EKSPERIMEN *NAÏVE BAYES* PADA DETEKSI BERITA *HOAX* BERBAHASA INDONESIA

NAÏVE BAYES'S EXPERIMENT ON HOAX NEWS DETECTION IN *INDONESIAN LANGUAGE*

Faisal Rahutomo[1], Inggrid Yanuar Risca Pratiwi[2], Diana Mayangsari Ramadhani[3]
[1,2,3]Politeknik Negeri Malang
Jalan Soekarno Hatta No. 9 Malang
email : faisal@polinema.ac.id [1], inggrid_yanuar@polinema.ac.id [2],
diana_mayangsari@polinema.ac.id [3]
(Diterima: 25-02-2019; Direvisi: 18-04-2019; Disetujui terbit: 25-6-2019)

Abstrak

Website dan blog terkenal sebagai media penayangan berita dalam berbagai bidang seperti penayangan berita. Validitas artikel berita dapat bersifat valid ataupun palsu. Berita palsu disebut juga dengan hoax news. Tujuan pembuatan berita hoax ini adalah untuk membujuk, memanipulasi, mempengaruhi pembaca berita untuk melakukan hal-hal yang bertentangan atau mencegah tindakan yang benar. Pada penelitian ini mengusulkan untuk melakukan eksperimen klasifikasi naïve Bayes pada deteksi berita hoax berbahasa Indonesia. Penelitian ini menggunakan dataset sendiri sebanyak 600 berita antara berita valid dan hoax. Tiga pembaca berita melakukan klasifikasi manual. Sistem yang dibangun dapat mengklasifikasikan berita daring berbahasa Indonesia dengan fitur term frequency dan algoritma klasifikasi naïve Bayes dengan menggunakan komponen library PHP-ML atau PHP-Machine Learning. Berdasarkan hasil uji coba secara statis, sistem ini menghasilkan akurasi sebesar 82,6% dan pengujian secara dinamis persentase kesesuaian dengan sistem 68,33%. Dataset disediakan terbuka sehingga dapat diakses oleh peneliti lainnya dan dapat dijadikan baseline pada penelitian-penelitian berikutnya.
Kata kunci : deteksi berita hoax, klasifikasi naïve Bayes.

Abstract

Website and blog are popular as a media to spread news. The validity of an article of news's can either be valid or fake. A fake article of news is usually called a hoax news article. The purpose of making hoax news is to persuade, manipulate, affect to people to do something that contradicts or prevents the right action. A hoax news usually used threats or misleading information to make them believe things that are not real. This research proposes an experiment using naïve Bayes to detect hoax news in Bahasa Indonesia. In this research, we use our own dataset consisting of a total of 600 valid and hoax articles. We asked three reviewers to conduct manual classification for our dataset. Final tagging was obtained by adopting the maximum score from the three reviewers. In our experiment, we show that naïve Bayes can classify Indonesian online news articles with term frequency feature using the PHP-ML library component's. We obtained an accuracy is 82.6% with static testing and 68.33% with dynamic testing. We give free access to the dataset so the future research can replicate, comparing the result and make a baseline testing.
***Keywords** : Hoax News Detection, Naïve Bayes Classifier.*

PENDAHULUAN

Kecepatan beredarnya berita seiring dengan kemajuan teknologi. Kecepatan media cetak seperti surat kabar dan majalah kalah bersaing dengan media elektronik seperti televisi dan internet

(Ishwara 2005). Internet berkembang sebagai media informasi yang terkenal pada berbagai bidang seperti pencarian berita, gambar, ulasan berita dan produk, layanan masyarakat, film dan sebagainya. Semua disajikan dalam berbagi sumber seperti artikel berita, media sosial, dan blog. Website dan blog terkenal sebagai media penayangan berita. Berita memiliki sudut pandang positif, negatif ataupun netral.

Artikel berita yang tersebar di situs-situs website dapat diarahkan oleh penulis berita ke arah tipuan atau sesuatu yang tidak benar. Informasi palsu dan menyesatkan ini berbahaya karena dapat menyesatkan persepsi manusia dengan menyampaikan informasi yang tidak benar sebagai sebuah kebenaran sehingga sangat memungkinkan untuk membawa dampak negatif pada pemikiran manusia (Ishak, Chen and Yong 2012), (Rasywir and Purwarianti 2016), dan (Shu, et al. 2017).

Tujuan pembuatan berita *hoax* ini adalah untuk membujuk, memanipulasi, mempengaruhi pembaca berita untuk melakukan hal-hal yang bertentangan atau mencegah tindakan yang sudah benar. Biasanya menggunakan ancaman, menyesatkan atau membuat mereka mempercayai hal-hal yang tidak nyata (Hernandez, et al. 2002) (Vuković, Pripužić and Belani 2009). Beberapa *hoax* diciptakan dengan berbagai cara. Bahkan mereka dapat memperoleh data pribadi dengan meyakinkan para korban bahwa data tersebut diperlukan untuk tujuan resmi (Ishak, Chen and Yong 2012).

Melalui internet, berita dapat disebarluaskan tanpa penyeleksian terlebih dahulu atau penelusuran kebenarannya. Penerima berita perlu mengklasifikasikan berita tersebut sebelum membagikannya kepada orang lain atau mempercayai berita tersebut. Hal ini akan berdampak kurang baik serta dapat menimbulkan beragam persepsi.

Adapun tujuan dari penelitian ini adalah melakukan eksperimen pada algoritma klasifikasi *naïve Bayes* untuk deteksi berita *hoax*, mengklasifikasikan berita dalam jaringan (daring) bahasa Indonesia yaitu valid dan *hoax* dengan menggunakan *term frequency* (TF) sebagai fitur pada eksperimen ini.

Pada penelitian-penelitian sebelumnya, telah membahas tentang deteksi *hoax* pada berbagai bidang. Seperti deteksi virus (Hernandez, et al. 2002), deteksi email palsu (Vuković, Pripužić and Belani 2009), deteksi penipuan pada gaya penulisan daring (Sadia, Brennan and Greenstadt 2012), klasifikasi ulasan asli dan palsu pada toko daring (Banerjee, Chua and Jung-Jae 2015) dan klasifikasi berita *hoax* berbasis pembelajaran mesin (Rasywir and Purwarianti 2016).

