ANALISA DAN DETEKSI KONTEN HOAX PADA MEDIA BERITA INDONESIA MENGGUNAKAN MACHINE LEARNING

Munirul, Ula[1], Mulia Mahendra[2] Alvanof[3], Rahmat Triandi[4]
[1,2)]Teknik Informatika Universitas Malikussaleh Lhokseumawe
Jl. Cot Tgk Nie-Reulet, Aceh Utara, 141 Indonesia
email : muliamahendraalvanof@gmail.com, rahmattriandi83@gmail.com

ABSTRAK

Sekarang ini konten Hoax yang mengandung informasi tidak benar malah sering kali menjadi konsumsi massal pengguna internet. Hal ini merupakan sesuatu yang buruk karena dapat meningkatkan rasa tidak percaya terhadap berita dan informasi yang ada di internet hingga menimbulkan kebingungan pada masyarakat dalam menentukan informasi mana yang benar.

Dalam Penelitian ini, percobaan yang dilakukan bertujuan untuk memilih algoritma terbaik dalam membedakan berita hoax dan berita asli menggunakan metode text mining serta pendekatan dengan machine learning dan 150 artikel berbahasa Indonesia (50 artikel hoax dan 100 artikel asli) sebagai data yang akan digunakan.

Penelitian ini akan dimulai dengan tahap preprocessing teks yang terdiri dari tokenizing, case folding, filtering, stopword removal, stemming dan weighting TF-IDF menggunakan penggabungan fitur unigram dan bigram baru kemudian diolah menjadi teks klasifikasi. Hasil dari penelitian ini didapatkan kesimpulan bahwa algoritma Random Forest memiliki akurasi terbaik dalam mengklasifikasikan berita hoax dan berita asli dibandingkan dengan algoritma Multilayer Perceptron, Naïve Bayes,dan Support Vector Machine dengan nilai akurasi 75.37%.

Kata kunci : Klasifikasi, Berita, Hoax, Text mining, Machine learning

PENDAHULUAN

Kemajuan teknologi informasi memiliki berdampak pada semakin maraknya media berita online yang dapat diakses dengan mudah. Namun Seiring berkembangnya penyebaran berita online, kualitas berita yang disebarkan juga semakin berkurang. Berita yang tersebarluas tidak semuanya benar, termasuk didalamnya berita palsu atau hoax yang biasanya berakibat merugikan untuk piha tertentu.

Hoax adalah informasi atau berita yang mengandung hal-hal yang tidak pasti atau yang tidak berdasarkan fakta atas sesuatu yang benar terjadi. Hoax juga dapat diidentifikasi dengan beberapa ciri: berita datang dari sumber yang tidak jelas /tidak dipercaya. Gambar, foto atau video digunakan adalah hasil rekayasa, menggunakan kalimat provokatif, mengandung politik maupun ras. Penyebaran hoax di kalangan masyarakat dapat menyebabkan efek negatif, seperti kerusakan,kerugian, baik materiil dan psikologis, hilangnya kepercayaan masyarakat, dan sebagainya.

Dampak dari penyebaran berita hoax akan memiliki konsekuensi dan bahaya buruk untuk banyak pihak, yang mana hoax dapat menyebabkan kerugian dari berbagai aspek, baik waktu dan ekonomi, publik panik, memburuknya hubungan sosial dan sebagainya. Untuk menghindari dampak buruk ini, penelitian ini akan membantu untuk mengklasifikasikan berita sebelum terpengaruh oleh hoax atau bahkan sampai ikut menyebarkan hoax. Teknik yang digunakan dalam penelitian ini adalah dengan sistem klasifikasi teks menggunakan pendekatan berbasis machine learning. Algoritma yang digunakan adalah: Multilayer Perceptron (MLP), Naive Bayes (NB), Random Forest (RF), dan Support Vector Machine (SVM).

