Prediksi Tweet Netizen Menggunakan Random Forest, Decision Tree, Naïve Bayes, dan Ensemble Algorithm

Vivi nadenia Harahap[*1], Deci Irmayani[2], Syaiful Zuhri Harahap[3]
[1,2,3]Universtias Labuhan Batu, Rantau Prapat, Indonesia
Email: nadeniaharahap@gmail.com[*1], deacyirmayani@gmail.com[2],
syaifulzuhriharahap@gmail.com[3]

Abstrak

Gubernur DKI Jakarta saat ini, meski sudah terpilih sejak tahun 2017 selalu menarik untuk dibicarakan atau bahkan dikomentari. Komentar yang muncul berasal dari media secara langsung atau melalui media sosial. Twitter menjadi salah satu media sosial yang sering digunakan sebagai media untuk mengomentari gubernur terpilih bahkan bisa menjadi trending topic di media sosial Twitter. Netizen yang berkomentar pun beragam, ada yang selalu menge-Tweet kritik, ada yang berkomentar Positif, dan ada pula yang hanya me-retweet. Dalam penelitian ini, prediksi apakah Netizen aktif akan cenderung selalu menimbulkan komentar Positif atau Negatif akan dilakukan dalam penelitian ini. Model algoritma yang digunakan adalah Decision Tree, Naïve Bayes, Random Forest, dan juga Ensemble. Data Twitter yang diolah harus melalui preprocessing terlebih dahulu sebelum dilanjutkan menggunakan Rapidminer. Dalam uji coba menggunakan Rapidminer dilakukan dalam empat kali uji coba dengan membagi menjadi dua bagian yaitu data testing dan data latih. Perbandingan yang dilakukan adalah 10% data pengujian: 90% data pelatihan, kemudian 20% data pengujian: 80% data pelatihan, kemudian 30% data pengujian: 70% data pelatihan, dan yang terakhir adalah 35% data pengujian: 65% data pelatihan. Rata-rata Akurasi untuk algoritma Decision Tree adalah 93,15%, sedangkan untuk algoritma Naïve Bayes Akurasinya adalah 91,55%, kemudian untuk algoritma Random Forest adalah 93,41, dan yang terakhir adalah algoritma Ensemble dengan Akurasi sebesar 93,42%. sini. 65% data pelatihan. Rata-rata Akurasi untuk algoritma Decision Tree adalah 93,15%, sedangkan untuk algoritma Naïve Bayes Akurasinya adalah 91,55%, kemudian untuk algoritma Random Forest adalah 93,41, dan yang terakhir adalah algoritma Ensemble dengan Akurasi sebesar 93,42%. sini. 65% data pelatihan. Rata-rata Akurasi untuk algoritma Decision Tree adalah 93,15%, sedangkan untuk algoritma Naïve Bayes Akurasinya adalah 91,55%, kemudian untuk algoritma Random Forest adalah 93,41, dan yang terakhir adalah algoritma Ensemble dengan Akurasi sebesar 93,42%. sini.

Kata kunci- *Pohon Keputusan, Naïve Bayes, Hutan Acak, Set, Twitter*

Abstract

The current governor of DKI Jakarta, even though he has been elected since 2017, is always interesting to talk about or even comment on. The comments that appear come from the media directly or through social media. Twitter is one of the social media that is often used as a medium for commenting on the elected governor, it can even become a trending topic on Twitter social media. Netizens who commented also varied, some always tweeted criticism, some commented positively, and some just retweeted. In this study, predictions of whether active netizens will tend to always cause positive or negative comments will be carried out in this study. The algorithm model used is Decision Tree, Naïve Bayes, Random Forest, and also Ensemble. Twitter data that is processed must go through preprocessing before continuing to use Rapidminer. In the trial using Rapidminer, it was carried out in four trials by dividing it into two parts, namely testing data and training data. The comparison made is 10% test data: 90% training data, then 20% test data: 80% training data, then 30% test data: 70% training data, and the last is 35% test data: 65% training data. The average accuracy for the Decision Tree algorithm is 93.15%, while for the Naïve Bayes algorithm the

accuracy is 91.55%, then for the Random Forest algorithm is 93.41, and the last one is the Ensemble algorithm with an accuracy of 93.42%. here. 65% training data. The average accuracy for the Decision Tree algorithm is 93.15%, while for the Naïve Bayes algorithm the accuracy is 91.55%, then for the Random Forest algorithm is 93.41, and the last one is the Ensemble algorithm with an accuracy of 93.42%. here. 65% training data. The average accuracy for the Decision Tree algorithm is 93.15%, while for the Naïve Bayes algorithm the accuracy is 91.55%, then for the Random Forest algorithm is 93.41, and the last one is the Ensemble algorithm with an accuracy of 93.42%. here.