Pada penelitian lainnya, membahas tiga tipe artikel berita palsu yaitu pemalsuan serius, *hoax* dalam skala besar, dan tipuan lucu (Rubin, Chen and Conroy 2015). Adapula yang melakukan klasifikasi postingan Facebook ke dalam *hoax* dan non-*hoax* (Tacchini, et al. 2017).

Penelitian ini melakukan eksperimen *naïve Bayes* pada deteksi berita *hoax* pada berita berbahasa Indonesia.

LANDASAN TEORI

Pada penelitian-penelitian sebelumnya yang dilakukan oleh (Hernandez, et al. 2002) tentang mendeteksi virus *hoax*. Selanjutnya terdapat penelitian seputar *hoax* yang dilakukan oleh (Vuković, Pripužić and Belani 2009) yaitu tentang klasifikasi email *hoax* dengan menggunakan *feed forward neural network* dan

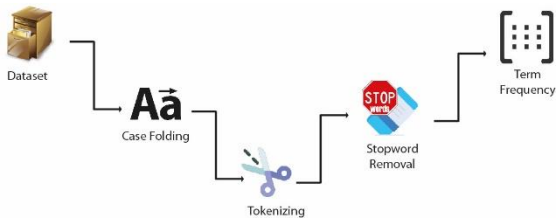
advanced text processing. Deteksi penipuan pada gaya penulisan daring menggunakan algoritma *k-Nearest Neighbor*, *naïve Bayes*, J48 *Decision Tree*, regresi logistik dan SVM dengan kernel RBF yang dilakukan dengan menggunakan perangkat WEKA (Sadia, Brennan and Greenstadt 2012). Klasifikasi ulasan asli dan palsu pada toko daring dengan sepuluh algoritma dilakukan karena semakin tumbuhnya metode belanja secara daring membuat calon pembeli cenderung untuk menelusuri ulasan produk terlebih dahulu sebelum melakukan pembelian. Namun tidak semua ulasan selalu asli atau autentik (Banerjee, Chua and Jung-Jae 2015). Pada penelitian lainnya, membahas tiga tipe artikel berita palsu yaitu pemalsuan serius, *hoax* dalam skala besar, dan tipuan lucu (Rubin, Chen and Conroy 2015). Pada klasifikasi berita *hoax* berbasis pembelajaran mesin memilih *naïve Bayes*, SVM, dan C4.5 sebagai algoritmanya (Rasywir and Purwarianti 2016). Adapula penelitian yang mengklasifikasikan postingan Facebook dalam klasifikasi *hoax* dan non-*hoax* (Tacchini, et al. 2017).

Web Crawler

Salah satu dasar ekstraksi web adalah memproses informasi, baik berupa informasi tekstual (*textual*) maupun intekstual (*hypertextual*) dari halaman web. Informasi tersebut kemudian disimpan di lokasi penyimpanan dan diberi indeks berdasarkan kata kunci (*keyword*) (Sari, Wicaksana and Burhan 2011). Prinsip kerja *crawler* adalah melakukan pemindaian (*scanning*) terhadap keberada-an *hyperlink* yang terhubung ke halaman lain. *Web crawler* adalah proses pencarian atau perayapan halaman atau halaman informasi dari sebuah halaman. Tidak hanya merangkak, tapi juga merayap web serta mengambil informasi dari halaman. Fungsi utama *crawler* web adalah mencari atau merayapi informasi dari sebuah halaman (Muzad and Rahutomo 2016).

Praproses Teks

Sebelum melakukan klasifikasi, data uji dan latih akan melalui tahap praproses yang terdiri dari *case folding*, *tokenizing*, *stopword removal*, dan *term frequency* untuk mendapatkan matriks *term* dan frekuensi dari masing-masing *term* dalam sebuah dokumen.



Gambar 1 Tahap Praproses
sumber : peneliti

Case Folding

Salah satu strategi umum yang dilakukan dalam proses teks adalah *case folding* yaitu mengubah semua huruf besar atau kapital menjadi huruf kecil (Pratiwi, Rahutomo and Asmara 2017). Contoh kalimat “Bapak Budi bekerja di BUMN yaitu PT. PLN”. Hasil pada tahapan case folding adalah “bapak budi bekerja di bumn yaitu pt pln”. Pada *case folding* tidak hanya mengubah huruf besar atau kapital menjadi huruf kecil tetapi juga menghilangkan tanda baca (Pratiwi, Rahutomo and Asmara 2017). Contoh kata C.A.T yang adalah sebuah akronim. Pada tahap ini, normalisasi akan melakukan penghilangan tanda baca sekaligus mengubah huruf besar menjadi huruf kecil. C.A.T → CAT. CAT → cat.

Tokenizing

Tokenizing adalah memotong kalimat menjadi beberapa bagian kata sekaligus

membuang karakter tertentu, seperti tanda baca berdasarkan spasi (Pratiwi, Rahutomo and Asmara 2017). Sebagai contoh (Christopher, Raghavan and Schütze 2009):

- Masukan : *friends, romans, countrymen, lend me your ears*;
- Luaran : *friends | romans | countrymen | lend | me | your | ears*

Token secara luas disebut sebagai istilah atau kata-kata. Pada saat proses *tokenizing*, secara bersamaan menghilang-kan karakter tertentu seperti tanda baca, angka, dan karakter selain huruf alphabet, karena karakter-karakter tersebut dianggap sebagai pemisah kata (*delimiter*) dan tidak memiliki pengaruh terhadap proses teks. Tetapi pada beberapa kasus pemrosesan teks, angka tidak dihilangkan karena masih dianggap penting. *Token* adalah turunan dari urutan karakter dalam beberapa dokumen tertentu yang dikelompokkan bersama sebagai unit semantik yang berguna selanjutnya untuk diproses.