Penelitian serupa telah dilakukan sebelumnya oleh Rasywir dan Purwarianti[6], bedanya antara penelitian ini dan penelitian sebelumnya adalah dalam penggunaan algoritma dimana Penelitian Rasywir dan Purwarianti menggunakan tiga algoritma,yaitu algoritma Naïve Bayes, Support Vector Machine, dan Decision Tree, sementara penelitian ini akan membandingkan empat algoritma yaitu Multilayer Perceptron, Naïve Bayes, Support Vector Machine, and Random Forest.

TINJAUAN PUSTAKA

1. Berita

Berita adalah laporan informasi tentang peristiwa dan pendapat yang aktual, penting dan menarik untuk disampaikan kepada publik dalam bentuk surat kabar, radio dan media online. Salah satu syarat dari berita adalah bahwa berita tersebut harus didasarkan pada keadaan atau peristiwa yang benar-benar terjadi. Tetapi seiring dengan kemajuan teknologi yang semakin pesat, penyebaran berita juga semakin tidak jelas apakah kebenarannya sesuai dengan fakta atau hanya hoax belaka. Hoax adalah manipulasi berita yang disengaja dan bertujuan untuk memberikan informasi atau pemahaman yang salah. Hoax sering ditemukan di berita, baik melalui media cetak maupun media sosial. Tujuan dari hoax itu sendiri sangat beragam, mulai dari menyebarkan ujaran kebencian, menimbulkan kecemasan di masyarakat, memengaruhi persepsi masyarakat, dan sebagainya

2. Text Mining

Text mining adalah salah satu cabang dari data mining yang menganalisis data dalam bentuk teks. Text mining itu sendiri adalah suatu proses data mining berupa teks dengan sumber data yang biasanya diperoleh melalui dokumen, dan bertujuan untuk menemukankata-kata yang dapat mewakili isi dari dokumen.

3. Text Preprocessing

Text preprocessing adalah tahap awal dari data mining. Pada tahap ini, proses persiapan untuk dokumen dan data dilakukan agar dokumen / data siap diproses dan proses klasifikasi dapat dilakukan dengan benar. Adapun tahapan dalam Text Preprocessing, yaitu:

a. Tokenizing

adalah proses membagi ataumemotong kalimat menjadi kata-kata dengan menggunakan spasi, contohnya: "Text preprocessing merupakan tahap awal dari text mining ", akan dipecah menjadi: "Text", "preprocessing", "merupakan", "tahap", "awal", "dari", "text","mining".

b. Case folding

Case folding adalah tahap di mana semua karakter dalam teks diproses dan diubah menjadi huruf kecil. Misalnya: "TEKNOLOGI" akan diubah menjadi "teknologi", "Teks" akan dikonversi menjadi kata "teks", dan sebagainya.

c. Filtering

adalah tahap penghapusan tanda baca

d. Stopwords removing

Dalam tahap Stopwords removing kata yang bukan kata unik atau kata-kata yang sering muncul namun tidak penting akan dihapus. Contoh kata yang dimaksud adalah kata depan, kata sambung, kata keterangan dan kata pengganti, seperti: "yang", "ke", "di","sebuah", "pada", "oleh", "ini", "dari", dll.

e. Stemming

Proses Stemming adalah proses yang dilakukan untuk mendapatkan kata dasar dari sebuah kata dengan menghilangkan akhiran dan atau akhiran dari kata tersebut.

f. TF - IDF weighting

TF (Term Frequency) adalah frekuensi Munculnya suatu istilah dalam suatu dokumen. Semakin besar jumlahkemunculan istilah (TF tinggi) padadokumen, semakin besar "bobot" atau nilai kesesuaiannya. IDF (Inverse Document Frequency) adalah perhitungan bagaimana istilah tersebut didistribusikan dalam koleksi dokumen terkait. IDF menunjukkanhubungan antara ketersediaan istilah dalam semua dokumen.