Keywords- *Decision Tree, Naïve Bayes, Random Forest, Set, Twitter*

1. PENDAHULUAN

Twitter adalah layanan jejaring sosial dan microblogging online yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks (Wikipedia, 2019). Begitu juga dengan akun Twitter resmi Gubernur DKI Jakarta, @aniesbaswedan. Bapak Anies Rasyid Baswedan dan Sandiaga Salahuddin Uno menjabat sebagai Gubernur dan Wakil Gubernur Provinsi DKI Jakarta periode 2017-2022. Jumlah berita tentang gubernur termasuk Tweet yang diposting di akun Bpk. Anies Rasyid Baswedan, baik yang memiliki Positif, Negatif, dan Netral, seperti pada Gambar 1 dan Gambar 2 di bawah ini:

Gambar 1 Contoh berita tentang Gubernur DKI Jakarta

Gambar 2 Contoh Tweet dari Netizen

Penelitian ini akan mengambil data atau Tweet yang diposting oleh Netizen di akun Twitter resmi Bpk. Anies Rasyid Baswedan, @aniesbaswedan. Pengambilan data menggunakan aplikasi Rapidminer, dilakukan secara efektif dan efisien yang kemudian dilakukan pelabelan Positif dan Negatif oleh pihak ketiga yaitu sebanyak 100 responden. Pelabelan berguna untuk menganalisis pendapat seseorang, penilaian seseorang, sikap seseorang, dan emosi seseorang ke dalam bahasa tulisan, dalam hal ini bisa disebut sentimen. Salah satu disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari data yang besar adalah data mining. Data

mining adalah proses penggalian untuk mendapatkan informasi penting yang tersirat dan tidak diketahui sebelumnya, dari data (Witten et al., 2011). Sangat menarik (non-sepele, implisit, sebelumnya tidak diketahui, dan berpotensi berguna) pola atau pengetahuan dari sejumlah besar data (Jiawei Han & Kamber, 2013). Data mining sering dianggap sebagai bagian dari Knowledge Discovery in Database (KDD), yang merupakan proses menemukan pengetahuan yang berguna dari data. Selain itu, data mining juga dikenal sebagai ekstraksi pengetahuan, analisis pola, pengumpulan informasi, dan intelijen bisnis.

Ada 5 peran utama data mining, yaitu: Estimasi, Prediksi, Klasifikasi, Klasifikasi, dan Asosiasi. Algoritma data mining yang sering digunakan dalam klasifikasi antara lain Naïve Bayes, K-Nearest Neighbors, Decision Tree, ID3, CART, Linear Discriminant Analysis, Logistic Regression, Ensemble, dan lain-lain. Namun dalam penelitian ini penulis hanya akan menggunakan algoritma Random Forest, Decision Tree, Naïve Bayes, dan Ensemble untuk mengolah, mengklasifikasikan, dan pengetahuan saya dari dataset Twitter pada akun @aniesbaswedan.

Dalam penambangan data, penelitian tentang klasifikasi posting Twitter telah dilakukan oleh peneliti lain. Sebagian besar

2. METODE PENELITIAN

Fase Pemahaman Bisnis / Riset

Tahap ini merupakan pemahaman terhadap objek penelitian. Dalam penelitian ini penulis menggunakan data Twitter dari akun resmi gubernur DKI Jakarta yaitu @aniesbaswedan pada periode 28 Sept 2019 S/d 09 November 2019. Pengambilan data Twitter Twitter menggunakan aplikasi pada Rapidminer. Pada tahap ini juga dilakukan pemahaman untuk menemukan label Positif dan Negatif pada teks yang diposting oleh pengguna. Selain label teks, juga dapat diperoleh Aktif dan Pasif dari pengguna Twitter.

Fase Pemahaman Data

Tahap ini merupakan proses memahami data yang akan dijadikan bahan yang akan diteliti untuk dilakukan ke tahap selanjutnya yaitu Preprocessing. Di bawah ini adalah langkah-langkah yang akan dilakukan.

