Perbandingan Pembobotan Fitur TF-IDF dan TF-ABS Dalam Klasifikasi Berita Online Menggunakan *Support Vector Machine* (SVM)

1st Iklima Apriani
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
iklimmaa@student.telkomuniversity.ac.id

2nd Yuliant Sibaroni
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
yuliant@telkomuniversity.ac.id

3rd Irma Palupi
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
irmapalupi@telkomuniversity.ac.id

Abstrak —Berita terjadi karena adanya informasi atau kabar yang berhubungan dengan fakta dan sedang terjadi untuk kemudian disampaikan kepada masyarakat. Seiring dengan perkembangan teknologi kini penyebaran informasi dilakukan melalui media sosial yaitu *website* yang bisa diakses dengan media dekstop ataupun handphone. Pemilihan berita untuk dimasukan pada kategori tertentu jika dilakukan oleh manusia bisa menyebabkan *human eror,* terlebih berita yang dipakai sangat banyak bisa menyebabkan kurang efisien. Maka dari itu, sistem klasifikasi otomatis akan menjadi solusi pada permasalahan ini. Dalam klasifikasi, *fitur extraction* merupakan proses dasar dalam kategorisasi yang penting untuk dilakukan dan diketahui. Fitur tersebut kemudian akan direpresentasikan kedalam bentuk vektor, nilai vektor diperoleh dari pembobotan kata. Penelitian ini membandingkan pembobotan *Term Frequency - Inverse Document Frequency* (TF.IDF) dan *Term Frequency Absolute* (TF.ABS) yang dikombinasikan dengan *fitur extraction* unigram dengan metode klasifikasi *Support Vector Machine* (SVM). Dari hasil penelitian menunjukkan pembobotan TF-IDF mendapat akurasi sebesar 96,63% dengan hasil dengan hasil *f1-score* mendapat 97,06%. Sedangkan pembobotan TF-ABS mendapat akurasi sebesar 89,66% dengan hasil *f1-score* 96,63%. Dengan menggunakan pembobotan TF-IDF dapat menaikkan akurasi sebesar 6,97%. daripada mengunakan TF-ABS.

Kata kunci—berita, klasifikasi, *support vector machine*, TF-IDF, TF-ABS

I. PENDAHULUAN

Teknologi telah banyak membawa perubahan, salah satunya teknologi informasi yang menjadi bagian tidak terpisahkan dari kehidupan manusia, seolah melekat seperti pakaian. Dengan adanya teknologi, kehidupan manusia sangat terbantu dalam berbagai hal seperti berkomunikasi jarak jauh, mencari informasi yang tidak bisa dijangkau hingga terciptanya teknologi baru. Teknologi informasi merupakan sebuah bantuan yang diciptakan untuk membantu, mengubah, menyimpan, berkomunikasi dan menyebarkan informasi secara luas ke khalayak umum. Sedangkan teknologi komunikasi merupakan sebuah alat bantu untuk memproses dan mengirim data dari satu perangkat ke perangkat lain [1]. Selain menggunakan koran dan televisi, kini penyebaran informasi bisa melalui situs web (*website*) yang bisa diakses pada media desktop dan handphone kapan pun dan dimana pun.

Berita merupakan sebuah informasi atau kabar yang berhubungan dengan fakta dan sedang terjadi untuk kemudian disampaikan kepada masyarakat [2]. Berita yang termuat dalam situs web (*website*) dinamakan artikel. Banyak media seperti kompas, kumparan, detik.com, BBC Word beralih tempat dari koran ke situs web agar masyarakat bisa mengakses dan mengetahui informasi terbaru. Ada banyak jenis kategori pada artikel yang dimuat, meliputi berita kesehatan, politik, olahraga, hiburan, hingga teknologi terbaru.

Dalam beberapa tahun terakhir, banyaknya artikel yang diunggah menyebabkan masalah pada pengkategorian topik berita, sehingga diperlukannya pengklasifikasian topik berita agar tidak menyebabkan *overload* informasi untuk kemudian pembaca mudah mengakses berita [3]. Klasifikasi dilakukan manual dengan data yang besar membutuhkan waktu lama serta memiliki risiko *human error*, pengguna perlu memikirkan suatu artikel untuk dimasukkan pada kategori tertentu [3]. Maka dari itu perlu dilakukan pengklasifikasian otomatis untuk mengorganisasikan informasi berupa artikel tadi dan pencari informasi bisa paham berdasarkan kategorinya, serta bisa mempermudah pengolahan dan penggunaannya sesuai kebutuhan yang diinginkan.

Klasifikasi merupakan sebuah proses membangun suatu model yang mengkategorikan suatu objek sesuai dengan kelompoknya [4]. Tujuan klasifikasi yaitu mengelompokkan data yang mana data tersebut memiliki target ataupun kelas yang telah ditentukan. Penelitian yang melakukan klasifikasi menggunakan berita online dilakukan oleh siti [1] dengan data sebanyak 500 artikel berita. Penelitiannya membandingkan metode SVM dan KNN, perolehan akurasi SVM kernel *Polynomial* unggul mendapat 93.2% daripada akurasi KNN mendapat 60%.

Dalam proses klasifikasi mengenal istilah *fitur extraction*, *fitur extraction* merupakan proses dasar dalam kategorisasi yang penting untuk diketahui. Adapun pada penelitian [1] memiliki kekurangan yang tidak melibatkan proses *fitur extraction* tersebut. Fitur penting diketahui karena menjadi proses dasar yang mencerminkan informasi menganai konten dan konteknya [5]. Dalam *machine learning* suatu fitur akan direpresentasikan kedalam bentuk vektor, nilai vektor tersebut diperoleh dari pembobotan kata [6]. Penggunaan *fitur extraction* yang umum digunakan yaitu Unigram, Bigram dan Trigram.

Selain pembobotan TF.IDF yang umum dipakai pada [1] penelitian lain telah melakukan pembobotan berbeda, penelitian [8] melakukan perbandingan pembobotan TF.ABS dan TF.CHI menggunakan metode SVM. Dari hasil penelitiannya diperoleh TF.ABS dan TF.CHI memiliki hasil yang sama-sama baik dengan perolehan akurasi yang sama yaitu 95.87%. Penelitian lainnya [9] membandingkan TF.ABS, TF.CHI2, TF.RF, dan TF.IDF menggunakan sebanyak 360 data dengan tujuan penelitian yaitu untuk mengetahui akurasi optimal dikombinasikan dengan *Decision Tree*. Dari hasil penelitian tersebut akurasi tertinggi diperoleh menggunakan pembobotan TF.ABS sebesar 82.22% dibandingkan TF.CHI2 mendapat akurasi sebesar 80.83%, TF.RF 65.56%, dan TF.IDF 50.56%. Metode TF.ABS melakukan kinerja pembobotan dengan melihat kemunculan kata dan kemungkinan kata yang tidak muncul dalam dokumen [8]**.** Adapun pada penelitian ini pembobotan TF-ABS dipilih sebagai pembanding karena pada penelitian [8][9] untuk meningkatkan pembobotan mendapat kesimpulan bahwa TF-ABS lebih baik daripada metode pembobotan lainnya.