Stopword Removal

Sering kali beberapa kata yang umum digunakan tampaknya bernilai sedikit dalam membantu memilih dokumen yang sesuai dengan kebutuhan pengguna (Christopher, Raghavan and Schütze 2009). Kata-kata ini disebut dengan *stopword.* Cara umum untuk menentukan daftar *stopword* adalah mengurutkan berdasarkan frekuensi kemunculan istilah kata pada kumpulan dokumen. Kemudian diambil kata-kata atau istilah yang paling sering muncul dan dikumpulkan menjadi *stopword list.* Anggota dari *stopword list* ini akan dibuang dari teks pada pengindeksan kata saat pelatihan dan pengujian. Sebagai contoh *stopword list* berbahasa Indonesia (Rasywir and Purwarianti 2016) diantaranya : "yang", "ini", "dari", "ke", "di", "dari". Penggunaan *stopword list* secara signifikan dapat mengurangi jumlah kumpulan kata yang harus disimpan dalam sebuah basis data.

Term Frequency

Term frequency (*TF*) merupakan salah satu fitur dalam proses teks. Perhitungan *TF* dilakukan dengan menghitung jumlah atau frekuensi kemunculan setiap kata dalam sebuah dokumen. Hasil dari perhitungan *TF* disimpan dalam basis data sistem untuk digunakan pada proses lebih lanjut.

Klasifikasi *Naïve Bayes*

Klasifikasi *Naïve Bayes* adalah klasifikasi probabilitas yang sederhana yang menerapkan teorema *Bayes*. Metode ini mudah, kuat dan berdiri sendiri atau independen. Jika "d" adalah vektor masukan fitur dan "c" adalah label kelas, *naïve Bayes* menuliskan dengan P(d|c). Notasi ini merupakan probabilitas kelas "c" didapatkan setelah fitur-fitur "d" ditemukan. Notasi P(d|c) disebut probabilitas akhir dan P(c) disebut probablitias awal. Pada proses pelatihan harus dilakukan pembelajaran probabilitas akhir berdasarkan informasi yang didapat dari data latih.

Naïve Bayes adalah algoritma untuk mengklasifikasikan kemungkinan yang berarti untuk dokumen "d", dari semua kelas "c". $\widehat{\ }$ adalah perkiraan tentang kelas yang benar (Thabtah, et al. 2009)

$$\widehat{Xnb} = \text{argmax}\, \mathrm{P(c)} \prod \mathrm{P(d|c)}$$

Formula *Naïve Bayes* untuk klasifikasi adalah :

$$\mathrm{P(d|c)} = \frac{\mathrm{Tct} + \lambda}{\mathrm{Nc} + \lambda \mathrm{V}}$$

diketahui :

T_{ct} : Berapa kali kata "d" muncul dalam kelas "c"

λ : Nilai konstanta positif. Biasanya 1 untuk menghindari probabilitas nol

N_c : Jumlah kata dalam kelas "c"

V : Jumlah seluruh kata

Akurasi

Akurasi dihitung dari jumlah prediksi yang benar dibagi total jumlah dokumen yang diprediksi (Pratiwi, Rahutomo and Asmara 2017).

$$\text{Akurasi} = \frac{tp+tn}{tp+fp+fn+tn}$$

Diketahui maksud dari variabel-variabel di atas yaitu "*tp*" untuk *true positive, "tn*" untuk *true negative, "fp*" untuk *false positive*, dan "*fn*" adalah *false negative.*

Precision

Nilai *precision* diperoleh dari jumlah prediksi yang benar (tp) dibagi jumlah yang didapat (tp+fp) (Pratiwi, Rahutomo and Asmara 2017).

$$Precision = \frac{tp}{tp+fp}$$

Recall

Nilai *recall* diperoleh dari jumlah prediksi yang benar dibagi jumlah seluruh data yang ada (Pratiwi, Rahutomo and Asmara 2017).

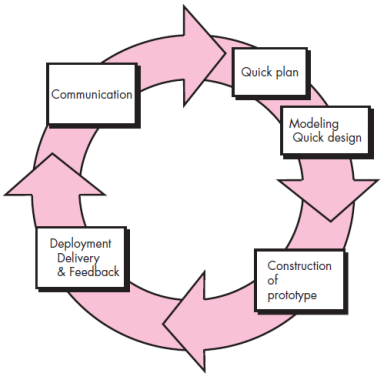
$$Recall = \frac{tp}{tp+fn}$$

METODE PENELITIAN

Metode penelitian yang digunakan peneliti dalam melakukan ekperimen ini adalah mengumpulkan dataset, melakukan manual *tagging* pada setiap data berita yang telah dikumpulkan dengan *voting* diantara responden, dan merancang bangun sistem yang dibuat. Dalam merancang bangun sistem ini, peneliti menggunakan metode *prototyping*.

Model Perancangan Sistem

Dalam merancangan sistem eksperimen *naïve Bayes* pada deteksi berita *hoax* berbahasa Indonesia ini, peneliti menggunakan model pengembangan perangkat lunak *prototyping.* Dengan menggunakan metode ini, pengguna berinterakasi dengan pihak pengembang selama proses pembuatan sistem (Muzad and Rahutomo 2016).



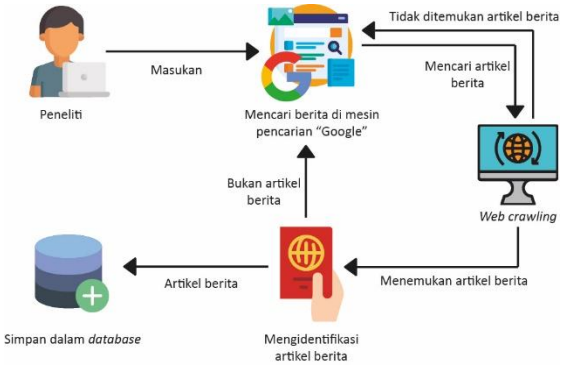
Gambar 2. Tahapan Metode *Prototyping*
sumber : Software and Engineering A Practitioner's Approach oleh Roger S. Pressman, 2010

Dimulai dengan komunikasi, pengembang bertemu dengan pengguna untuk mendefinisikan tujuan keseluruhan suatu sistem, mengidentifikasi kebutuhan sistem, dan deskripsi sistem secara garis besar. Pada tahap *qiuck plan*, pengembang membuat representasi dari aspek-aspek dari sistem yang akan dihasilkan, seperti desain antar muka sistem. Pada tahap ini pengembang fokus untuk membangun sebuah *prototype. Prototype* yang dibangun akan digunakan dan dievaluasi oleh pengguna dan pengguna akan memberikan umpan balik yang digunakan untuk fokus pembuatan sistem. Tahap ini akan berulang dikomunikasikan dua arah antar pengguna dan pengembang untuk lebih memahami apa yang perlu dilakukan.