4. Tipe Algoritma

Algoritma yang digunakan dalam penelitian ini meliputi :

a. Multilayer Perceptron (MLP)

Metode Multilayer Perceptron adalah salah satu topologi artificial neural networks yang paling umum digunakan dimana perceptron akan dikoneksikan sehingga membuat multiple layers. MLP terdiri atas input, hidden dan output layer dan karena terdiri atas beberapa layer, proses perhitungan pada algoritma ini akan melalui beberapa tahapan pula.

$$hidden_k = \frac{1}{1 + exp^{-hidden\_Net\,j}}$$

dimana nilai *hidden_Netj* dapat diambil dari

$$hidden_{Net_j} = \Sigma_i w_{jk} \times mf_{ij}(y_{ij}) + bias_k$$

Keterangan:

w_{jk} = "Bobot" neuron j layer sebelumnya ke neuron k pada hidden layer

$mf_{ij}(y_{ij})$ = "Derajat Anggota" atribut i ke kelas j.

$bias_k$ = "Bias" neuron k.

$$Output_l = \frac{1}{1 + exp^{-output\_Net\,k}}$$

dimana nilai $output\_Net_k$ dapat diambil dari persamaan

$$Output_{Net_k} = \Sigma_j w_{kl} \times hidden_k + bias_l$$

Keterangan::

w_{kl} = Bobot" neuron k layer sebelumnya ke neuron l pada hidden layer

$bias_l$ = Bias neuron l.

b.Naïve Bayes

Algoritma yang dikembangkan ini telah banyak digunakan sebagai metode untuk mengklasifikasikan teks yang memiliki "chance" sederhana. Algoritma ini menggunakan probabilitas dan statistik untuk memprediksi probabilitas masa depan berdasarkan pengalaman yang ada.

Persamaan teorema Naïve Bayes theorem daat ditulis sebagai berikut :

$$P(X|Y) = \frac{P(Y|X) \times P(X)}{P(Y)}$$

Keterangan

P(X|Y) = Kemungkinan X Berdasarkan Kondisi Y.
P(Y|X) = Kemungkinan Y Perdasarkan Hipotesis X.
P(X) = Kemungkinan X.
P(Y) = Kemungkinan Y.

c.Support Vector Machine

Metode SVM asalah model universal dari machine learning yang menggunakan 'fungsi pembatasan linear' sebagai dasar.

Persamaan untuk perhitungan pada algoritma SVM untuk data yang mungkin belum terkelompokan secara benar (Asiyah&Fithriasari, 2016) adalah:

$$min\frac{1}{2}||w||^2 + C\sum_{i=1}^{\lambda}\xi i$$

Keterangan:

W = Hyperplane parameter dicari
C = Parameter data error masukan pengguna.
Λ = Jumlah Partisi/data.
I = Nilai Awal

d.Random Forest

Metode Random Forest adalah pengembangan dari metoder CART, Random Forest adalah sebuah metode yang dapat meningkatkan akurasi karena metode ini akan membuat node-node acak untuk setiap node.

$$Entropy(Y) = -\Sigma_i P(C|\mathrm{Y})\log_2 P(C|\mathrm{Y})$$

Keterangan:

Y = Junlah kasus.

P(C|Y) = Perbandingan nilaibY terhadap kelas C

METODE PENELITIAN

1. Sumber Data

Data yang digunakan dalam penelitian ini diambil dari artikel berita media online yang sudah memiliki label hoax dan non-hoax. dengan 50 artikel berlabel hoax dan 100 artikel berlabel non-hoax. Jadi total keseluruhan artikel yang digunakan sebagai data pada penelitian ini adalah 150 artikel.

2. Tahapan Analisa

Dalam penelitian ini, beberapa tahapan analisa yang digunakan yaitu:

a. Pengumpulan berita dari situs berita online diantaranya :
viva.co.id, detik.com, kompas.com, liputan6.com, metrotvnews.com, beritasatu.com, cnnindonesia.com, idntimes.com, republika.co.id, prokal.co, cekfakta.com, jpnn.com, okezone.com, sindonews.com, solopos.com, tempo.co, merdeka.com, tribunnews.com

b. data berita yang telah dikumpulkan kemudian disimpan dalam format CSV dan kemudian akan melewati tahap *text preprocessing*, tahap pertama adalah *tokenizing* dimana semua kalimat dari data berita akan dipisahkan berdasarkan tanda spasi.