Support Vector Machine banyak digunakan untuk mengklasfikasikan berita dalam klasifikasi teks, pada penelitian [10] membandingkan tiga metode berbeda yaitu SVM, *Neural Network* dan *Naive Bayes*. Klasifikasi mengenai berita Nepali ini menggunakan sebanyak 4.964 data. Metode SVM mendapat klasifikasi yang unggul, ini ditunjukkan dalam perolehan akurasi yaitu SVM Kernel RBF mendapat akurasi sebesar 74.65%, SVM Linier mendapat 74,62%, NN mendapat 72.99% dan NB hanya mendapat 68.31%. Penelitian lain [11] melakukan klasifikasi *automatic* multilabel untuk artikel berbahasa Indonesia, penelitiannya berfokus mengkomparasikan beberapa seleksi fitur, pembobotan fitur, pendekatan multilabel dan metode klasifikasi, dengan tujuan klasifikasi untuk mengurangi dimensi fitur. Dari hasil percobaan didapat SVM memperoleh akurasi 85.13%. Berdasarkan penelitian [10][11] metode SVM mendapat klasifikasi terbaik karena dapat diimplementasikan pada data berdimensi tinggi seperti teks dan volume yang besar kemudian bisa mengurangi dimensi fitur.

Oleh karena itu, penelitian ini bertujuan membangun model klasifikasi untuk menentukan kategori artikel berita online berdasarkan metode pembobotan *Term Frequency - Inverse Document Frequency* (TF.IDF) dan *Term Frequency Absolute* (TF.ABS) dengan algoritma *Support Vector Machine* (SVM). Pendekatan fitur yang digunakan adalah Unigram. Adapun evaluasi dari penelitian ini menggunakan *confusion matrix*. Batasan pada penelitian ini adalah menggunakan tiga buah kernel SVM yakni kernel *linear, polynomial* dna RBF. Jumlah data sebanyak 2225 artikel berita dengan kategori 5 topik yaitu *business, entertainment, politics, sport* dan *tech.*

II. KAJIAN TEORI

Penelitian [12] melakukan perbandingan pembobotan menggunakan metode TF-ABS, TF-BNS, TF-GR, TF-IDF, TF-IG, TF-OR, dan TF-QUI dengan algoritma SVM dengan tujuan untuk mendapatkan hasil performansi yang optimal menggunakan data dari Federal District Legislative Assembly di Brazil diambil pada tahun 2003-2004. Dari hasil pengujian yang dilakukan, pembobotan TF-ABS memperoleh akurasi yang lebih baik daripada pembobotan lainnya.

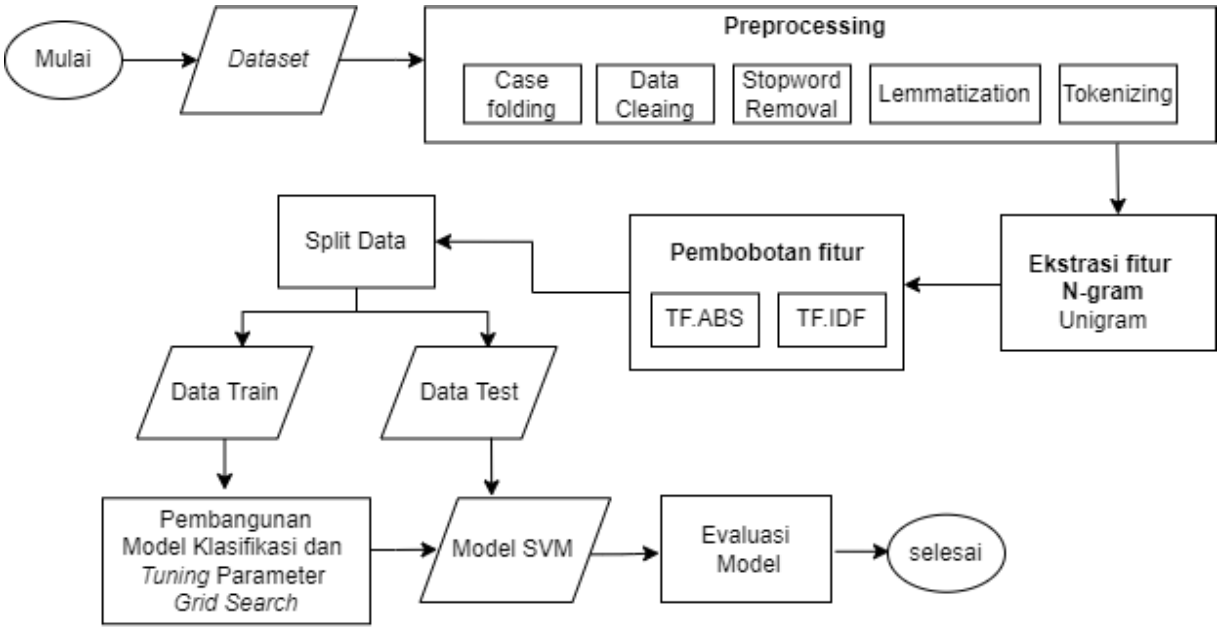
Kemudian, penelitian [13] melakukan klasifikasi untuk menemukan model terbaik mengenai informasi berita *hoax* dan tidak *hoax* menggunakan metode *Support Vector Machine* dengan pembobotan *Term Frecquency-Inverse Document Frequency*. Percobaan dengan *k-fold=10* didapatkan bahwa menggunakan kernel *linear* mampu mendapat akurasi yang lebih tinggi yaitu 95.83 % daripada kernel *Polynomial* yang hanya mendapat 85.83% kemudian kernel RBF dan *sigmoid* hanya mendapat 55%.

Penelitian [3] melakukan perbandingan pembobotan *Term Frecquency-Inverse Document Frequency* dan *Singular Value Decomposition* menggunakan Algoritma *Multinomial Naive Bayes*, *Multivariate Bernoulli Naive Bayes* dan *Support Vector Machine* mengklasifikasikan data mengenai berita artikel berbahasa Indonesia. Dari hasil percobaannya menggunakan kombinasi TF-IDF dengan *Multinomial Naïve Bayes* memberikan nilai *precision* dan *recall* yang tinggi yaitu 98.4% dibandingkan kombinasi TF-IDF dengan *Multivariate Bernoulli Naive Bayes* sebesar 98.2%. Percobaan SDV tidak bekerja dengan baik sehingga tidak memiliki pengaruh terhadap klasifikasi.

Penelitian berikutnya [15] mengenai pengaduan tiket helpdesk pada Direktorat Jendral Kekayaan Negara Kementrian Keuangan menggunakan data sebanyak 10.537 dan memiliki delapan kategori dengan tujuan penelitiannya adalah membandingkan performa KNN menggunakan pembobotan TF-ABS dan TF-IDF. Diperoleh pembobotan TF-ABS mendapat akurasi yang tinggi yaitu 90,04% ketika jumlah fitur 15% saat kondisi K=3 dibandingkan dengan metode TF-IDF mendapat akurasi sebesar 72.11% ketika jumlah fitur sama berada pada 15% dengan K=9

III. METODE

Sistem yang dibangun pada penelitian ini merupakan sistem yang dapat melakukan klasifikasi berita *online* BBC mengggunakan bahasa Inggris dengan Algoritma SVM dengan pembobotan TF-IDF dan TF-ABS. Berikut gambaran sistem yang akan dibangun.