Idealnya *prototyping* berfungsi sebagai sebuah mekanisme untuk mengidentifikasi persyaratan pembuatan perangkat lunak. Jika *prototype* dibuat dan diterapkan dalam proses pembuatan sistem, memungkinkan pekerjaan cepat selesai.

Teknik Pengumpulan dan Sumber Data

Pada penelitian ini, peneliti mengumpulkan *dataset* sendiri secara manual melalui mesin pencarian Google. Berikut di bawah ini gambaran proses pengumpulan *dataset.*



Gambar 3. Tahap Pengumpulan *Dataset*
sumber : peneliti

Peneliti mengumpulkan *dataset* di portal-portal berita secara manual melalui internet, dimana Google sebagai mesin pencari. Peneliti memasukkan kata kunci berita pada Google. Google akan mencari berita sesuai dengan kata kunci yang dimasukkan.

Peneliti juga mengidentifikasi apakah tautan berita yang dikeluarkan oleh google termasuk dalam berita atau opini masyarakat. Berdasarkan penelitian sebelumnya (Muzad and Rahutomo 2016) terdapat situs berita yang divalidasi, diantaranya adalah kompas.com, merdeka.com, tribunnews.com, tempo.co, viva.co.id, republika.co.id, dan metrotvnews.com. Sehingga tolak ukur kevalidan sebuah berita berdasarkan artikel berita yang diterbitkan pada situs-situ berita tersebut.

Jika yang ditemukan tersebut adalah sebuah berita, peneliti menyimpan isi berita dalam format teks dan pada basis data serta selanjutnya akan dilakukan penandaan manual berita tersebut. Jika bukan sebuah berita, akan kembali ke tahap awal.

Manual *Voting Tagging*

Dataset yang terkumpul tidak memiliki inisialisasi tanda. Sebagai bahan pembelajaran sistem, sistem membutuhkan inisialisasi awal tanda dari masing-masing berita. Untuk mendapatkan tanda kelompok berita tersebut, peneliti menggunakan proses *voting* dari responden.

Luaran hasil klasifikasi dari sistem ini adalah berita valid dan hoax, sehingga nilai tanda yang digunakan dalam voting adalah berita valid dan hoax serta jumlah responden harus berjumlah ganjil. Hal ini dikarenakan untuk menghindari nilai sama atau seri atau *draw* di antara responden sehingga dapat diambil nilai terbanyak dari setiap pilihan tanda artikel beritanya. Sebagai contoh ditampilkan proses ini di Tabel 1.

Tabel 1. Manual *Voting Tagging* Berita

Situs Berita	*Reviewer*			Hasil
	1	2	3	
http://tekno.liputan6.com/read/2110066/iphone-6-plus-melengkung-jadi-olok-olokan-di-internet	1	1	2	V
http://www.tribunnews.com/tribunners/2016/02/24/menelusuri-kebenaran-lele-pemicu-kanker	1	2	1	V
http://www.madiunpos.com/2015/11/10/pembakaran-reog-ribuan-seniman-ponorogo-turun-jalan-tuntut-permintaan-maaf-660052	2	2	1	X
http://www.viva.co.id/berita/nasional/976681-dpr-pelemparan-alquran-di-mako-brimob-hoaks	1	1	1	V
https://tekno.kompas.com/read/2018/04/17/14333587/viral-kabar-facebook-bakal-diblokir-24-april-hoaks-atau-fakta	1	2	1	V

Keterangan :
V : Valid
X : *Hoax*

HASIL PENELITIAN DAN PEMBAHASAN

Dataset

Berdasarkan proses pengumpulan *dataset* yang telah dilakukan oleh peneliti, terdapat 600 *dataset* berita yang terdiri dari 12 kata kunci berita, masing-masing kata kunci memiliki 10 artikel berita. 12 kata kunci tersebut diantaranya adalah (1) Ikan lele mengandung sel kanker. (2) Tusuk jari dengan jarum membantu pasien stroke. (3) Iphone 6 mudah melengkung. (4) Reog Ponorogo dibakar di Fhilipina. (5) Peserta aksi 212 tidak bisa masuk masjid Istiqlal. (6) Sikat gigi dari bulu babi. (7) Permen dot berbahaya. (8) Pokemon "GO" sama dengan "Aku Yahudi". (9) Pelemparan Al-Qur'an di Mako Brimob. (10) CTO Traveloka *walk out* saat Gubernur Anies pidato. (11) Facebook akan tutup di Indonesia. (12) Gaji Presiden naik.

Dataset berita yang telah dikumpulkan oleh peneliti disediakan secara terbuka sehingga dapat diakses oleh peneliti lainnya pada situs

https://data.mendeley.com/datasets/p3hfgr5j3m/1. Diharapkan *dataset* ini dapat menjadi *baseline* pada penelitian-penelitian selanjutnya.

Tabel 2 Hasil Manual *Tagging*

Tagging	Jumlah Berita
Valid	372
Hoax	228
Total	600

Manual *Voting* Tanda Kelompok Berita

Proses *voting* berita yang dilakukan oleh 3 orang responden menghasilkan jumlah berita untuk masing-masing kelompok. Data ditunjukkan oleh Tabel 2.

Tahapan Praproses

Setiap *dataset* berita latih dan uji melalui tahap praproses. Hal ini dilakukan untuk mendapatkan model yang dikehendaki. Seperti yang sudah dipaparkan pada bab sebelumnya, tahapan praproses tersebut adalah *case folding, tokenizing, stopword removal,* dan *term frequency.*

Sebagai contoh, terdapat 10 berita yang dimiliki. 7 berita sebagai data latih dan 3 berita sebagai data uji. Untuk mendapatkan model matriks kata, diperlukan proses latihan pada data latih untuk mendapatkan hasil model klasifikasi. Sehingga pada tahap uji, data uji akan mencocokan dengan model yang telah dimiliki.

1) Case Folding

Case folding adalah merubah huruf besar menjadi huruf kecil dan menghapus tanda baca. Sebagai contoh setelah melalui tahap *case folding*, maka akan diperoleh hasil pada Tabel 4 dari teks asli di Tabel 3.