c. Kemudian pada proses *folding case*, semua huruf diubah menjadi huruf kecil, kemudian kata-kata yang memiliki arti yang sama akan dinormaisasi menjadi satu kata yang sama.

d. Pada proses *filtering*, karakter selain huruf dan angka akan dihapuskan, misalnya tanda baca (.).

e. Pada tahap *stopwords removing*, kata yang tidak penting dan tidak unik akan dihapuskan.

f. Selanjutnya pada proses *stemming*, awalan dan atau akhiran kata dihilangkan, sehingga didapat kata dasarnnya saja.
g. Tahap terakhir pada text preprocessinf adalah menghitung bobot tiap kata menggunakan pembobotan TF-IDF dengan penggunaan kombinasi fitur unigram dan bigram.
h. Data kemudian dibagi 2, training data (80%) dan test data (20%). pemilihan training dan test data akan dilakukan secara acak oleh program.
i. Lakukan Klasifikasin teka menggunakan algoritma machine learing, seperti Multilayer
   Perceptron, Support Vector Machine, Naïve Bayes, dan Random Forest.
j. Bandingkan hasil antara algoritma berdasarkan tingkat akutasi, presisi, recall dan score F-1 dari tiap algoritma.

HASIL DAN PEMBAHASAN

1. Precision

Nilai Precision adalah tingkat akurasi berdasarkan informasi yang diminta pengguna dengan jawaban yanh diberikan sistem.
Perhitungan akurasi adalah sebagai berikut (Manning et al, 2009):

$$Precision = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})}$$

dapat pula ditulis sebagai berikut:

$$Precision = \frac{\text{Relevant data found}}{\text{All data found}}$$

Keterangan:
TP = True Positive adalah jumlah data relevan yang secara benar diklasifikasian sebagai kecocokan oleh sistem.
FP = False Positive adalah jumlah data yang tidak relevan, namjn diklasifikasikan sebagai kecocokan oleh sistem

Adapun nilai Presisi yang didapat dari tiap algoritma dapat dilihat pada tabel dibawah

Tabel 1 Nilai Precision

Algoritma	Precision
Multilayer Perceptron	69.12%
Naïve Bayes	79.79%
Support Vector Machine	70.16%
Random Forest.	79.34%

2. Recall

Recall adalah nilai kesuksesan sistem dalam menemukan kembali sebuah informasi, dengan kata lain Recall menunjukan selengkap apa hasil relevan yang ditampulkan sistem.

$$Recall = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}$$

Keterangan:

TP = True Positive adalah jumlah data relevan yang secara benar diklasifikasian sebagai kecocokan oleh sistem.

FN = False Negative adalah jumlah data relevan, namun tidak diklasifikasikan sebagai kecocokan data oleh sistem.

Adapun nilai Recall yang didapat dari tiap algoritma dapat dilihat pada tabel dibawah

Tabel 2 Nilai Recall

Algoritma	Recall
Multilayer Perceptron	67.16%
Naïve Bayes	66.86%
Support Vector Machine	62.31%
Random Forest.	73.82%

3. Skor F-1

Skor F-1 adalah perhitungan nilai performa yang dilakukan untuk melihat hasil yabg didapat dari proses klasifikasi berdasarkan nilai Presisi dan recall yang telah didapat sebelumnya.

$$F1 = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

Perbandingan Skor F-1 dari keempat algortima dapat dilihat pada tabel berikut.

Tabel 3 Nilai F-1

Algoritma	Scor F-1
Multilayer Perceptron	67.09%
Naïve Bayes	67.26%
Support Vector Machine	71.08%
Random Forest.	74.24%

4. **Accuracy**

Level Accuracy adalah level kedekatan antara nilai prediksi dengan nilai aktual. Perhitungan nilai akurasibdapat ditulis sebagai berikut(Manning et al, 2008):

$$Accuracy = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

Perhitungan akurasi pada persamaan diatas juga dapat ditulis sebagai berikut :

$$Accuracy = \frac{\text{data yang diklasifikasikan dengan benar}}{\text{total data diuji}}$$

Keterangan :
TP = True Positive adalah jumlah data relevan yang secara benar diklasifikasian sebagai kecocokan oleh sistem.
TN = True Negatif adalah jumlah data tidak relevan yang diklasifikasikan sebagai tidak cocok dengan benar oleh sisitem.
FP = False Positive adalah jumlah data yang tidak relevan, namjn diklasifikasikan sebagai kecocokan oleh sistem
FN = False Negative adalah jumlah data relevan, namun tidak diklasifikasikan sebagai kecocokan data oleh sistem.