GAMBAR 1
SISTEM KLASIFIKASI BERITA *ONLINE*

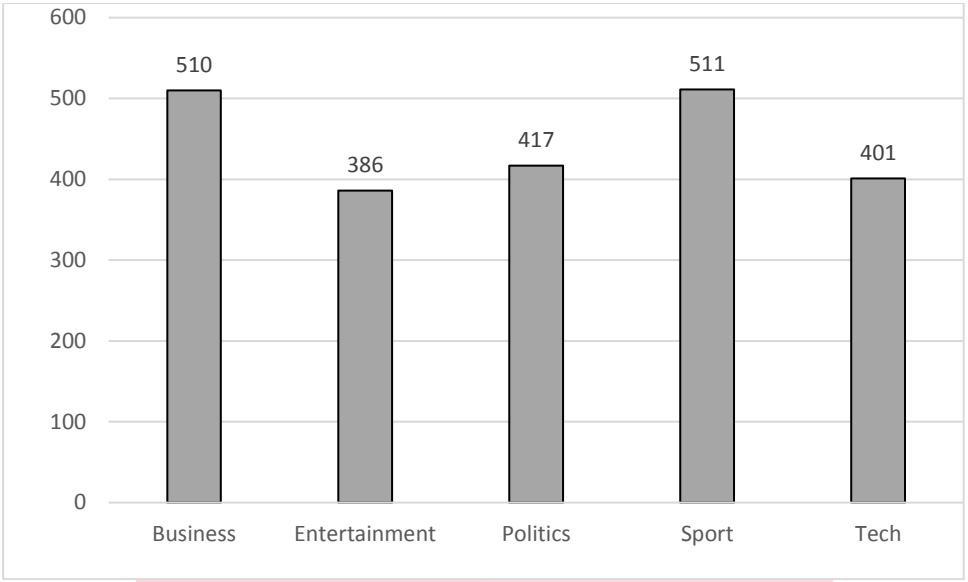
A. Dataset

Penelitian ini menggunakan dataset yang diambil dari BBC *News* mengenai berita *online* [16]. Diperoleh data sebanyak 2225 menggunakan bahasa Inggris yang telah memiliki kategori berdasarkan topiknya dari tahun 2004 - 2005. Adapun data ini merupakan bukan data mengenai suatu produk yang mana tahun menjadi hal yang dipertimbangan dan akan memerlukan kerelevanan pada data. Penelitian ini menggunakan. data yang bersifat umum. Berikut representasi data yang dipakai :

TABEL 1
RESPRESENTASI DATASET

No	Text	Kategori
1	*worldcom ex-boss launches defence lawyers defending former worldcom chief bernie ebbers against a battery of fraud charges have called a company whistleblower as their first witness.*	*Business*
2	*lifestyle governs mobile choice faster better or funkier hardware alone is not going to help phone firms sell more handsets research suggests.*	*Tech*
3	*howard truanted to play snooker conservative leader michael howard has admitted he used to play truant to spend time with his school friends at a snooker hall.*	*Politics*

Adapun jumlah setiap ketegori yang digunakan dapat dilihat pada gambar 2. Perbedaan jumlah data antara satu kategori dengan kategori lain tidak terlalu jauh sehingga perbadaan ini masih *relative balance*.



GAMBAR 2
JUMLAH DATA BERDASARKAN KATEGORI

B. Preprocessing

Tahap awal untuk memulai pemrosesan data adalah dilakukannya *preprocessing* yaitu sebuah proses memasukkan data ke model klasifikasi agar data mendapat hasil kinerja yang baik. Data tersebut diolah untuk menjadi data *train* dan *test*. Tujuan *preprocessing* yaitu membuat dokumen menjadi seragam dan bisa dibaca saat proses klasifikasi [17].

1.*Case Folding*

Case folding merupakan proses mengubah huruf besar menjadi huruf kecil dari dokumen. Huruf-huruf yang diubah berdasarkan huruf alfabet dari A - Z [18]. Ini dilakukan karena tidak semua teks yang terdapat dalam dokumen memiliki huruf yang seragam yaitu huruf besar dan kecil. Maka dari itu, *Case Folding* berperan untuk melakukan konversi teks agar menjadi huruf kecil.

2.*Data Cleaning*

Data cleaning merupakan proses untuk menghilangkan *noise* seperti tanda baca *tag, html, link* dan *script*. Ini dilakukan agar data yang diproses menjadi bersih dari kompenen-komponen yang tidak berhubungan dan tidak diperlukan.

3.*Stopword Removal*

Stopword Removal merupakan proses untuk menghilangkan kata - kata yang tidak memiliki pengaruh pada pengklasifikasian. Teks yang diproses dimasukkan ke daftar kamus *stopword* menggunakan *library stopword* berbahasa Inggris. Ini dilakukan karena kata yang sering muncul tidak memberikan informasi penting pada tiap dokumen. Sehingga tujuan *stopword* yaitu mengetahui kata yang tidak memiliki arti dan atau dianggap tidak penting.

4.*Lemmatization*

Lemmatization merupakan tahapan untuk menemukan kata agar dikembalikan pada kata semula atau membentuk data dasar. *Lemma* mengumpulkan kata-kata yang memiliki arti mirip dengan kata lain yang sama. Ini dilakukan dengan tujuan untuk mengubah *infinite tense* dan *noun* pada kata dalam Bahasa Inggris yang sama [19].

5.*Tokenizing*

Tokenizing merupakan proses untuk menghilangkan spasi dan angka agar menjadi token. Tujuan *tokenizing* yaitu memisahkan kata - kata pada sebuah paragraf, kalimat ataupun halaman agar menjadi kata tunggal.

