JIMP : Jurnal Informatika Merdeka Pasuruan
Vol.7 No.2 Agustus 2022, P-ISSN : 2502-5716, E-ISSN : 2503-1945

Terakreditasi Peringkat Sinta 4 berdasarkan Petikan dari Keputusan Menteri Riset dan Teknologi/ Kepala Badan Riset dan Inovasi Nasional Nomor 200/M/KPT/2020, masa berlaku mulai Vol.3 No.1 tahun 2018 s.d Vol.7 No.2 tahun 2022

Studi Komparasi Metode SVM, Logistic Regresion dan Random Forest Clasifier untuk Mengklasifikasi Fake News di Twitter

Anosa Putri Ruise[1], Ahmad Sanusi Mashuri[2], Muhammad Sulaiman[3], Fazrur Rahman[4]
[1, 2, 3, 4] PJJ Magister Teknik Informatika, Universitas Amikom, Indonesia
[1] putriruise@students.amikom.ac.id
[2] sanusi@students.amikom.ac.id
[3] muhammadsulaiman@students.amikom.ac.id
[4] fazrurrahman@students.amikom.ac.id

Received: 30-05-2022; Accepted: 27-08-2023; Published: 09-09-2023

Abstrak— ***Media sosial adalah salah satu platform utama untuk mendapatkan berita dan informasi. Internet adalah media yang paling cerdas dan mudah didapat, juga secara signifikan membantu dalam mengembangkan kehidupan kita. Namun, itu juga memberikan kemudahan bagi tersebar luas berita palsu. Alasan di balik fakenews adalah menciptakan hype untuk mendapatkan perhatian audiens dan membangun dampak negatif pada saat itu. Deteksi berita palsu diperlukan untuk memurnikan lingkungan Internet. Maka perlu dilakukan pengklasifikasian fakenews menggunakan teknik klasifikasi dengan data mining dan menggunakan 3 metode yaitu SVM, Logistic Regresion dan Random Forest Clasifier dengan didapatkan tingkat akurasi SVM = 98%, Logistic Regresion = 97% dan Random Forest Clasifier = 97%.***

Kata kunci— **Natural Language Processing, SVM, Logistik Regresion, Random Forest Clasifier, Term of Matriks**

Abstract— Social media is one of the main platforms for getting news and information. Internet is the most intelligent and accessible medium, it also significantly helps in developing our lives. However, it also makes it easy for fake news to spread. The reason behind fakenews is to create hype to get the audience's attention and build a negative impact at the time. Fake news detection is necessar to purify the Internet environment. So it is necessary to classify fakenews using a classification technique with data mining and using 3 methods, namely SVM, Logistic Regression and Random Forest Classifier, with an accuracy rate of SVM = 98%, Logistic Regression = 97% and Random Forest Classifier = 97%.

Keywords— Natural Language Processing, SVM, Logistic Regression, Random Forest Classifier, Term of Matrix

I. PENDAHULUAN

Salah satu konsekuensi dari teknologi adalah berita palsu. Sekarang ini sedang maraknya penyebaran informasi atau berita melalui berbagai macam media online yang belum diketahui lisensinya. Berita tersebut dapat di bagikan oleh siapa saja pengguna internet. Hal ini tentunya memberikan kebingungan antar sesama masyarakat akan kebenaran berita karena sebagian berita tersebut dibagikan secara mentah-mentah tanpa keterangan yang dapat memicu kesalahpahaman dan terindikasi berita hoax [1]. Dengan perkembangan teknologi saat ini banyak pengguna sosial media khususnya pengguna twitter yang tidak bertanggung jawab memanfaatkannya untuk mendapatkan perhatian audiens dan membangun dampak negatif pada saat itu dan berkedok penipuan lainnya.

Peneliti sebelumnya bertujuan mengidentifikasi dan mendeskripsikan bagaimana cara hoax bekerja. Dari hasil penelitian para ahli akan menghasilkan strategi dan formula baru untuk menghindari hoax [2].