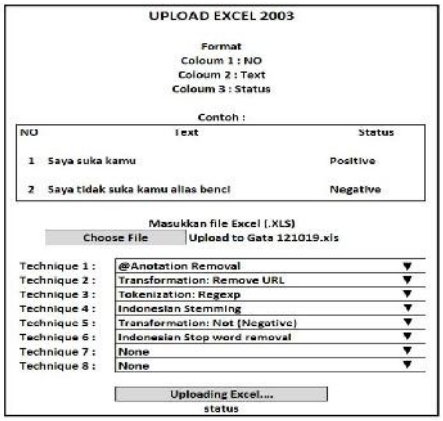
Siapkan Total Data Set dari akun pribadi Twitter @aniesbaswedan dan download data sebanyak 29.340 tweet, data tweet di download menggunakan tools dari Rapidminer, kemudian lanjutkan dalam format Excel. Dataset yang sudah disimpan di Excel diproses lebih lanjut untuk mengidentifikasi tweet duplikat di postingan, dengan kata lain Dataset dibersihkan dengan proses yang disebut pembersihan data. Setelah dilakukan Cleaning diperoleh data sebanyak 12.027 yang dapat digunakan. Pada penelitian ini hanya mengambil 10.000 data yang terdiri dari Label Positif sebanyak 5.000 dan Label Negatif sebanyak 5.000 data. Ini.

pelabelan melibatkan 100 responden menggunakan metode pelabelan Crowdsourced, yaitu metode pelabelan data yang melibatkan partisipasi khalayak umum. Proses pelabelan untuk dataset yang tidak memerlukan keahlian khusus atau peserta studi dalam memberikan pelabelan (Rachmat & Lukito, 2016). Banyak responden akan mempercepat proses pemberian label dan juga akan lebih netral dalam label. Hal lain yang juga diuntungkan penulis dari banyaknya responden dalam proses ini, adalah tidak perlu biaya yang besar jika dibandingkan dengan menggunakan Bantuan tenaga ahli untuk melakukannya.

Persiapan dan Pemodelan Data

Langkah selanjutnya adalah mempersiapkan data sebelum data tersebut akan dimodelkan atau disebut dengan Data Preparation. Untuk tahap 2 ini mempersiapkan data untuk melakukan langkah-langkah yang dikenal dengan text preprocessing, menggunakan dua aplikasi preprocessing, pertama menggunakan Gataframework diakses melalui Link http://Gataframework.com/textmining yang dapat digunakan secara gratis juga mudah digunakan karena tidak harus membuat akun untuk menggunakan makanan dan melanjutkan praproses Rapidminer.

Tahap selanjutnya adalah preprocessing Rapidminer dengan urutan yang ditunjukkan pada Gambar 4 di bawah ini:



Gambar 4 Gambar tampilan Gataframework

1. @Penghapusan Anotasi

   Langkah pertama adalah teks diurai dengan spasi, semua anotasi yang terdapat dalam Tweet akan dihilangkan dan huruf kecil atau ubah huruf dalam teks menjadi huruf kecil semua, seperti contoh berikut Tabel 1:

Tabel 1. Perbandingan Teks sebelum dan sesudah Proses Penghapusan

Text	Sesudah Proses Penghapusan
Tenangkan Massa @ganjarpranowo Turun ke Tengah Mahasiswa, kalau @aniesbaswedan? #AniesGaBener https://t.co/t9lqYpmGLx	Tenangkan Massa @ganjarpranowo Turun ke Tengah Mahasiswa, kalau @aniesbaswedan? #AniesGaBener https://t.co/t9lqYpmGLx
@aniesbaswedan @DKIJakarta @pln_123 @PT_TransJakarta @DishubDKI_JKT @dinaslhdki Hati hati Pak @aniesbaswedan sampai saat ini saja wakilnya belum ada... kasihan @PKSejahtera di zholimi terus. https://t.co/K294APLorx	hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus. https://t.co/k294aplorx
hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus. https://t.co/k294aplorx	hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus. https://t.co/k294aplorx
hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus. https://t.co/k294aplorx	hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus. https://t.co/k294aplorx

2. Transformasi: Hapus URL

   Seringkali URL muncul dari data Twitter Twitter membuat data tidak efektif dan tidak berarti. Untuk itu perlu untuk menghapus URL URL atau bias juga untuk menghapus link internet, seperti Tabel 2 berikut ini:

Tabel 2. Perbandingan Teks sebelum dan sesudah proses @ Transformasi: Hapus URL

	Teks		**Transformasi: Hapus URL**
tenangkan massa turun ke tengah			tenangkan massa turun ke tengah
mahasiswa,	kalau	#aniesgabener	
https://t.co/t9lqypmglx			
hati hati pak sampai saat ini saja			hati hati pak sampai saat ini saja

wakilnya belum ada... panggang di zholimi terus. https://t.co/k294aplorx	wakilnya belum ada... panggang di zholimi terus
siapapun presiden indonesia... pasti	siapapun presiden indonesia... pasti

3. Tokenisasi: Regexp

The tokenization process is performed after the transform cases. All unnecessary characters will be discarded. Includes excessive white space and all punctuation. This process will be done on any documents entered from the document collection. So it is obtained a unique word and can represent documents, such as the following Table 3 example:

Table 3.Table Comparison of Text before and after the @ tokenization process: regexp

Text	Tokenization: Regexp
tenangkan massa turun ke tengah mahasiswa, kalau #aniesgabener	tenangkan massa turun ke tengah mahasiswa kalau aniesgabener
hati hati pak sampai saat ini saja wakilnya belum ada... kasihan di zholimi terus.	hati hati pak sampai saat ini saja wakilnya belum ada kasihan di zholimi terus

4. Indonesian Stemming

After the result of the transformation not Negative will be followed by the steaming process is to remove the suffix that is found in each word so that it is a basic word using Indonesian stemming for a Tweet - speaking Indonesia, such for example Table 4 follows:

Tabel 4. Table Comparison of Text before and after the Indonesian stemming process

Text	Indonesian Stemming
tenangkan massa turun ke tengah mahasiswa kalau aniesgabener	tenang massa turun ke tengah mahasiswa kalau aniesgabener
hati hati pak sampai saat ini saja wakilnya belum ada kasihan di zholimi terus	hati hati pak sampai saat ini saja wakil belum ada kasihan di zholimi terus
siapapun presiden indonesia pasti ngutang gubernur nya aja ngutang	siapa presiden indonesia pasti ngutang gubernur nya aja ngutang
alhamdulillah mudahan dosanya pak anies terhapus karena fitnah ini	alhamdulillah mudah dosa pak anies hapus karena fitnah ini

5. Transformation: Not (Negative)

From the results of Tokenization (Regexp), the next process is transformation not Negative. For this example, in the text used previously, there was no change because there were no words made by Transformation Not Negative. But to clarify the purpose of the process, another text from the same local data is used, such as the example in Table 5 below:

Tabel 5. Comparison Of Text Before And After The Transformation Process: Not (Negative)

Text	Transformation: Not (Negative)
untuk erat tali saudara yg cerai berai oleh radikalisme pak gimana pak jawab saya udah keren belum	untuk erat tali saudara yg cerai berai oleh radikalisme pak gimana pak jawab saya udah keren belum_
tanya saya simple anda boleh orang	tanya saya simple anda boleh orang
dagang trotoar yang notabene buat untuk jalan kaki bukan untuk dagang iya atau	dagang trotoar yang notabene buat untuk jalan kaki bukan_untuk dagang iya atau

tidak	tidak_
lo me bicara harga juga semua dapat harga lo liat lapang sini lo orang jakarta bukan	lo me bicara harga juga semua dapat harga lo liat lapang sini lo orang jakarta bukan_
lha itu jpo sdh ada atap dul ngapain di lepas lain halnya kalo emang dr dolo ga ada kita trima dgn lapang dada buat bijak itu yg manfaat jgn buat bijak yg lebih tdk manfaat dr belum	lha itu jpo sdh ada atap dul ngapain di lepas lain halnya kalo emang dr dolo ga ada kita trima dgn lapang dada buat bijak itu yg manfaat jgn buat bijak yg lebih tdk manfaat dr belum_

6. Indonesian Stop Words removal
   This stop word stage will refine the token by length filter Stage. Words consisting of more than 3 letters and included in the stop words will be discarded. Because the word does not reflect the contents of the document even though it frequently appears, such as Table 6 example follows:

Tabel 6. Comparative Text before and after the Indonesian Stop Word removal process

Text	**Indonesian Stop word removal**
tenang massa turun ke tengah mahasiswa kalau aniesgabener	tenang massa turun mahasiswa aniesgaben er
hati hati pak sampai saat ini saja wakil belum_ada kasihan di zholimi terus	hati hati wakil belum_ada kasihan zholimi
siapa presiden indonesia pasti ngutang gubernur nya aja ngutang	presiden indonesi ngutang gubernur a ngutang
gw maklum lah lho pantas aja lho mati nyinyir pak sahabtanya dan buzzer lain	maklum lah lho mati nyinyir sahabtanya buzzer

