" "before" "i" "was" "a" "student" "here" "an" "older" "grad" "student" "who" "supposedly" "knew" "of" "the" "affair" "told" me*

Lampiran 5.

Kalimat	*Remove Stop Word*
"there" "were" "rumors" "unsubstantiated" "just" "gossip" "that" "my" "advisor" "had" "an" "affair" "with" "one" "of" "his" "female" "graduate" "students" "before" "i" "was" "a" "student" "here" "an" "older" "grad" "student" "who" "supposedly" "knew" "of" "the" "affair" "told" me	*"rumors" "'unsubstantiated" "gossip" "advisor" "affair" "one" "female" "graduate" "students" "student" "older" "grad" "student" "supposedly" "knew" "affair" "told"*

Lampiran 6.

Kalimat	***Stemming***
“rumors” “'unsubstantiated” “gossip” “advisor” “affair” “one” “female” “graduate” “students” “student” “older” “grad” “student” “supposedly” “knew” “affair” “told”	*'rumor', 'unsubstanti', 'gossip', 'advisor', 'affair', 'one', 'femal', 'graduat', 'student', 'student', 'older', 'grad', 'student', 'suppos', 'knew', 'affair', 'told'*

Lampiran 7.

	Prediction		
Actual		*Hostile Work Environment*	*Quid Pro Quo*
	Hostile Work Environment	*TP*	-
	Quid Pro Quo	-	*TP*

Lampiran 8.