TABEL 2
CONTOH PROSES *PREPROCESSING*

Kalimat awal : *Enron bosses in $168m payout Eighteen Former enron directors have agreed a $168m (Â£89m) settlement deal in a shareholder lawsuit over the collapse of the energy firm.*	
Proses preprocessing	**Hasil proses**
case folding	*enron bosses in $168m payout eighteen former enron directors have agreed a $168m (a£89m) settlement deal in a shareholder lawsuit over the collapse of the energy firm.*
Data Cleaning	*enron bosses in 168m payout eighteen former enron directors have agreed a 168m a89m settlement deal in a shareholder lawsuit over the collapse of the energy firm*
stoword removal	*enron bosses payout eighteen former enron directors agreed settlement deal shareholder lawsuit over collapse energy firm*
Lemmatization	*enron boss payout eighteen former enron director agree settlement deal shareholder lawsuit over collapse energy firm*
Tokenizing	*"enron", "boss", "payout", "eighteen", "former", "enron", "director", "agree", "settlement", "deal", "shareholder", "lawsuit", "over", "collapse", "energy", "firm"*

C. Pembobotan *(Term Weighting)*

1. Term *Frequency - Inverse Document Frequency* (TF.IDF)

Setelah mendapat hasil dari *preprocessing* dan N-gram maka selanjutnya kata atau fitur akan dibobotkan menggunakan TF-IDF. *Term Frequency - Invers Document Frequency* (TF.IDF) merupakan proses penempatan suatu bobot kata berdasarkan frekuensi dokumen (DF). Semakin banyak kata yang muncul pada dokumen maka kepentingan kata tersebut menjadi sedikit sehingga bobot yang diberikan pun kecil begitupun sebaliknya apabila kemunculan kata semakin sedikit maka bobot yang diberikan menjadi besar sehingga kata tersebut menjadi penting [7][21]. Kemunculan frekuensi sebuah kata menggunakan semua kata – kata yang ada dalam dokumen [22] Sedangkan fitur yang digunakan yaitu semua kata – kata yang ada pada data.

Term Frequency (TF) merupakan cara untuk mencari bobot dari sebuah dokumen. Semakin banyak jumlah kemunculan sebuah kata maka akan mempengaruhi besar bobotnya, sedangkan *Inverse Documen Frequency* (IDF) merupakan proses untuk menghitung penyebaran kata dalam dokumen. Penyebaran kata yang tidak sesuai bisa mempengaruhi hasil dari perhitungan bobot pada dokumen [23]**.** Berikut merupakan persamaan TF-IDF [21].

$$TF - IDF(t,d) = tf_{t,d}\ x \log\left(\frac{N}{df_t}\right) \quad (1)$$

Di mana ;

N : merupakan jumlah dokumen dalam data

$tf_{t,d}$: merupakan kemunculan atau frekuensi kata t pada dokumen ke d

$df_{t,d}$: merupakan jumlah dokumen yang mengandung kata t

TABEL 3
CONTOH TF-IDF DAN TF-ABS

Dokumen Ke	Kalimat
1	*marc marquez crashes grand prix motorcycle racing mandalika (sport)*
2	*future technology implemented Grand Prix motorcycle racing (tech)*
3	*political power marc marquez (politics)*

TABEL 4
PERHITUNGAN TF-IDF PADA FITUR UNIGRAM

Term	$TF(t,d)$			df_t	IDF_t	$TF(t,d) * IDF_t$		
	D1	D2	D3			D1	D2	D3
marc	1	0	1	2	log(3/2)= 0,18	0,18	0	0,18
marquez	1	0	1	2	log(3/2)= 0,18	0,18	0	0,18
crash	1	0	0	1	log(3/1)= 0,48	0,48	0,00	0
grand	1	1	0	2	log(3/2)= 0,18	0,18	0,18	0
prix	1	1	0	2	log(3/2)= 0,18	0,18	0,18	0
motorcycle	1	1	0	2	log(3/2)= 0,18	0,18	0,18	0
racing	1	1	0	2	log(3/2)= 0,18	0,18	0,18	0
mandalika	1	0	0	1	log(3/1)= 0,48	0,48	0	0
future	0	1	0	1	log(3/1)= 0,48	0	0,48	0
technology	0	1	0	1	log(3/1)= 0,48	0	0,48	0
implement	0	1	0	1	log(3/1)= 0,48	0	0,48	0
political	0	0	1	1	log(3/1)= 0,48	0	0	0,48
power	0	0	1	1	log(3/1)= 0,48	0	0	0,48

2. *Term Frequency Absolute (TF.ABS)*

Term Frequency Absolute (TF.ABS) merupakan penggabungan metode TF dan ABS. Sama seperti pada metode TF-IDF, TF-ABS memiliki metode TF yaitu setiap kata memiliki kepentingan yang sesuai dengan kemunculan kata pada tiap dokumen. Dengan arti lain, nilai bobot suatu kata memiliki jumlah yang sama dengan kemunculan kata pada tiap dokumen. Sedangkan metode TF-ABS merupakan kemunculan setiap kata pada tiap dokumen yang ada dalam kategori dan melihat kemungkinan kata yang tidak muncul pada tiap dokumen yang ada dalam kategori [8][9].

Pada tabel 6 terlebih dahulu dilakukan pencarian untuk masing-masing n_{ij} kemudian setiap n_{ij} ditambahkan dengan 0,5 sesuai dengan persamaan 2. Hasil TF ada pada tabel 5. Berikut persamaan 2 untuk menentukan ABS :

$$ABS(t_j, c_i) = \left|\ln\left(\frac{(n_{ij}+0{,}5)+(n_{\bar{i}\bar{j}}+0{,}5)}{(n_{\bar{i}j}+0{,}5)+(n_{i\bar{j}}+0{,}5)}\right)\right| \quad (2)$$

Di mana ;

t_j : merupakan kata t_j

c_i : merupakan kategori c_i

n_{ij} : jumlah dokumen pada kategori c_i yang mengandung term t_j

$n_{\bar{i}\bar{j}}$: jumlah dokumen dengan tidak dalam kategori c_i yang tidak mengandung term t_j

$n_{\bar{i}j}$: jumlah dokumen yang tidak dalam ketegori c_i yang mengandung term t_j

$n_{i\bar{j}}$: jumlah dokumen dalam ketegori c_i yang tidak mengandung term t_j

Setelah hasil ABS diperoleh, selanjutnya dilakukan perkalian matrik antara TF dan ABS untuk mendapatkan nilai bobot TF.ABS. Berikut persamaan dari TF.ABS.

$$TF.ABS = TF(t,d) * ABS(t_j, c_i) \quad (3)$$

TABEL 5
PERHITUNGAN TF.ABS PADA FITUR UNIGRAM

Term	$n_{ij} + 0.5$	$n_{\bar{i}\bar{j}} + 0.5$	$n_{\bar{i}j} + 0.5$	$n_{i\bar{j}} + 0.5$	$ABS(t_j, c_i)$	$TF.ABS = TF(t,d) * ABS(t_j, c_i)$		
						D1	D2	D3
marc	1,5	0,5	0,5	1,5	0,81	0,81	0	0,81
marquez	1,5	0,5	0,5	1,5	0,81	0,81	0	0,81
crash	1,5	0,5	1,5	0,5	1,39	1,39	0	0
grand	1,5	0,5	0,5	1,5	0,81	0,81	0,81	0
prix	1,5	0,5	0,5	1,5	0,81	0,81	0,81	0
motorcycle	1,5	0,5	0,5	1,5	0,81	0,81	0,81	0
racing	1,5	0,5	0,5	1,5	0,81	0,81	0,81	0
mandalika	1,5	0,5	1,5	0,5	1,39	1,39	0	0
future	1,5	0,5	1,5	0,5	1,39	0	1,39	0
technology	1,5	0,5	1,5	0,5	1,39	0	1,39	0
implement	1,5	0,5	1,5	0,5	1,39	0	1,39	0
political	1,5	0,5	1,5	0,5	1,39	0	0	1,39
power	1,5	0,5	1,5	0,5	1,39	0	0	1,39