Tabel 3. Data Latih

No	Berita	*Tagging*
1	iPhone 6 Plus dikeluhkan sejumlah pengguna memiliki kelemahan mudah melengkung bila terlalu lama disimpan di saku celana.	*Hoax*
2	iPhone 6 Plus mudah melengkung	*Hoax*
3	iPhone 6 Plus melengkung jadi bahan ejekan di dunia maya	*Hoax*
4	iPhone yang melengkung merupakan kejadian langka	Valid
5	Fitur iPhone baru memiliki sisipan baja atau titanium untuk memperkuat keliling body ponsel	Valid
6	Hasil uji coba menunjukkan bahwa iPhone 6 Plus tidaklah serentan yang dikeluhkan belakangan ini	Valid
7	Melengkungnya body iPhone 6 Plus bukanlah suatu kekurangan, melainkan memang salah satu fitur khusus yang sengaja ditambahkan pada smartphone ini	Valid

Tabel 4. Hasil *Case Folding*

No	Berita
1	iphone 6 plus dikeluhkan sejumlah pengguna memiliki kelemahan mudah melengkung bila terlalu lama disimpan di saku celana.
2	iphone 6 plus mudah melengkung
3	iphone 6 plus melengkung jadi bahan ejekan di dunia maya
4	iphone yang melengkung merupakan kejadian langka
5	fitur iphone baru memiliki sisipan baja atau titanium untuk memperkuat keliling body ponsel
6	hasil uji coba menunjukkan bahwa iphone 6 plus tidaklah serentan yang dikeluhkan belakangan ini
7	melengkungnya body iphone 6 plus bukanlah suatu kekurangan, melainkan memang salah satu fitur khusus yang sengaja ditambahkan pada smartphone ini

2) Tokenizing

Proses *tokenizing* adalah memecah kalimat menjadi token. Dari contoh diatas, didapatkan hasil *tokenizing* pada tabel 5.

Tabel 5. Hasil *Tokenizing*

No	Berita
1	iphone | 6 | plus | dikeluhkan | sejumlah | pengguna | memiliki | kelemahan | mudah | melengkung | bila | terlalu | lama | disimpan | di | saku | celana
2	iphone | 6 | plus | mudah | melengkung
3	iphone | 6 | plus | melengkung | jadi | bahan | ejekan | di | dunia | maya
4	iphone | yang | melengkung | merupakan | kejadian | langka
5	fitur | iphone | baru | memiliki | sisipan baja | atau | titanium | untuk | memperkuat | keliling | body | ponsel
6	hasil | uji | coba | menunjukkan | bahwa | iphone | 6 | plus | tidaklah | serentan | yang | dikeluhkan | belakangan | ini
7	melengkungnya | body | iphone | 6 | plus | bukanlah | suatu | kekurangan | melainkan | memang | salah | satu | fitur | khusus | yang | sengaja | ditambahkan | pada | smartphone | ini

Tabel 6. Hasil *Stopword Removal*

No	Berita
1	iphone | 6 | plus | dikeluhkan | pengguna | memiliki | kelemahan | mudah | melengkung | disimpan | saku | celana
2	iphone | 6 | plus | mudah | melengkung
3	iphone | 6 | plus | melengkung | bahan | ejekan | dunia | maya
4	iphone | melengkung | kejadian | langka
5	fitur | iphone | memiliki | sisipan baja | titanium | memperkuat | keliling | body | ponsel
6	hasil | uji | coba | menunjukkan | iphone | 6 | plus | serentan | dikeluhkan
7	melengkungnya | body | iphone | 6 | plus | kekurangan | fitur | khusus | sengaja | smartphone

3) Stopword Removal

Proses *stopword removal* kata pada *dataset* disesuaikan dengan koleksi *stopword list* yang dimiliki. Pada penelitian ini menggunakan *stopword list* dari Talla F. Z. Kata yang dihapus dari *dataset* adalah kata yang terdapat dalam daftar *stopword*. Berikut hasil

penghapusan *stopword* dijelaskan pada tabel 6.

Fitur *TF*

Setelah *dataset* melalui tahap pra proses, *dataset* akan dihitung frekuensi kemunculan kata pada dokumen atau yang disebut dengan *term frequency* (*TF*). Didapatkan hasil pada Tabel 7.

Tabel 7. Hasil *TF*

	Kata	*TF Hoax*	*TF* Valid
1	iphone	3	4
2	6	3	2
3	plus	3	2
4	dikeluhkan	1	0
5	pengguna	1	0
6	memiliki	1	1
7	kelemahan	1	0
8	mudah	2	0
9	melengkung	3	1
10	disimpan	1	0
11	saku	1	0
12	celana	1	0
13	bahan	1	0
14	ejekan	1	0
15	dunia	1	0
16	maya	1	0
17	kejadian	0	1
18	langka	0	1
19	fitur	0	2
20	sisipan	0	1
21	baja	0	1
22	titanium	0	1
23	memperkuat	0	1
24	keliling	0	1
25	body	0	1
26	ponsel	0	1
27	hasil	0	1
28	uji	0	1
29	coba	0	1
30	menunjukkan	0	1
31	serentan	0	1
32	dikeluhkan	0	1
33	melengkungnya	0	1
34	kekurangan	0	1
35	khusus	0	1
36	sengaja	0	1
37	smartphone	0	1

Probabilitas Kata

Setelah mendapatkan frekuensi kata pada seluruh dokumen latih, pada tahap ini menghitung nilai probabilitas $P(c_i)$. Masing-masing kelompok dihitung nilai probabilitasnya berdasarkan banyak dokumen dalam kategori per seluruh dokumen.

$$P(hoax) = \frac{\text{jumlah dokumen } hoax}{\text{jumlah seluruh dokumen}}$$

$$= \frac{3}{7} = 0{,}429$$

$$P(\text{valid}) = \frac{\text{jumlah dokumen valid}}{\text{jumlah seluruh dokumen}}$$

$$= \frac{4}{7} = 0{,}571$$

Berdasarkan persamaan p(d|*class*) didapatkan hasil pada Tabel 8.