RESULT

Model Decision Tree experiments and testing results

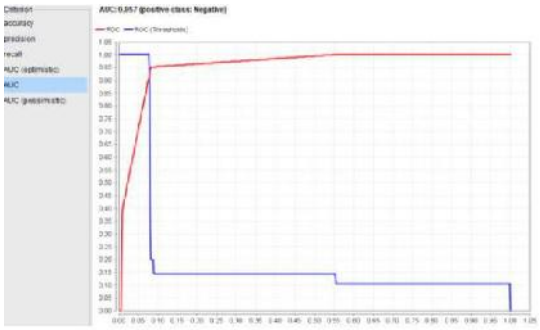
Of the 10,000 Text data that was posted and processed using the Decision Tree algorithm on Rapidminer with data testing comparison and data training 10:90 There are as many as 4152 data in positive predictions and fact positive, 4128 negative predicted data and reality negative, 372 data predicted positive but Negative and 348 negative predicted data But reality Positive, as in the following Table 7 below:

Table 7. Confusion Matrix Decision Tree data testing 10% and data training 90%.
accuracy: 92.00%

	true Positive	true Negative	class precision
pred.			
pred.			

class	72.72%	91.73%	

The ROC curve measurement by using the UnderCurve Area (AUC) resulting in an AUC value of 0. 957, as in Figure 5 below.



Source: Rapidminer Tools

Fig. 5 Images Under Curve Area graph (AUC) of Decision Tree, data testing 10% and training Data 90%

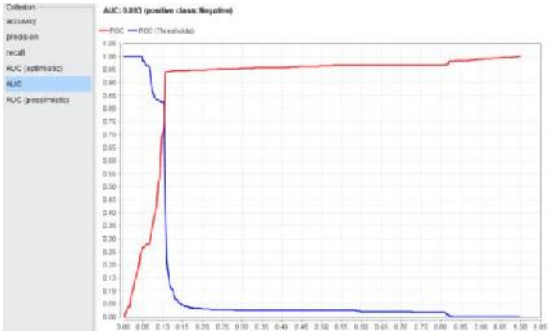
Test results of Naïve Bayes Model experiments and testing

Accuracy value obtained by testing 10% Data Testing comparison: The training data is 90%; Accuracy = 91.64%. Of the total 10,000 datasets were processed, as many as 4024 the amount of data predicted positive and positive, 4224 negative predicted data and negative, 276 predicted data positive but negative, and 476 negative predicted data But Positive as in Table 8 below.

Table 8 Confusion Matrix Naïve Bayes data testing 10% and data training 90% accuracy: 91.64%

	true Positive	true Negative	class precision
pred. Positive	4024	276	93.58%
pred. Negative	476	4224	89.87%
class recall	89.42%	93.87%	

ROC curve measurements using the Area Under Curve (AUC) which produces an AUC value of 0.893 as shown in Figure 6.



Source: Rapidminer Tools Fig. 6 Images Under Curve Area graph (AUC) Naïve Bayes algorithm, 10% data testing, and 90% training Data.

Results of experimental and Model Ensemble Vote

With a comparison of the 10% Data Set tester and the 90% training Data Set, it is generated by accuracy. Accuracy results of 91.44%. Of the total 10,000 datasets were processed, as many as 4152 the amount of data predicted positive and positive, 4224 negative predicted data and negative, 227 predicted data positive but negative, and 348 negative predicted data But Positive as in the Table 9 below:

Table 9
Confusion Matrix Ensemble, testing 10% data testing and 90% training Data accuracy: 91.44%

	true Positive	true Negative	class precision
pred. Positive	4152	277	93.75%
pred. Negative	348	4224	92.39%
class recall	92.27%	93.84%	

ROC curve measurements using the Area Under Curve (AUC) which produces an AUC value of 0.905 as shown in Figure 7

Source: Rapidminer Tools Fig. 7 Images Under Curve Area graph (AUC) of Ensemble algorithm, 10% data testing and 90% training Data

The results of the experiment and test Model of Random Forest

With a comparison of the 10% Data Set tester and the 90% training Data Set, it is generated by accuracy. Accuracy results of 93.08%. Of the total 10,000 datasets are processed, as many as 4153 amount of data predicted positive and positive, 4224 negative predicted data and negative, 226 predicted data positive but negative, and
347 negative predicted data But reality positive as in Table 10 below:

Table 10 Confusion Matrix Random Forest, testing 10% data testing and 90% training Data accuracy: 93.08%

	true Positive	true Negative	class precision
pred.			
pred.			
class			

ROC curve measurements using the Area Under Curve (AUC) which produces an AUC value of 0.962 as shown in Figure 8.