Berita hoax sudah beredar sedari lama. Beberapa orang percaya akan berita tersebut walaupun sudah terbukti salah. Oleh karena itu, mendeteksi berita palsu bisa jadi sulit, apalagi tanpa badan pengawas di internet. Itu pertumbuhan kekhawatiran mengenai deteksi berita yang tidak dapat diandalkan baru-baru ini. Sulit bagi manusia untuk mendeteksi berita secara manual, bahkan dengan adanya semua topik yang ditampilkan di media sosial. Oleh karena itu, diperlukan cara yang efisien untuk membantu kita membedakan informasi palsu dari yang benar yang diposting di sosial media. Selain itu, untuk meningkatkan literasi digital terdapat strategi personal yang dapat dilakukan menurut Potter (2004:378). Yang pertama, kembangkan kesadaaran akurat yang akan memaparkan informasi dengan memilih sumber terpercaya dan resmi. Selanjutnya yang kedua, tingkatkan ilmu pengetahuan agar pemikiran kita lebih terbuka untuk tidak langsung percaya kepada suatu argumen tanpa ada bukti yang jelas.yang ketiga, bandingkan informasi yang kita peroleh dari suatu media dengan media lainnya agar mendapatkan sudut pandang yang berbeda. Kemudian yang keempat, perhatikan opini kita secara pribadi, apakah opini tersebut dapat dipertanggungjawabkan dengan segala sumber informasi yang dimiliki. Yang terakhir, tumbuhkan kebiasaan untuk memverifikasi kebenaran dan mengoreksi berita hoax yang beredar [3]. Akan tetapi, selain solusi tersebut terdapat solusi yang efisien adalah dengan mengklasifikasikan berita menggunakan algoritma pembelajaran mesin.

Di dalam penelitian sebelumnya yang dilakukan oleh Zeinab Shahbazi dan Cheol Byun [4], yang telah menyajikan kombinasi blockchain dan teknik pembelajaran mesin untuk memberikan solusi dan desain arsitektur berbasis kepercayaan terhadap berita bersama secara online. Derngan menerapkan teknik pembelajaran penguatan dan algoritma berbasis pembelajaran, untuk membuat arsitektur pengambilan keputusan yang kuat dan menggabungkannya dengan kerangka kerja blockchain, kontrak pintar, dan

DOI http://dx.doi.org/10.51213/jimp.v7i2.472

algoritma konsensus yang sangat cocokm untuk disesuaikan. Selain itu, pada penelitian yang dilakukan oleh Long Ying, dkk [5], Metode mereka dievaluasi pada dua set data dunia nyata yaitu WEIBO dan PHEME, dan hasil eksperimen menunjukkan strategi pendekatan MTMN yang diusulkan dapat mengungguli SOTA garis dasar.

Penelitian yang diusulkan berupa Implementasi Metode Svm, Logistic Regresion Dan Random Forest Clasifier Untuk Mengklasifikasi Fake News Di Twitter Dengan Mengimplementasikan 3 Metode ML sekaligus yaitu Implementasi Metode Svm, Logistic Regresion Dan Random Forest Clasifier.

II. METODOLOGI PENELITIAN

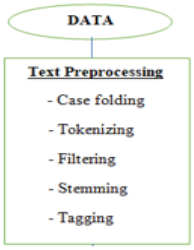
Dengan pertumbuhan besar-besaran konten media sosial di Internet, cara mengenali dan mendeteksi berita palsu menjadi semakin menantang. Para peneliti telah bekerja tentang deteksi berita palsu dan usulkan banyak metode berbeda yang secara kasar dapat ditinjau dari dua aspek: deteksi berita palsu modal tunggal (misalnya, teks atau gambar) dan deteksi berita palsu multi-modal. Dalam analisis modalitas tunggal, metode yang ada terutama mengekstrak fitur tekstual atau fitur visual dari konten teks atau informasi gambar postingan, yang telah dieksplorasi dalam berbagai deteksi berita palsu bekerja.

Pada penelitian ini menggunakan 3 metode yaitu SVM, Logistic Regresion dan Random Forest Clasifier. Datasets yang digunakan pada penelitian ini bersumber dari https://www.kaggle.com. Penelitian ini menerapkan beberapa tahapan untuk sampai pada kesimpulan penelitian. Tahapan-tahapan yang harus dilakukan adalah sebagai berikut, pengambilan (mining) dataset fakenews, prepocessing, klasifikasi dataset dan hasil klasifikasi dataset

III. HASIL DAN PEMBAHASAN

A. *Text Preprocessing*

Pra-pemrosesan data diperlukan sebelum mengekstraksi fitur. Data ini mungkin berisi karakter khusus, angka, dan ruang yang tidak perlu. Adapun proses tahapannya yaitu :



Gambar. 1 Tahapan Text Processing

B. *Klasifikasi*

Pada pembahasan selanjutnya kita akan membahas tentang klasifikasi yang mana kita menggunakan 3 metode yaitu, SVM, Logistic Regresion dan Random Forest Clasifier.

1) Support Vector Machine (SVM) merupakan sistem pembelajaran yang menggunakan ruang hipotesis yang berupa fungsi-fungsi linear didalam sebuah fitur yang memiliki dimensi tinggi dan dilatih dengan menggunakan algoritma pembelajaran berdasarkan teori optimasi *[6].*

```
#2. Support Vector Machine(SVM)
svm_model = SVC(kernel='linear')

#Fitting training set to the model
svm_model.fit(xv_train,y_train)

#Predicting the test set results based on the model
svm_y_pred = svm_model.predict(xv_test)

#Calculate the accuracy score of this model
score = accuracy_score(y_test,svm_y_pred)
print('Accuracy of SVM model is ', score)

Accuracy of SVM model is  0.9860418743768694
```

Gambar. 2 Hasil metode Support Vector Machine

Didapatlah kesimpulan dengan menggunakan metode SVM menghasilkan nilai akurasi sebesar 0,98.