Tabel 8. Nilai Probabilitas

No	Kata	P(d|*hoax*)	P(d|valid)
1	iphone	0.0755	0.0862
2	6	0.0755	0.0517
3	plus	0.0755	0.0517
4	dikeluhkan	0.0377	0.0172
5	pengguna	0.0377	0.0172
6	memiliki	0.0377	0.0345
7	kelemahan	0.0377	0.0172
8	mudah	0.0566	0.0172
9	melengkung	0.0755	0.0345
10	disimpan	0.0377	0.0172
11	saku	0.0377	0.0172
12	celana	0.0377	0.0172
13	bahan	0.0377	0.0172
14	ejekan	0.0377	0.0172
15	dunia	0.0377	0.0172
16	maya	0.0377	0.0172
17	kejadian	0.0189	0.0345
18	langka	0.0189	0.0345
19	fitur	0.0189	0.0517
20	sisipan	0.0189	0.0345
21	baja	0.0189	0.0345
22	titanium	0.0189	0.0345
23	memperkuat	0.0189	0.0345
24	keliling	0.0189	0.0345
25	body	0.0189	0.0345
26	ponsel	0.0189	0.0345
27	hasil	0.0189	0.0345
28	uji	0.0189	0.0345
29	coba	0.0189	0.0345
30	menunjukkan	0.0189	0.0345

No	Kata	P(d|*hoax*)	P(d|valid)
31	serentan	0.0189	0.0345
32	dikeluhkan	0.0189	0.0345
33	melengkungnya	0.0189	0.0345
34	kekurangan	0.0189	0.0345
35	khusus	0.0189	0.0345
36	sengaja	0.0189	0.0345
37	smartphone	0.0189	0.0345

Sampel Uji

Pada sub bab ini akan dicontohkan proses uji *dataset* dengan *naïve Bayes*. Terdapat 3 *data* berita yang digunakan sebagai data uji. Data uji disajikan pada Tabel 9. Setiap data uji akan melalui tahap pra proses seperti yang telah dilakukan pula pada data latih yaitu *case folding*, *tokenizing*, dan *stopword removal* sehingga menghasilkan matriks kata. Matriks kata ini akan dicocokkan dengan kumpulan kata yang telah didapatkan sebelumnya. Jika terdapat kata yang dicari maka akan diambil bobot P(w|valid) dan P(w|*hoax*). Sebaliknya, bila tidak terdapat kata uji pada kumpulan kata maka kata tersebut tidak dihitung atau dihiraukan (Christopher, Raghavan and Schütze 2009). Kata-kata yang sudah mendapatkan bobot P(w|valid) dan P(w|*hoax*) akan dihitung nilai probabilitas-nya atau kemungkinannya dengan *naïve Bayes*.

Tabel 9. Data Uji

No	Berita	*Tanda*
8	Apple dinilai melakukan kesalahan besar memilih material campuran metal-aluminium alloy yang tidak terlalu solid sebagai bahan dasar rangka iPhone 6 Plus.	*Hoax*
9	Pihak iphone membantah bahwa produknya dapat melengkung	Valid
10	Banyak pengguna Twitter, Path, serta jejaring sosial berbagai foto lainnya yang mengejek kelemahan iPhone 6 Plus itu.	*Hoax*

Berdasarkan pencocokan matriks kata data uji dengan model yang telah dilatih, didapatkan bobot P(w|*hoax*) dan P(w|valid) yang selanjutnya menghitung nilai probabilitas dengan *naïve Bayes*.

Tabel 10 menunjukkan pada berita ke-9, kesimpulan yang dihasilkan oleh perhitungan dengan *naïve Bayes* tidak sama dengan hasil tanda kelompok manual. Ini dikarenakan kata-kata pada data uji yang dicocokkan dengan matriks kata berita yang telah didapatkan sebelumnya hanya terdapat dua kata yang ditemukan. Sehingga bila dihitung probabilitasnya menghasilkan kesimpulan bahwa berita tersebut termasuk berita *hoax.* Semakin banyak *dataset* yang dilatih maka akan menghasilkan matriks kata yang lebih banyak.

Tabel 10. Hasil Uji dan Klasifikasi

Berita 8		
Kata	**P(w|*hoax*)**	**P(w|valid)**
bahan	0.0377	0.0172
iphone	0.0755	0.0862
6	0.0755	0.0517
plus	0.0755	0.0517
Argmax(*hoax*)		6.95×10^{-6}
Argmax(*valid*)		2.27×10^{-6}
Kesimpulan		***Hoax***
Berita 9		
Kata	**P(w|*hoax*)**	**P(w|valid)**
iphone	0.0755	0.0862
melengkung	0.0755	0.0345
Argmax(*hoax*)		0.002441133
Argmax(*valid*)		0.001698658
Kesimpulan		***Hoax***
Berita 10		
Kata	**P(w|*hoax*)**	**P(w|valid)**
pengguna	0.0377	0.0172
kelemahan	0.0377	0.0172
iphone	0.0755	0.0862
6	0.0755	0.0517
Argmax(*hoax*)		3.47×10^{-6}
Argmax(*valid*)		7.57×10^{-7}
Kesimpulan		***Hoax***

Pengujian Statis

Dari 600 berita di dalam *dataset*, dibagi menjadi tiga proses latih dan uji dengan persentase data latih dan uji sebesar 60%:40%, 70%:30%, dan 80%:20%. Perbandingan ini digunakan untuk melakukan eksperimen *naïve Bayes* pada deteksi berita *hoax*. Semakin banyak data latih yang digunakan, akan menghasilkan akurasi yang lebih besar pula atau berbanding lurus. Karena semakin banyak data yang dilatih, semakin tepat pula hasil ujinya.

Tabel 11. Data Latih dan Uji

Dataset	Perbandingan Berita (%)		Jumlah Berita	
	Latih	Uji	Latih	Uji
600	60	40	360	240
	70	30	420	180
	80	20	480	120

Proses perhitungan probabilitas klasifikasi dengan metode *naïve Bayes* di dalam penelitian ini menggunakan *library* PHP - *machine learning* (PHP-ml). *Library* ini bersifat *open source* sehingga *developer* dapat menggunakannya sesuai dengan kebutuhan.

Dari hasil latih dan uji pada masing-masing perbandingan data, didapatkan nilai akurasi, *precision* dan *recall* berturut-turut sebagaimana ditunjukkan pada Tabel 12, 13, dan 14.