E. *Support Vector Machine* (SVM)

Support vector machine (SVM) merupakan metode klasifikasi yang bekerja dengan mencari *hyperplane* terbaik. Sedangkan *hyperplane* adalah garis pembatas atau pemisah data dari kelas yang dimiliki [24]. Kelas pertama terletak pada bidang pembatas pertama dan kelas kedua terletak pada bidang pembatas kedua. *Hyperplane* dapat ditemukan dengan margin *hyperplane*. Margin adalah jarak antar *hyperplane* dengan pola terdekat antar kelas. Data atau pola yang ada dibidang pembatas tersebut dinamakan *vectors* terdekat dengan *hyperplane* yang disebut *Support Vector*. *Vector* tersebut berasal dari hasil pembobotan kata yang telah dilakukan. Berikut persamaan *hyperplane* :

$$\vec{w}.\vec{x} + b = 0 \quad (4)$$

Di mana ;

$\vec{w}$: parameter bobot
$\vec{x}$: vector input
b : bias

Dalam hyperplane memiliki dua kelas yaitu kenas naik dan kelas turun. Sample yang bernilai positif masuk pada kelas +1 dan sample kelas negatif masuk pada kelas -1

$$\vec{w}.\vec{x} + b \geq 1, untuk\ Y1 = +1 \quad (5)$$
$$\vec{w}.\vec{x} + b \geq 1, untuk\ Y1 = -1 \quad (6)$$

SVM menjadi salah satu metode klasifikasi *supervised learning* yang menyelesaikan permasalahan pada kasus biner, seperti model yang memiliki 2 kelas, positif dan negatif. Pada klasifikasi tidak biner yang memiliki kelas lebih dari 2 seperti kasus berita *online* yaitu bisnis, olahraga dan hiburan, perlu dilakukan pendekatan menggunakan *multiclass* SVM dengan konsep *One Againts one* (OAO) dan *One Againts All* (OAA). OAA menyelesaikan permasalahn dengan membagi data menjadi dua kelas dimana, kelas *entertaiment* akan diberi label +1, kelas lainnya diberi label -1. Begitupun dengan kelas selanjutnya untuk kelas *sport* akan diberi label +1 dan kelas lainnya diberi label -1. Setelah diterapkan pada semua kelas maka akan didapatkan *hyperplane*. Penelitian ini menggunakan *multiclass* dengan konsep *One Againts All.*

$$kelas\ g = \arg\max_{r=1\ldots n}((w^{(r)})^T.\ \varphi(x) + b^{(r)}) \quad (7)$$

Persamaan 7 digunakan pada tahap akhir proses *testing* setelah *support vectore* pada data *testing* didapatkan. $((w^{(r)})^T$ merupakan *hyperplane* atau bobot kata hasil dari proses TF-IDF dan TF-ABS. $b^{(r)}$ merupakan bias yang bernilai bebas. Nantinya hasil klasifikasi ditentukan dengan $\arg max$ bahwa nilai tertinggi yang akan diambil dari hasil perhitungan semua *hyperplane* sebagai kelas prediksinya [26][27]. Dalam penerapannya SVM tidak selalu menyelesaikan permasalahan kasus data secara linear, maka dibutuhkan *kernel trick* untuk menyelesaikan data secara non linear. Adapun kernel umum yang digunakan yaitu *linear, polynomial dan radal basis function* (RBF) [26][28]. Dalam fungsi kernel trick mengenal parameter C dan Gamma. C atau *complexcity* berfungsi untuk mengontrol kesalahan margin dalam klasifikasi sedangkan gamma betugas utuk menentukan seberapa banyak lengkungan yang diinginkan dalam batas keputusan.

F. Model *Tuning* Parameter

Tuning parameter dilakukan untuk mendapatkan parameter terbaik terhadap data klasifikasi. Salah satu metode yang dapat dilakukan yaitu dengan menggunakan *Grid Search.* Cara kerja dari *Grid Search* adalah mengkombinasikan seluruh nilai parameter yang telah ditentukan. Adapun parameter yang dibutuhkan yaitu, nilai C, Gamma, dan kernel. Misalnya ketika akan mencari nilai dari parameter SVM dengan memasukkan nilai A = [1,2] dan B = [3,4] maka *Grid Search* melakukan semua kombinasi dari A dan B yaitu [1,3], [1,4], [2,3] dan [2,4] kemudian akan dicari kombinasi terbaik dari nilai tersebut.

G. K - Fold Cross Validation

K-*Fold Cross Validation* merupakan metode evaluasi yang menggunakan keseluruhan data dengan membagi data *train* dan data *test* [1] seperti ilustrasi pada tabel 8. Ada k buah partisi acak kemudian dilakukan sebanyak k-kali eksperimen yang mana setiap eksperimen ini menggunakan partisi ke-k sebagai data *test* dan sisa partisi lainnya sebagai data *train*. Ini dilakukan agar pengujian terhadap data mendapat perlakuan secara merata. Perbandingan jumlah data *train* dan *test* yaitu 80:20 dengan nilai k=5. Nilai akurasi dari setiap *fold* akan diambil kemudian dihitung rata-ratanya. Berikut merupakan contoh penerapan *k-fold*

TABEL 6
K-FOLD CROSS VALIDATION DENGAN K=5

Iterasi 1	Test	Train	Train	Train	Train
Iterasi 2	Train	Test	Train	Train	Train
Iterasi 3	Train	Train	Test	Train	Train
Itreasi 4	Train	Train	Train	Test	Train
Iterasi 5	Train	Train	Train	Train	Test

H. Evaluasi Confussion Matrix

Untuk mengevaluasi hasil klasifikasi analisis berita online, maka penelitian ini akan dihitung nilai akurasi dan F1 – Score, untuk mempermudah melihat hasil

performansi maka digunakan *confussion matrix*. Berikut contoh penerapan *confusion matrix* :

TABEL 7
CONFUSSION MATRIX

Kelas asli	Kelas prediksi		
		Positif	Negatif
	Positif	*TP*	*FP*
	Negatif	*FN*	*TN*

True Positif (TP) yakni hasil klasifikasi sesuai harapan dimana hasil prediksi bernilai positif dengan keadaan sebenarnya adalah benar. True Negative (TN) hasil klasifikasi tidak ada yang benar dimana prediksi bernilai negatif dengan keadaan sebenarnya juga negatif. False Negatif (FP) yakni hasil yang tidak diharapkan dimana prediksi bernilai positif dengan keadaan sebenarnya adalah salah. Dan False Negative (FN) merupakan hasil yang melesat dimana prediksi bernilai negatif dengan keadaan sebenarnya adalah benar

A. Akurasi digunakan untuk mengevaluasi banyaknya kelas prediksi yang sesuai dengan dengan kelas asli, akurasi mengambarkan keakuratan model klasifikasi yang dibangun.