2) Logistic Regresion, Menurut *(Hosmer dan Lemesshow, 2000)* Regresi Logistik adalah suatu metode analisis statistika untuk mendeskripsikan hubungan antara peubah respon (*dependent variable*) yang memiliki dua kategori atau lebih dengan satu atau lebih peubah penjelas (*independent variable*) berskala kategori atau interval *[7].*

```
#1. Logistic Regression - menggunakan ini karena paling cocok untuk klasifikasi biner
LR_model = LogisticRegression()

#Fitting training set to the model
LR_model.fit(xv_train,y_train)

#Predicting the test set results based on the model
lr_y_pred = LR_model.predict(xv_test)

#Calculate the accurracy of this model
score = accuracy_score(y_test,lr_y_pred)
print('Accuracy of LR model is ', score)

Accuracy of LR model is  0.9750747756729811
```

Gambar. 3 Hasil metode Logistic Regresion

Didapatlah kesimpulan dengan menggunakan metode Logistic Regression menghasilkan nilai akurasi sebesar 0,97.

3) Random Forest Clasifier, Menurut (Breiman, 2001) random Forest didefinisikan sebagai prinsip umum suatu ansambel acak dari suatu pohon keputusan [8].

```
[ ]  #3. Random Forest Classifier
     RFC_model = RandomForestClassifier(random_state=0)

     #Fitting training set to the model
     RFC_model.fit(xv_train, y_train)

     #Predicting the test set results based on the model
     rfc_y_pred = RFC_model.predict(xv_test)

     #Calculate the accuracy score of this model
     score = accuracy_score(y_test,rfc_y_pred)
     print('Accuracy of RFC model is ', score)

     Accuracy of RFC model is  0.9740777666999003
```

Gambar. 4 Hasil metode Random Forest Clasifier

Didapatlah kesimpulan dengan menggunakan metode random Forest Classifier menghasilkan nilai akurasi sebesar 0,97.

C. *Term of Matrix*

Pada penjelasan terakhir kita akan menghitung sebuah term menggunakan matriks.

DOI http://dx.doi.org/10.51213/jimp.v7i2.472

```
#Vectorization
vectorization = TfidfVectorizer()
xv_train = vectorization.fit_transform(x_train)
xv_test = vectorization.transform(x_test)

print(xv_train)
print(xv_test)
```

Gambar. 5 Barisan kode menghitung sebuah term menggunakan matrix

```
  (0, 9325)     0.01468771617844318
  (0, 5030)     0.010169524803284733
  (0, 11805)    0.01018538929046713
  (0, 23968)    0.009642133244048895
  (0, 13694)    0.02586377629914996
  (0, 25866)    0.019459069989260218
  (0, 22031)    0.011257050125580887
  (0, 4406)     0.01468771617844318
  (0, 24978)    0.014909316537537987
  (0, 15300)    0.009170144459456632
  (0, 19962)    0.01680159390413113
  (0, 2021)     0.023497631852905754
  (0, 6656)     0.015830395270339456
  (0, 4797)     0.014408273680626317
  (0, 10694)    0.019060794225997536
  (0, 7310)     0.01832413912431574
  (0, 2081)     0.03651977385021212
  (0, 21917)    0.04269116984798302
  (0, 28132)    0.01749874425307546
  (0, 22195)    0.017314067035870268
  (0, 18269)    0.027253632036356553
  (0, 19900)    0.03835323614364853
  (0, 18206)    0.012187305718234011
  (0, 23581)    0.030161263030315934
  (0, 6144)     0.029341257707297494
  :     :
```

```
(3005, 5538)  0.016647624631831406
(3005, 28084) 0.029473619139075834
(3005, 28584) 0.052167994522333505
(3005, 1106)  0.01315903180450025
(3005, 28813) 0.02728926392815848
(3005, 25898) 0.006962044146289029
(3005, 433)   0.01305653905151897
(3005, 14680) 0.0145852568338427
(3005, 5350)  0.00934815319885797
(3005, 22746) 0.009004889413927141
(3005, 6641)  0.02611307810303794
(3005, 14280) 0.007955656758130989
(3005, 749)   0.03478119610879882
(3005, 22171) 0.08073955115054242
(3005, 28046) 0.009054947086207132
(3005, 13572) 0.02248308176373447
(3005, 17396) 0.007428102479979547
(3005, 10042) 0.023185964327690302
(3005, 10267) 0.016010047920387994
(3005, 18573) 0.01340423445128166
(3005, 3602)  0.03857735085200095
(3005, 17412) 0.0071557010801967195
(3005, 26614) 0.006895817070097255
(3005, 20960) 0.009922950580554347
(3005, 26081) 0.010826559629354566
```

```
(0, 28886)    0.01760599518780924
(0, 28841)    0.03034985119453701
(0, 28813)    0.0056897101947820826
(0, 28812)    0.20180056797534365
(0, 28810)    0.11087411520208126
(0, 28659)    0.011397541758326519
(0, 28626)    0.06042549425405788
(0, 28623)    0.012347032465422689
(0, 28622)    0.01329698315341827
(0, 28614)    0.023445136417246104
(0, 28613)    0.0390767314805139
(0, 28602)    0.029891343898318785
(0, 28584)    0.04350733256220688
(0, 28487)    0.043745193435098945
(0, 28447)    0.02158268827527316
(0, 28291)    0.030742941727568234
(0, 28196)    0.019966892151627086
(0, 28177)    0.03195212431265121
(0, 28145)    0.015847217202435884
(0, 28132)    0.0123280019071955
(0, 28127)    0.02154947795921512
(0, 28046)    0.05286183961526466
(0, 28009)    0.01959763294194984
(0, 27971)    0.07519118239528301
(0, 27954)    0.01062331305945196
:     :
```