Dari 600 data berita diambil persentase perbandingan data latih dan uji masing-masing 60%:40%, 70%:30% dan 80%:20%. Masing-masing persentase dilakukan tiga kali waktu uji sehingga didapatkan nilai rata-rata akurasi 82,3%, 82,7% dan 83%. Pada persentase data latih dan uji 80%:20%, menghasilkan rata-rata akurasi paling besar dibanding persentase data lainnya. Metode *naïve Bayes* dapat mengklasifikasikan berita daring berbahasa Indonesia dengan akurasi rata-rata 82,6% untuk 600 data berita.

Nilai *precision* yang didapatkan seperti pada Tabel 13. Dari 600 data berita diambil persentase perbandingan data latih dan uji masing-masing 60%:40%, 70%:30% dan 80%:20%. Masing-masing persentase dilakukan tiga kali waktu uji sehingga didapatkan nilai rata-rata *precision hoax* adalah 81%, 80% dan 72,5% dan untuk nilai rata-rata *precision* alid adalah 83,7%, 84,6% dan 89%.

Nilai *recall* yang didapatkan ditunjukkan pada Tabel 14. Dari 600 data berita diambil persentase perbandingan data latih dan uji masing-masing 60%:40%, 70%:30% dan 80%:20%. Masing-masing persentase dilakukan tiga kali waktu uji sehingga didapatkan nilai rata-rata *recall hoax* adalah 67,5%, 76,5% dan 80,8% dan untuk nilai rata-rata *recall* valid adalah 90,6%, 87,5% dan 84,3%.

Tabel 12. Nilai Akurasi 600 *Dataset*

Waktu	Latih : Uji	Akurasi (%)	Rata-Rata(%)
1		82	
2	60 : 40	82	82,3
3		83	
1		81,3	
2	70 : 30	81,8	82,7
3		85	
1		81,8	
2	80 : 20	83,4	83
3		84	

Tabel 13. Nilai *Precision* 600 *Dataset*

Waktu	Latih : Uji (%)	*Precision*		Rata -Rata(%)	
		Hoax	Valid	*Hoax*	Valid
1		92	79,5		
2	60 : 40	78	84	81	83,7
3		73	87,8		
1		76,5	83,8		
2	70 : 30	76,5	86	80	84,6
3		87	84		
1		74,5	87		
2	80 : 20	74	88	72,5	89
3		69	92		

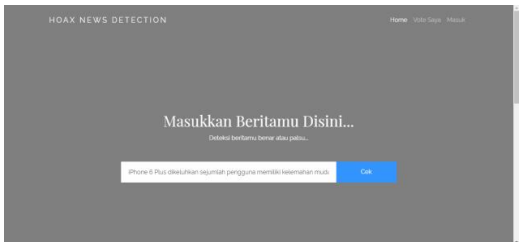
Tabel 14. Nilai *Recall* 600 *Dataset*

Waktu	Latih : Uji *(%)*	*Recall*		Rata-Rata (%)	
		Hoax	Valid	*Hoax*	Valid
1		55	97		
2	60 : 40	72,7	88	67,5	90,6
3		75	87		
1		74	86,8		
2	70 : 30	83,7	82	76,5	87,5
3		72	93,9		
1		74	82		
2	80 : 20	78	86,2	80,8	84,3
3		90,6	84,8		

Pengujian Dinamis

Pada pengujian dinamis ini sistem mendapatkan inputan berita dari pengguna. Pengguna dapat menekan tombol "cek" yang disediakan pada tampilan dan sistem akan mengeluarkan hasil probabilitas berita, apakah berita tersebut termasuk valid atau *hoax.*

Sebagai contoh, pengguna memasukkan berita "*iPhone 6 Plus dikeluhkan sejumlah pengguna memiliki kelemahan mudah melengkung bila terlalu lama disimpan di saku celana. Apple dinilai melakukan kesalahan besar dengan memilih material campuran metal-aluminium alloy yang tidak terlalu solid sebagai bahan dasar rangka iPhone 6 Plus*." pada sistem. Berita akan melalui tahapan praproses hingga penghitungan nilai probabilitas berita. Gambar 4 menunjukkan gambar masukan berita dari pengguna.



Gambar 4. Halaman Deteksi Berita
sumber : peneliti

Gambar 5. Hasil Cek Berita
sumber : peneliti

Sistem akan memanggil model yang telah dimiliki dari proses latih dan menghitung nilai probabilitas dengan algoritma *naïve Bayes.* Sehingga menghasilkan luaran bahwa berita tersebut adalah valid, seperti ditunjukkan pada Gambar 5.

Pada pengujian dinamis ini dilakukan proses uji pada 12 kata kunci berita yang sama dengan *dataset* dan mengambil masing-masing 5 artikel berita selain dari kumpulan berita *dataset*. Sehingga terdapat 60 artikel berita yang diujicobakan. Nilai kesesuaian hasil sistem dengan penandaan manual ditunjukkan oleh Tabel 15.

Tabel 15. Data Uji Dinamis

Total *dataset* berita uji	60 berita
Total hasil sesuai sistem dengan *voting*	41 berita
Total hasil tidak sesuai sistem dengan *voting*	19 berita
Persentasi tidak sesuai	31.67 %
Persentasi sesuai	68.33 %

PENUTUP

Kesimpulan

Berdasarkan hasil penelitian dan pengujian yang telah dilakukan dapat ditarik kesimpulan sebagai berikut :

1. Algoritma *naïve Bayes* dapat digunakan sebagai algoritma klasifikasi berita daring berbahasa Indonesia dengan dua probabilitas yaitu berita valid dan *hoax*.