$$Akurasi = \frac{TP+TN}{TP+FN+FN+FP} \quad (7)$$

B. Recall merupakan keberhasilan model dalam mengklasifikasikan dokumen dalam menemukan sebuah informasi.

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

C. Precission merupakan hasil keakuratan data yang diminta oleh hasil prediksi dengan yang diberikan oleh model klasifikasi

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

D. Sedengkan F1-Score merupakan nilai evaluasi yang menjadi nilai perbandingan rata rata hasil dari precision dan recall.

$$F1 - Score = 2 \times \frac{prcission \times recall}{precission + recall}, \quad (10)$$

IV. HASIL DAN PEMBAHASAN

Tujuan penelitian ini adalah menentukan kategori artikel berita *online* berdasarkan pembobotan *Term Frequency - Inverse Document Frequency* (TF-IDF) dan *Term Frequency Absolute* (TF-ABS) dengan metode *Support Vector Machine* (SVM). Setelah dilakukan *preprocessing*, N-gram - Unigram menggunakan pembobotan TF-IDF dan TF-ABS kemudian dilakukan klasifikasi menggunaka metode *Support Vector Machine* dan selanjutnya adalah melakukan pengujian menggunakan *K-fold cross validation* sama dengan 5.

A. Pengujian Parameter *Grid Search*

Penelitian ini terlebih dahulu melakukan pengujian *hyperparameter tuning* menggunakan *grid search* dengan tujuan untuk menentukan kombinasi parameter agar mendapatkan hasil yang optimal. Parameter yang diuji meliputi C, gamma dan kernel. Adapun nilai parameter yang diuji ada pada tabel 10

TABEL 8
NILAI *TUNING* PARAMETER

Parameter	Nilai yang diuji
C	0.1, 1, 10, 100, 1000
Gamma	1, 0.1, 0.01, 0.001, 0.0001
Kernel	Polynomial, Linear, RBF

Tabel 11 - 16 merupakan hasil rata – rata kombinasi *tuning* parameter SVM berdasarkan nilai parameter sesuai pada tabel 10. Angka bertanda warna abu - abu memiliki nilai akurasi terbesar yang dihasilkan pada masing – masing C, gamma dan kernel berdasarkan TF-IDF dan TF-ABS.

1. *Grid Search* TF-IDF

TABEL 9
TF-IDF *KERNEL POLYNOMIAL*

C / γ	0.1	1	10	100	1000
1	23,18%	72,34%	74,66%	74,66%	74,66%
0.1	23,18%	23,18%	23,18%	23,18%	72,34%
0.01	23,18%	23,18%	23,18%	23,18%	23,18%
0.001	23,18%	23,18%	23,18%	23,18%	23,18%
0.0001	23,18%	23,18%	23,18%	23,18%	23,18%

Berdasarkan tabel 11, menggunakan pembobotan TF-IDF dengan *kernel polynomial* kombinasi optimal diperoleh ketika C=10, 100, 1000 digabungkan dengan gamma=1 mendapat akurasi sebesar 74,66%. Akurasi terendah didapat sebesar 23,18%. Dari kombinasi parameter menggunakan *kernel polynomial* didapat semakin besar nilai C dan gamma yang dipilih maka dapat berpengaruh terhadap hasil akurasi.

TABEL 10
TF-IDF *KERNEL LINEAR*

C / γ	0.1	1	10	100	1000
1	88,44%	97,74%	97,92%	97,92%	97,92%
0.1	88,44%	97,74%	97,92%	97,92%	97,92%
0.01	88,44%	97,74%	97,92%	97,92%	97,92%
0.001	88,44%	97,74%	97,92%	97,92%	97,92%
0.0001	88,44%	97,74%	97,92%	97,92%	97,92%

Berdasarkan tabel 12, menggunakan pembobotan TF-IDF dengan *kernel linear* kombinasi optimal diperoleh ketika C=10,100,1000 digabungkan dengan gamma=1, 0.1, 0.01, 0.001, dan 0.0001 mendapat akurasi yang sama sebesar 97.92%. Adapun akurasi terendah didapat sebesar

88,44%. Dari kombinasi parameter menggunakan *kernel linear* didapat semakin besar nilai C maka dapat berpengaruh terhadap hasil akurasi.

TABEL 11
TF-IDF *KERNEL RBF*

C / γ	0.1	1	10	100	1000
1	31,86%	97,08%	97,22%	97,22%	97,22%
0.1	23,18%	94,66%	97,80%	97,86%	97,86%
0.01	23,18%	23,18%	95,50%	97,92%	97,92%
0.001	23,18%	23,18%	23,18%	95,50%	97,92%
0.0001	23,18%	23,18%	23,18%	23,18%	95,50%

Berdasarkan tabel 13, menggunakan pembobotan TF-IDF dengan *kernel rbf* kombinasi optimal diperoleh ketika C=100 dan 1000 digabungkan dengan gamma=0.01 mendapat akurasi sebesar 97.92%. akurasi terendah didapat sebesar 23,18%. Dari kombinasi parameter menggunakan *kernel rbf* semakin besar nilai C yang ditentukan maka dapat menaikkan hasil akurasi.

2. *Grid Search* TF-ABS

TABEL 12
TF-ABS *KERNEL POLYNOMIAL*

C / γ	0.1	1	10	100	1000
1	87,34%	90,50%	91,76%	91,76%	91,76%
0.1	23,18%	23,18%	47,88%	87,32%	90,50%
0.01	23,18%	23,18%	23,18%	23,18%	23,18%
0.001	23,18%	23,18%	23,18%	23,18%	23,18%
0.0001	23,18%	23,18%	23,18%	23,18%	23,18%

Berdasarkan tabel 14, menggunakan pembobotan TF-ABS dengan *kernel polynomial* kombinasi optimal diperoleh ketika C=10, 100 dan 1000 digabungkan dengan gamma=1 mendapat akurasi sebesar 91,75%. Akurasi terendah didapat sebesar 23,18% Dari kombinasi parameter menggunakan *kernel polynomial* semakin besar nilai C dan gamma yang ditentukan maka dapat menaikkan hasil akurasi.