Nilai Akurasi dari keemlat algoritma dapat dilihat pada tabel dibawah ini.

Tabel 4 Nilai Accuracy

Algoritma	Acuraccy
Multilayer Perceptron	68.63%
Naïve Bayes	74.51%
Support Vector Machine	60.00%
Random Forest.	75.37%.

Secara keseluruhan perbandingan perbandingan antara keempat algoritma: Multilayer Perceptron, Support Vector Machine, Naïve Bayes, dan Random Forest berdasarkan presisi, recall, skor F-1 dan akurasi dapat dilihat pada tabel dibawah.

Tabel 5 Perbandingan Algoritma

Algoritma	P	R	F-1	A
Multilayer Perceptron	69.12%	67.16%	67.09%	68.63%
Naïve Bayes	79.79%	66.86%	67.26%	74.51%
Support Vector Machine	70.16%	62.31%	71.08%	60.00%
Random Forest.	79.34%	73.82%	74.24%	**75.37%.**

KESIMPULAN

Pada studi ini, sistem klasifikasi berita hoax di Indonesia telah dibuat menggunakan machine learing dengan 4 macam algoritma, antara lain : Multilayer Perceptron, Naïve Bayes, Support Vector Machine, dan Random Forest. dengan total 150 artikel ssbagai data yang terdiri atas 50 artikel hoax dan 100 artikel non hoax. Penelitian ini juga dilakukan dengan melalui proses *tokenizing, case folding, normalization, filtering, stopwords removing, stemming, and TF-IDF weighting* menggunanakn fitur *unigram* dan *bigram*.

Dapat disimpulkan bahwa hasil klasifikasi terbaik didapat dengan algoritma Random Forest jika dibandingkan dengan algoritma Multilayer Perceptron algorithm, Naïve Bayes, dan Support Vector Machine dengan akurasi sebesar 75.37%.

Berdasarkan perbandinga pada tabel 6, penulis lebih memilih untuk menggunakan algoritma dengan akurasi paling tinggi, karena secara rata-rata jika tingkat akurasinya tinggi, maka tingkat presisi dan recallnya juga akan tinggi pula.

DAFTAR PUSTAKA

[1] Asiyah. S. N., &Fithriasari. K., (2016): *Klasifikasi Berita Online Menggunakan Metode Support Vector Machine dan K-Nearest Neighbor*, Surabaya: Jurnal Sains dan Seni ITS.

[2] Binarwati. L., Mukhlash. I., &Soetrisno. S., (2017): *Implementasi Algoritma Genetika untuk Optimalisasi Random Forest dalam Proses Klasifikasi Penerimaan Tenaga Kerja Baru :Studi Kasus PT.XYZ*, Surabaya: Jurnal Sains dan Seni ITS.

[3] Juditha. C., (2018): *Interaksi Simbolik dalam Komunitas Virtual Anti Hoak suntuk Mengurangi Penyebaran Hoaks*, Jakarta: Jurnal PIKOM, vol. 19, no. 1, Kementerian Komunikasi dan Informatika RI.

[4] Monohevita. L., (2017): *Stop Menyebarkan Hoax, Depok: Universitas Indonesia.*

[5] Negnevitsky. M., (2005): Artificial Intelligence: A Guide to Intelligent System (2nd Ed), Harlow: Pearson Education.

[6] Rasywir. E., & Purwarianti. A., (2015): *Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin.* Bandung: Jurnal Cybermatika, vol. 3, no. 2.