Source: Rapidminer Tools Fig. 8 Images Under Curve Area graph (AUC) of Ensemble algorithm, 10% data testing and 90% training Data

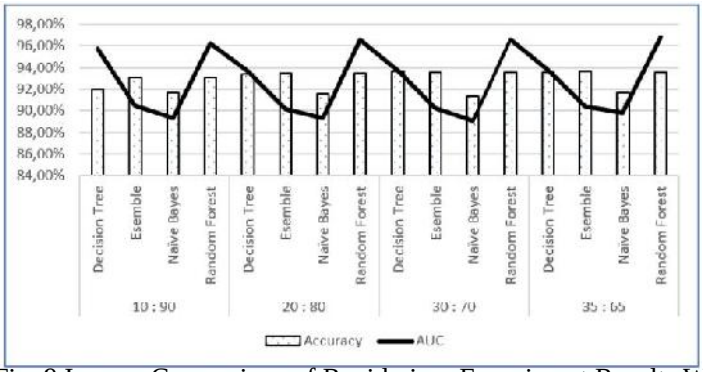
HASIL DAN PEMBAHASAN

Overall Model Comparison

The results of processing with Rapidminer above as a representative of 4 comparisons of data testing and data training in each algorithm. From table 3.7, we can see the comparison results of the four algorithms used in the research of this Random Forest, Decision Tree, Naïve Bayes, and Ensemble with Vote features, based on data sharing testing: The following data training; 10:90, 20:80, 30:70 and 35:65 on testing 10% data testing comparison and 90% training, the accuracy value of the Random Forest algorithm is 93.08% and higher than the other two algorithm, with the AUC value of 0.962. In comparative data testing 20% and training 80% training data, the accuracy value of the Random Forest algorithm is 93.45% with its AUC value of 0.966. In the last comparison with testing 30% Data testing and training of 70%, the Decision Tree algorithm with an accuracy of 93.60% and with an AUC value of 0937, the latter is a ratio of 35% to data testing and 65% for training data with the highest accuracy result in the Ensemble of 93.60% and the AUC of 0.904. For average overall experiments can be seen in Table 11 and Figure 9 below.

Table 11 Recapitulation for Rapidminer Test Data Set

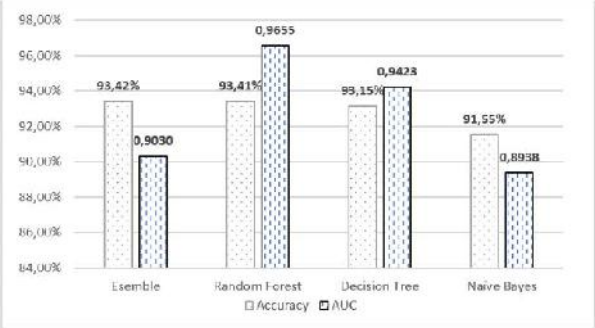
Algorithm	Testing	Training	Accuracy	AUC
Decision Tree	10%	90%	92,00%	0,957
Together	10%	90%	93,06%	0,905
Naïve Bayes	10%	90%	91,64%	0,893
Random Forest	10%	90%	93,08%	0,962
Decision Tree	20%	80%	93,40%	0,937



Source: Ms Excel Fig. 9 Images Comparison of Rapidminer Experiment Results Whereas if averaged from to four experiments on 4 algorithms, as in Table 12 and figure 10 follows.