TABEL 13
TF-ABS *KERNEL LINEAR*

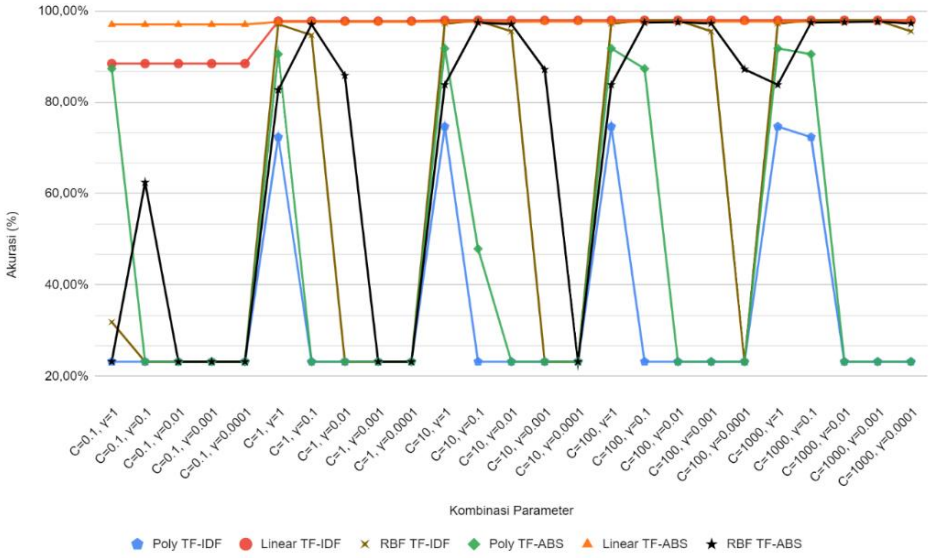
C / γ	0.1	1	10	100	1000
1	97,00%	97,56%	97,62%	97,62%	97,62%
0.1	97,00%	97,56%	97,62%	97,62%	97,62%
0.01	97,00%	97,56%	97,62%	97,62%	97,62%
0.001	97,00%	97,56%	97,62%	97,62%	97,62%
0.0001	97,00%	97,56%	97,62%	97,62%	97,62%

Berdasarkan tabel 15, menggunakan pembobotan TF-ABS dengan *kernel linear* kombinasi optimal diperoleh ketika C=10, 100 dan 1000 digabungkan dengan gamma=1, 0.1, 0.01, 0.001, 0.0001 mendapat akurasi sebesar 97,62%. Akurasi terendah didapat sebesar 97,00%. Dari kombinasi parameter menggunakan *kernel linear* semakin besar nilai C maka dapat menaikkan akurasi. Adapun pada gamma semakin besar parameter yang ditentukan hasil akurasi tidak memiliki perubahan yang signifikan.

TABEL 14
TF-ABS *KERNEL RBF*

C / γ	0.1	1	10	100	1000
1	23,24%	82,70%	83,82%	83,82%	83,82%
0.1	62,44%	97,02%	97,42%	97,42%	97,42%
0.01	23,18%	85,84%	97,08%	97,52%	97,52%
0.001	23,18%	23,18%	87,14%	97,24%	97,62%
0.0001	23,18%	23,18%	23,18%	87,20%	97,24%

Berdasarkan tabel 16, menggunakan pembobotan TF-ABS dengan *kernel rbf linear* kombinasi optimal diperoleh ketika C=1000 digabungkan dengan gamma=0,001 mendapat akurasi sebesar 97,62%. Akurasi terendah didapatkan sebesar 23.18%. Dari kombinasi parameter menggunakan *kernel rbf* semakin besar nilai C dengan nilai gamma yang tepat maka dapat memberikan pengaruh kenaikan terhadap hasil akurasi.



GAMBAR 3
VISUALISASI *GRID SEARCH*

B. Hasil Akurasi Support Vector Machine TF-IDF dan TF-ABS

Setelah dilakukan kombinasi parameter, diperoleh kombinasi paling optimal untuk setiap pembobotan TF-IDF dan TF-ABS yaitu ketika C=10, gamma=1 menggunakan kernel linear. Kombinasi inilah yang dijadikan pengujian pada data *test* untuk mendapatkan hasil akurasi.

TABEL 18
HASIL AKURASI *K-FOLD* = 5

K-Fold	**TF-IDF + SVM**	**TF-ABS + SVM**
1	100,00%	94,38%
2	97,75%	94,38%
3	92,13%	85,39%
4	97,75%	89,89%
5	95,51%	84,27%
Rata – Rata	**96,63%**	**89,66%**
F1-Score	**97,06%**	**96,63%**

Setelah diketahui parameter C, gamma dan kernel yang optimal berdasarkan tabel 17 pada masing – masing pembobotan. Selanjutnya melakukan pengujian pada artiel berita online menggunakan *k-fold* sama dengan 5. Berdasarkan hasil pengujian pada tabel 18 untuk pembobotan TF-IDF mendapat rata – rata akurasi sebesar 96,63% dengan hasil tertinggi sebesar 100,00% ketika K=1 kemudian akurasi terendah sebesar 97,75 % ketika K=4, *F1-Score* yang diperoleh sebesar 97,06%. Selanjutnya TF-ABS mendapat rata – rata akurasi sebesar 89,66% dengan hasil tertinggi sebesar 94,38% ketika K=1 dan 2, dan akurasi terendah sebesar 84,27% ketika K=4, *F1-score* mendapat 96,63%. Adapun kenaikan akurasi mendapat 6,97%.

C. Analisis hasil pengujian

Berdasarkan hasil pengujian yang telah dilakukan dengan data berita online, hasil *tuning* menggunakan C, gamma dan kernel SVM sangat berpengaruh pada pembobotan TF-IDF dan TF-ABS dengan algoritma *Support Vector Machine*. Untuk setiap pembobotan akurasi terbaik TF-IDF mendapat rata-rata 96,63% sedangkan TF-ABS SVM mendapat rata–rata 89,66% dengan kenaikan akurasi sebesar 6,97%. Berdasarkan hasil akurasi tersebut, kinerja klasifikasi dengan metode pembobotan TF-IDF menggunakan algoritma SVM mendapat hasil yang unggul dan cocok digunakan begitupun sebaliknya penggunaan metode TF-ABS tidak cocok apabila menggunakan Support Vector Machine dengan data berita online. Hal ini disebabkan karena metode TF-ABS memperhitungkan kategori untuk perhitungan bobotnya dengan menghitung jumlah dokumen yang masuk pada kategori tertentu dan menghitung jumlah dokumen yang tidak masuk pada suatu kategori dengan melihat muncul atau tidak munculnya kata pada pada masing – masing kategori, dalam hal ini kata – kata pada kategori akan dipertimbangkan untuk kemudian dibobotkan. Sebaliknya dengan TF-IDF semua kata pada semua kategori dijumlahkan kemudian dibobotkan dan metode SVM memiliki kelebihan dapat diimplementasikan pada ruang berdimenasi tinggi.

V. KESIMPULAN

Penelitian ini telah melakukan klasifikasi berita online BBC menggunakan Support Vector Machine menggunakan pembobotan *Term Frequency – Invers Document Frequency* (TF-IDF) dan *Term Frequency – Absolute* (TF-ABS). Dapat disimpulkan bahawa TF-IDF dapat memberikan hasil yang unggul digabungkan dengan algoritma *Support Vector Machine*, dimana rata-rata akurasi yang dihasilkan TF-IDF lebih tinggi dibandingkan TF-ABS. Besar peningkatan akurasi yang diperoleh lebih dari 6%. Akurasi terbaik yang diperoleh oleh sistem menggunakan TF-IDF adalah 96,63%. Hasil *experiment* ketika pengujian data train juga menunjukkan bahwa dengan menggunakan TF-IDF memperoleh hasil yang unggul dibandingkan TF-ABS, akurasi sistem mencapai nilai yang stabil ketika nilai C=10,100 dan 1000 dikombinasikan dengan keseluruhan gamma=0.0001, 0.001, 0.01, 0.1, dan 1. Dalam hal ini dengan nilai C yang

besar memiliki pengaruh untuk mengontrol kesalahan margin dalam klasifikasi, adapun nilai gamma tidak memiliki pengaruh terhadap hasil akurasi. Pada penelitian ini hasil *tuning* parameter dapat mempengaruhi hasil tingkat akurasi *Support Vector Machine* dengan kombinasi optimal untuk masing masing pembobotan yaitu C=10, gamma=1 dengan kernel linear. Hasil ketepatan dalam mengklasifikasikan dokumen dengan baik pada TF-IDF diperoleh oleh kelas target *entertainment* dan *sport* dibandingkan TF-ABS tidak ada kelas yang berhasil diklasifikasikan dengan benar.