2. Metode *naïve bayes* ini memiliki kelebihan dapat bekerja baik pada dataset yang jumlahnya sedikit, mudah dibuat, cepat dalam proses penghitungan. Sedangkan kekurangan dalam metode ini adalah tidak berlaku jika nilai probabilitasnya atau kemungkinannya adalah 0 (nol). Apabila nol maka kemungkinan prediksinya bernilai 0 juga. Serta mengasumsikan variabel bebas.
3. Pengujian statis dilakukan terhadap 600 berita yang menghasilkan rata-rata akurasi sebesar 82,6%.
4. Pengujian dinamis dilakukan dengan memasukkan isi berita pada sistem. Dari 60 berita yang diuji, 41 berita menghasilkan klasifikasi berita sama dengan tanda manual dan 19 berita menghasilkan klasifikasi yang berbeda dengan tanda manual. Persentase hasil sesuai yaitu 68,33% dan tidak sesuai yaitu 31,67%.
5. Kontribusi pada penelitian ini adalah pembuatan *dataset* berita berbahasa Indonesia yang telah memiliki penandaan awal berdasarkan beberapa responden serta *dataset* ini disediakan secara terbuka oleh peneliti.

Saran

Saran yang diberikan dari hasil penelitian ini untuk pengembangan sistem adalah sebagai berikut :

1. Pengembangan *dataset* dapat diperkaya lagi. Sehingga aplikasi dapat berjalan dengan maksimal.
2. Aplikasi ini dapat dikembangkan dengan mengumpulkan *dataset* secara *real-time* dengan proses *crawling* dari situs berita.
3. Uji coba dengan pemilihan tahapan praproses, pemilihan fitur, dan metode yang digunakan berbeda dari yang sudah dilakukan pada penlitian ini dan mengombinasikannya dapat menghasilkan klasifikasi yang berbeda dan mungkin lebih baik, mencapai nilai akurasi yang tinggi.
4. Penandaan manual kelompok berita dapat dikaji kembali agar mendapatkan penilaian yang tepat karena akan mempengaruhi proses pembelajaran aplikasi.

UCAPAN TERIMA KASIH

Penelitian ini masuk di dalam program penelitian swadana inovatif Politeknik Negeri Malang dengan dana DIPA Politeknik Negeri Malang Tahun 2018 dengan Nomer: DIPA 042.01.2.401004/ 2018 tanggal 05 Desember 2017 dengan Surat Perjanjian No: 6332/PL2.1/HK/ 2018. Para penulis mengucapkan terima kasih kepada Politeknik Negeri Malang, khususnya jajaran direktur, UPT P2M, pimpinan jurusan Teknologi Informasi, dan Teknik Elektro.

DAFTAR PUSTAKA

Afroz, Sadia, Michael Brennan, and Rachel Greenstadt. "Detecting Hoxes, Frauds, and Deception in Writing Style Daring." *IEEE Symposium on Security and Privacy*. IEEE, 2012. 461-475.

Aldwairi, Monther, and Ali Alwahedi. "Detecting Fake News in Social Media Networks." *The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks.* Elsevier, 2018. 215-222.

Banerjee, Snehasish, Alton YK Chua, and Kim Jung-Jae. "Using Supervised Learning To Classify Authentic and Fake Online Reviews." *Proceedings of*

the 9th International Conference on Ubiquitous Information Management and Communication. ACM, 2015. 88.

Christopher, Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval.* Cambridge: Cambridge University Press, 2009.

Hernandez, J. C., C. J. Hernandez, J. M. Sierra, and A. Ribagorda. "A first Step Towards Automatic Hoax Detection." *36th Annual 2002 International Carnahan Conference on Security Technology.* IEEE, 2002. 102-114.

Ishak, Adzlan, Y.Y. Chen, and Suet-Peng Yong. "Distance-base Hoax Detection System." *2012 International Conference on Computer Science & Information Science.* IEEE, 2012. 215-220.

Ishwara, Luwi. *Catatan - Catatan Jurnalisme Dasar.* Jakarta: Buku Kompas, 2005.

Muzad, Aad Miqdad Muadz, and Faisal Rahutomo. "Korpus Berita Daring Bahasa Indonesia dengan Depth First Focused Crawling." *Seminar Nasional Terapan Riset Inovatif Vol. 2 No. 1*, 2016.

Natali, Ruchansky, Sungyong Seo, and Yan Liu. "CSI: A Hybrid Deep Model for Fake News Detection." *CIKM '17.* Singapore: ACM, 2017. 797-806.

Pratiwi, Inggrid Yanuar Risca, Faisal Rahutomo, and Rosa Andrie Asmara. "Study of Hoax News Detection Using Naive Bayes Classifier in Indonesian Language." *2017 11th International Conference on Information & Communication Technology and System.* Surabaya: IEEE, 2017. 73-78.

Rasywir, Errissya, and Ayu Purwarianti. "Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin." *Jurnal Cybermatika 3.2.*, 2016.

Rubin, Victoria L, Yimin Chen, and Niall J. Conroy. "Deception Detection For News: Three Types of Fakes." *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community.* American Society for Information Science, 2015. 83.

Ruchansky, Natali, Sungyong Seo, and Yan Liu. "CSI: A Hybrid Deep Model for Fake News Detection." *CIKM '17.* Singapore: ACM, 2017. 797-806.

Sadia, Afroz, Michael Brennan, and Rachel Greenstadt. "Detecting Hoxes, Frauds, and Deception in Writing Style Online." *IEEE Symposium on Security and Privacy.* IEEE, 2012. 461-475.

Sari, Riri Fitri, Adi Wicaksana, and Burhan. *Teknik Ekstraksi Informasi di Web.* Yogyakarta: Andi, 2011.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Haun Liu. "Fake News Detection on Social Media : A Data Mining Perspective." *SIGKDD Explorations*, 2017: 22-36.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. "Fake News Detection on Social Media : A Data Mining Perspective." (ACM) 19, no. 1 (2017).

Tacchini, Eugenio, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. "Some Llike It Hoax: Automated Fake News Detection in Social Networks." *Proceedings of the Second Workshop on Data Science for Social Good (SoGood).* Macedonia: arXiv preprint arXiv:1704.07506, 2017.

Thabtah, Fadi, M. Eljinini, M. Zamzeer, and W. Hadi. "Naïve Bayesian Based

on Chi Square to Categorize Arabic data." *The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies.* Egypt: ibimapublishing, 2009. 4-6.

Vuković, Marin, Krešimir Pripužić, and Hrvoje Belani. "An Intelligent Automatic Hoax Detection System." *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems.* Berlin: Springer, 2009. 318-325.

Zhang, Xichen, and Ali A. Ghorbani. "An Overview of Online Fake News: Characterization, Detection, and Discussion." *Information Processing and Management*, 2019: 1-26.