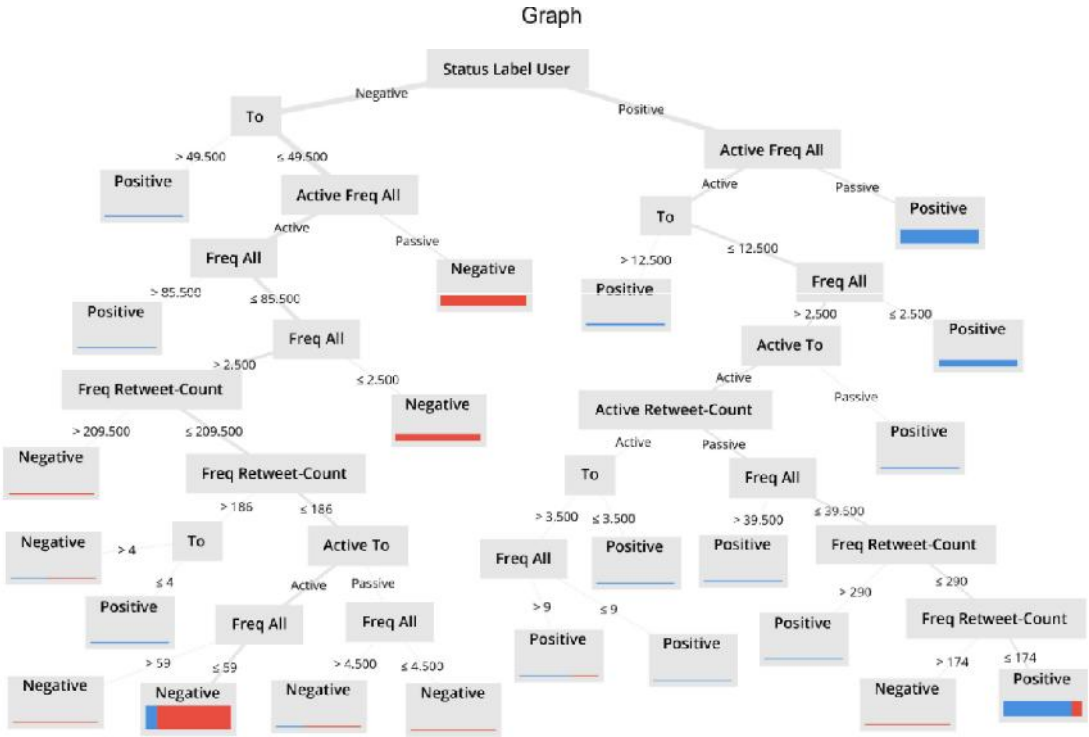
Table 12 Results on average dataset Rapidminer test datasets.

Algorithm	Accuracy	AUC
Together	93,42%	0,903
Random Forest	93,41%	0,9655
Decision Tree	93,15%	0,9423
Naïve Bayes	91,55%	0,8938



In table 3.8, we can see that the average of the four experimental data obtained an average accuracy and the average Under Curve Area (AUC) of the Ensemble algorithm with an accuracy of 93.42% and AUC of 0.9030. For the Random Forest Algorithm with an accuracy of 93.41 and an AUC of 0.9655. For the Decision Tree Algorithm with 93.15% accuracy and AUC of 0.9423. As for the Naïve Bayes algorithm with an accuracy of 91.55% and AUC of 0.8938.

Tweet Prediction Analysis Netizen



Source: Rapidminer Fig. 11 Image of Tree Diagram of a Random Forest model, with DataSet 2000

From the tree diagram in figure 11 above that can be from the Rapidminer process with a Random Forest model and the number of datasets as much as 2000 data. The Label Status of the User who is a netizen identity that posts a tweet can be predicted whether the tweet is Negative or Positive with the conditions and conditions as in the above image.

KESIMPULAN

From the comparison of Random Forest algorithms, Decision Tree, Naïve Bayes, and Ensemble, from the trial with the distribution of data testing: Data training 10%: 90%, 20%: 80%, 30%: 70% and 35%: 65%. Algorithms Random Forest excelled on test tests with compositions Data testing 10%: Training data 90% with 93.08% accuracy and AUC of 0.962, then superior to Data testing 20%: Training data 80% with accuracy 93.45 and AUC 0.966. While in test data testing 30%: Training Data 70%, superior to the Decision Tree with an accuracy of 93.60% and AUC 0.937. The final test is on data testing 35%: Training data training 65%, obtained an accuracy of 93.60%, and AUC 0.904 for Ensemble algorithm.

The Random Forest algorithm can be predicted to affect the Twitter or netizen user ID whether The Tweet posted to the account @aniesbaswedan the majority leads to sentiment Positive or Negative. The things that affect your user ID or netizen lead to positive or negative such as netizen Tweet frequency, then the contents of a tweet are positive or negative, also the type whether tweet or re-tweet.