Saran pada penelitian selanjutnya yaitu menggunakan *extrasi fitur* lain selain unigram kemudian menggunakan tuning parameter yang sesuai dengan kebutuhan kernel pada SVM.

REFERENSI

[1] S. N. Asiyah and K. Fithriasari, "Klasifikasi Berita Online Menggunakan Metode Support Vector Machine Dan K-Nearest Neighbor Online News Classification Using Support Vector Machine and K-Nearest," *Jurnal Sains dan Seni ITS*, vol. 5, no. 2, 2016.

[2] Sora N, "Pengertian Berita Dan Ciri-Ciri Berita Yang Baik".

[3] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, and Rudy, "News Article Text Classification in Indonesian Language," *Procedia Comput Sci*, vol. 116, pp. 137–143, 2017, doi: 10.1016/j.procs.2017.10.039.

[4] I. M. Parapat, M. T. Furqon, and Sutrisno, "Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 10, pp. 3163–3169, 2018.

[5] V. V Krzhizhanovskaya *et al.*, *Are n-gram Categories Helpful in Text Classification*. 2020. doi: 10.1007/978-3-030-50417-5.

[6] G. Domeniconi, G. Moro, R. Pasolini, and C. Sartori, "A study on term weighting for text categorization: A novel supervised variant of tf.idf," *DATA 2015 - 4th International Conference on Data Management Technologies and Applications, Proceedings*, pp. 26–37, 2015, doi: 10.5220/0005511900260037.

[7] A. Fauzi, E. B. Setiawan, and Z. K. A. Baizal, "Hoax News Detection on Twitter using Term Frequency Inverse Document Frequency and Support Vector Machine Method," *J Phys Conf Ser*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012025.

[8] M. A. Kurniawan, Y. Sibaroni, and K. L. Muslim, "Kategorisasi Berita Menggunakan Metode Pembobotan TF.ABS dan TF.CHI," *Indonesian Journal on Computing (Indo-JC)*, vol. 3, no. 2, p. 83, 2018, doi: 10.21108/indojc.2018.3.2.236.

[9] H. Tantyoko, Adiwijaya, and U. N. Wisesty, "Perbandingan Pembobotan untuk Klasifikasi Topik Berita menggunakan Decision Tree," *Jurnal Teknologia*, vol. 2, no. 1. pp. 97–113, 2019.

[10] T. B. Shahi and A. K. Pant, "Nepali news classification using Naïve Bayes, Support Vector Machines and Neural Networks," *Proceedings - 2018 International Conference on Communication, Information and Computing Technology, ICCICT 2018*, vol. 2018-Janua, no. February, pp. 1–5, 2018, doi: 10.1109/ICCICT.2018.8325883.

[11] D. Rahmawati and M. L. Khodra, "Automatic multilabel classification for Indonesian news articles," *ICAICTA 2015 - 2015 International Conference on Advanced Informatics: Concepts, Theory and Applications*, pp. 1–6, 2015, doi: 10.1109/ICAICTA.2015.7335382.

[12] L. A. Matsunaga and N. F. F. Ebecken, "Term Weighting Approaches for Text Categorization Improving," *Proceedings - 8th International Conference on Intelligent Systems Design and Applications, ISDA 2008*, vol. 1, pp. 409–414, 2008, doi: 10.1109/ISDA.2008.21.

[13] D. Maulina and R. Sagara, "Klasifikasi Artikel Hoax Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency-Inverse Document Frequency," *Jurnal Mantik Penusa*, vol. 2, no. 1, pp. 35–40, 2018.

[15] Riza Adrianti Supono and Muhammad Azis Suprayogi, "Perbandingan Metode TF-ABS dan TF-IDF Pada Klasifikasi Teks Helpdesk Menggunakan K-Nearest Neighbor," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 5, pp. 911–918, Oct. 2021, doi: 10.29207/resti.v5i5.3403.

[16] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," *ACM International Conference Proceeding Series*, vol. 148, no. 2004, pp. 377–384, 2006, doi: 10.1145/1143844.1143892.

[17] S. B. Setiawan and M. S. Mubarok, "Klasifikasi Topik Berita Menggunakan Metode Weighted K-Nearest Neighbor," vol. 5, no. 2, pp. 1–7, 2015.

[18] Y. Wibisono and M. L. Khodra, "Clustering Berita Berbahasa Indonesia," *Universitas (Stuttg)*, pp. 1–4, 2005.

[19] M. Allahyari *et al.*, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," Jul. 2017, [Online]. Available: http://arxiv.org/abs/1707.02919

[20] A. S. Nugraha and K. K. Purnamasari, "Penerapan Metode Support Vector Machine Pada Part of Speech Tag Bahasa Indonesia," no. 112, 2019.

[21] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," in *Procedia Computer Science*, 2013, vol. 17, pp. 26–32. doi: 10.1016/j.procs.2013.05.005.

[22] Rathinam Technical Campus. Department of Computer Science & Engineering, Rathinam Technical Campus. Department of Information Technology, Technically Enriched Software Engineers, Institute of Electrical and Electronics Engineers, Institute of Electrical and Electronics

Engineers. Madras Section, and IEEE Computational Intelligence Society, *A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification*.

[23] M. P. Shakina Rizkia, Erwin Budi Setiawan S.Si., M.T, Diyas Puspandari S.S., "Analisis Sentimen Kepuasan Pelanggan Terhadap Internet Provider Indihome di Twitter Menggunakan Metode Decision Tree dan Pembobotan TF-IDF," *e-Proceeding of Engineering*, vol. 6, no. 2, pp. 9683–9693, 2019.

[24] H. Khaulasari, "Combine Sampling - Least Square Support Vector Machine Untuk Klasifikasi Multi Class Imbalanced Data," 2016.

[26] A. S. Nugraha and K. K. Purnamasari, "Penerapan Metode Support Vector Machine Pada Part of Speech Tag Bahasa Indonesia," no. 112, 2019.

[27] E. Ramon *et al.*, "Klasifikasi Status Gizi Bayi Posyandu Kecamatan Bangun Purba Menggunakan Algoritma Support Vector Machine (SVM)," *Jurnal Sistem Informasi dan Informatika (Simika) P-ISSN*, vol. 5, pp. 2622–6901, 2022.

[28] B. Santosa, "Tutorial Support Vector Machine."