Gambar. 6 Hasil perhitungan sebuah term menggunakan matrix

IV. KESIMPULAN

Masalah berita palsu bukanlah hal baru, seperti disinformasi sudah lama beredar di surat kabar dan radio. Karena internet, berita bohong menyebar dengan cepat melalui media sosial dan blog. Penelitian ini menggunakan tiga algoritma ML yang berbeda yaitu Support Vector Machine (SVM), Logistic Regresion dan Random Forest Clasifier. Pengklasifikasian fakenews menggunakan teknik klasifikasi dengan data mining menggunakan 3 metode yaitu SVM, Logistic Regresion dan Random Forest Clasifier dapat disimpulkan menghasilkan nilai akurasi yang sangat baik yaitu sebagai berikut: akurasi Support Vector Machine (SVM) = 98%, Logistic Regresion = 97% dan Random Forest Clasifier = 97%. Sesuai dengan masing-masing nilai akurasi dapat disimpulkan pada kasus ini metode klasifikasi terbaik adalah metode Support Vector Machine (SVM).

UCAPAN TERIMA KASIH / ACKNOWLEDGMENT

Alhamdulillah penulis mengucapkan terimakasih atas segala rahmat Allah SWT dan semua pihak yang terlibat. Sehingga penulis dapat menyelesaikan jurnal yang berjudul "Studi Komparasi Metode Svm, Logistic Regresion Dan Random Forest Clasifier Untuk Mengklasifikasi Fake News Di Twitter".

REFERENCE

[1.] C. Juditha, "Interaksi Komunikasi Hoax di Media Sosial serta Antisipasinya," Jurnal Pekommas, vol. 3, pp. 31-44, 2018.

[2.] A. Justito, G. G. H. and N. Maharani, "Hoax, Reproduksi Dan Persebaran: Suatu Penelusuran Literatur," Jurnal Pengabdian

DOI http://dx.doi.org/10.51213/jimp.v7i2.472

Kepada Masyarakat , vol. 1, pp. 271-278, 2017.
[3.] A. R. Sabrina, “Literasi Digital Sebagai Upaya Preventif Menanggulangi Hoax,” Journal of Communication Studies, vol. 5, pp. 31-46.
[4.] Z. Shahbazi and C. Byun, “Fake Media Detection Based on Natural Language Processing and Blockchain Approaches,” IEEE Access, vol. 9, pp. 128443-128453, 2021.
[5.] L. Ying, H. Yu, J. Wang, Y. Ji and S. Qian, “Fake News Detection via Multi-Modal Topic Memory Network,” IEEE Access, vol. 9, pp. 132818-132828, 2021.
[6.] I. M. Parapat, M. T. Furqon and S. , “Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak,” Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 2, pp. 3163-3169, 2018.
[7.] R. Hendayana, “Penerapan Metode Regresi Logistik Dalam Menganalisis Adopsi Teknologi Pertanian,” Informatika Pertanian, vol. 22, pp. 1-9, 2013.
[8.] N. Wuryani and S. Agustiani, “Random Forest Classifieruntuk Deteksi Penderita COVID-19 berbasis Citra CT Scan,” Jurnal Teknik KomputerAMIK BSI, vol. 7, pp. 187-193, 2021.

DOI http://dx.doi.org/10.51213/jimp.v7i2.472