On existing data sets, accuracy is influenced by algorithm selection and data testing comparisons with training data. This is evidenced by 4 experiments carried out a result of 3 of 4 algorithms that can excel its accuracy. Other things that also take the rise of accuracy and AUC are pre-processing over downloaded data sets.

DAFTAR PUSTAKA

[1] Al-Rubaiee, H., Qiu, R., & Li, D. (2016). Analysis of the relationship between Saudi twitter posts and the Saudi stock market. 2015 IEEE 7th International Conference on Intelligent Computing and Information Systems, ICICIS 2015, December, 660–665. https://doi.org/10.1109/IntelCIS.2015.7397193

[2] Alhamad, A., Azis, A. I. S., Santoso, B., & Taliki, S. (2019). Heart Disease Prediction using methods of Machine Learning based on Ensemble – Weighted Vote. 5(3), 352 – 360.

[3] Blatnik, A., Jarm, K., & Meža, M. (2014). Movie sentiment analysis based on public tweets. Elektrotehniski Vestnik/Electrotechnical Review, 81(4), 160–166.

[4] Buntoro, G. A. (2017). Analysis of candidates for governor of DKI Jakarta 2017 on Twitter. Integer Journal March,1(1),32–41.

[5] Cureg, M. Q., De La Cruz, J. A. D., Solomon, J. C. A., Saharkhiz, A. T., Balan, A. K. D., & Samonte, M. J. C. (2019). Sentiment analysis on tweets with punctuations, emoticons, and negations. ACM International Conference Proceeding Series, Part F1483(1), 266–270. https://doi.org/10.1145/3322645.3322657

[6] Da Silva, N. F. F., Hruschka, E. R., & Hruschka, E. R. (2014). Tweet sentiment analysis with classifier ensembles. Decision Support Systems. https://doi.org/10.1016/j.dss.2014.07.003

[7] Flux, A. W., Pareto, V. (1897). Political Economy Course. The Economic Journal. https://doi.org/10.2307/2956966

[8] Gorunescu, F. (2011). Data mining Concepts, Models, and Techniques. Verlag Berlin

[9] Heidelberg: Springer Han, J., & Kamber, M. (2007). Data mining Concepts and Techniques. Morgan Kaufmann publisher.

[10] Jiawei Han, & Kamber, M. (2013). Data Mining: Concepts and Techniques Second Edition. In Morgan Kaufmann. https://doi.org/10.1017/CBO9781107415324.004

[11] Larose, D. T. (2005). Discovering Knowledge in Data. New Jersey: John Willey & Sons, Inc.

[12] Kartiko, M., & Sfenrianto. (2019). Accuracy for Sentiment Analysis of Twitter Students on ELearning in Indonesia using Naive Bayes Algorithm Based on Particle Swarm Optimization. Journal of Physics: Conference Series, 1179(1). https://doi.org/10.1088/1742-6596/1179/1/012027

[13] Mentari, N. D., Fauzi, M. A., & Muflikhah, L. (2018). 2013 curriculum sentiment analysis on Twitter social Media using the K-Nearest Neighbor method and the Feature Selection Query Expansion Ranking. Journal of Information Technology and Computer science development (J-Ptiik) Universitas Brawijaya, 2(8), 2739 – 2743.

[14] Pratama, B., Saputra, D. D., Novianti, D., Purnamasari, E. P., Kuntoro, A. Y., Hermanto, Gata, W., Wardhani, N. K., Sfenrianto, S., & Budamsono, S. (2019). Sentiment Analysis of the Indonesian Police Mobile Brigade Corps Based on Twitter Posts Using the SVM and NB Methods. Journal of Physics: Conference Series, 1201(1). https://doi.org/10.1088/1742-6596/1201/1/012038

[15] Puyalnithi, T., V, M. V., & Singh, A. (2016). Comparison of Performance of Various Data Classification Algorithms with Ensemble Methods Using Rapidminer. 6(5), 1–6.

[16] Rachmat, A., & Lukito, Y. (2016). Implementation of WEB based Crowdsourced Labelling system with Weighted Majority Voting method. ULTIMA Infosys Journal, 6(2), 76 – 82. https://doi.org/10.31937/si.v6i2.223

[17] Ratul, A. R., & Engineering, F. (n.d.). A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining.

[18] Witten, I. H., Frank, E., & Hall, M. a. (2011). Data Mining: Practical Machine Learning Tools and Techniques (Google eBook). In Complementary literature None. http://books.google.com/books?id=bDtLM8CODsQC&pgis=1