PROPOSAL PROYEK DATA MINING

Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer

Disusun oleh:

1. 12S17028 – Andri Reimondo Tamba
2. 12S17029 – Silvany Angelia Lumban Gaol
3. 12S17041 – Dewi Purnama Napitupulu

PROGRAM STUDI SARJANA SISTEM INFORMASI
FAKULTAS TEKNIK INFORMATIKA DAN ELEKTRO
INSTITUT TEKNOLOGI DEL
NOVEMBER 2020

DAFTAR ISI

1 Business Understanding

1.1 Determine Bussiness Objective

Proyek kali ini akan dilakukan dalam lingkup penelitian. Tujuan dari proyek ini adalah menolong para pembaca berita untuk lebih mudah memilih berita yang akan di bacanya. Hal tersebut dapat direalisasikan dengan mengklasifikasikan berita kedalam beberap kategori. Pada tahun 2007 algoritma nazief stemmer kemudiandikembangkan lagi oleh Jelita Asian, dengan menambahkan beberapa perbaikan yang bertujuan untuk meningkatkan hasil stemming yang diperoleh.Algoritma ini kemudian dikenal sebagai confix-stripping stemmer[1]. Penelitian ini juga menggunkan beberapa metode yaitu Support Vactor Machine (SVM) yang bertujuan adalah untuk membangun OSH (Optimal Separating Hyperplane), yang membuat fungsi pemisahan optimum yang dapat digunakan untuk klasifikasi. Penelitian ini juga menggunakan metode Naïve Bayes Classifier yang menyediakan data training. Tujuan adanya data training adalah untuk menghasilkan model dari NBC untuk mengetahui ketepatan dari klasifikasi selain itu pada data training juga memperhatikan waktu yang diperlukan pada pembentukan model.Berikut merupakan hasil dari data training.

1.2 Situation Assesment

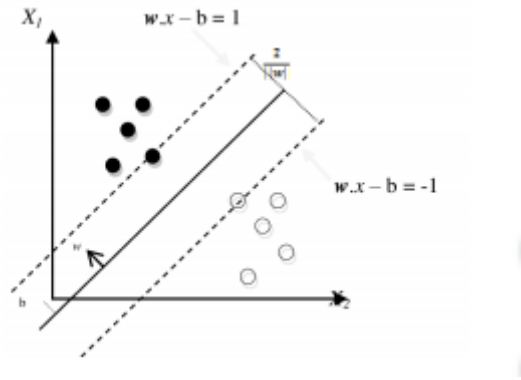
Pada tahun 2006 pertumbuhan dan pertukaran informasi sudah mencapai lebih dari 550 triliun dokumen dan 7,3 juta Internet page baru tiap harinya. Salah satu dampaknya adalah artikel berita yang diunggah di internet sangatlah banyak dan rentang waktu yang cepat. Selama ini pengkategorian berita masih menggunakan tenaga manusia atau manual.Kategori yang banyak beserta waktu yang cepat akan menyulitkan editor untuk mengkategorikan, terutama artikel yang tidak terlalu berbeda secara jelas. Beberapa kategori yang penggunaan bahasanya tidak berbeda terlalu jauh seperti nasional, internasional, sains, ekonomi, tekno, health, dan properti mengharuskan seorang editor mengetahui isi artikel yang akan diunggah secara keseluruhan untuk selanjutnya dimasukkan ke dalam kategori yang tepat.Akan lebih efesien apabila kategori berita dimasukkan secara otomatis dengan komputermenggunakan metode tertentu. [1]

1.3 Determine Data Mining Goal

Text mining adalah cara agar teks dapat diolah dengan menggunakan komputer untuk menghasilkan analisis yang bermanfaat. Praproses dalam text mining diantaranya adalah tokenizing, case folding, stopwords, dan stemming. Diantara keempat langkah tersebut yang paling penting adalah proses stemmingyang merupakan proses

menghilangkan imbuhan pada suatu kata untuk mendapatkan kata dasar dari kata tersebut. Penelitian kali ini akan menggunakan Text Mining untuk pengklasifikasian Betrita Indonesia. Berikut Startegi utama yang digunakan :

1. SVM adalah metode *supervised learning* yang digunakan untuk klasifikasi serta regresi. Klasifikasi dan regresi keduanya adalah sub kategori dari *supervised learning.* Klasifikasi adalah sesuatu yang dapat didefinisikan sebagai memprediksi label sedangkan regresi adalah tentang memprediksi kuantitas. Jadi tugas utama *classifier Support Vector Machine* adalah melakukan klasifikasi, yaitu mengklasifikasikan data dalam kelas yang berbeda dengan menggambar hyperplane yang membedakan antara kelas berbeda yang kita plot dalam ruang n-dimensi [2]



Gambar 1 *Hyperplane* pada SVM

Sumber : ***ICTACT Journal On Soft Computing:Detection On Twitter Data Using Support Vector Machine(***Ashima Garg and Neelam Duhan, 2020)

Hyperplane yang digambar oleh SVM digambar dengan bantuan fungsi matematika yang disebut kernel. Titik data yang paling dekat dengan *hyperplane* disebut vektor dukungan dan metode ini disebut *Support Vector Machine.* Kernel yang akan digunakan dengan penelitian yang ditunjukkan di persamaan di bawah ini:

$$K\left(\bar{w}\cdot\bar{x},x_d\right)=\left(X_i^T X_j+C\right),\gamma>0$$

1.4 Produce Project Plan

Project plan merupakan komponen rencana proyek yang dilakukan selama pengerjaan proyek. Pada proyek berikut, yang menjadi tahapan perencanaan dalam pengerjaan proyek:

- Pengumpulan dataset, berupa artikel bahasa Indonesia dan melakukan analisis terhadap data.
- Melakukan tahapan pra pemrosesan data untuk mencegah adanya redundansi pada data.
- Mengukur kketepatan dan waktu klasifikasi SVM terhadap data training.
- Melakukan pemenggalan kata, seperti pada gambar berikut:

No	Format Kata	Pemenggalan
1	berV..	Ber-V.. | be-rV..
2	berCAP..	Ber-CAP.. dimana C!- 'r' & P!-'er'
3	berCAerV..	Ber-CaerV.. dimana C!='r'
4	Belajar	Bel-ajar
5	beC$_1$erC$_2$..	beC$_1$erC$_2$.. dimana C$_1$!={'r'|'l'}
6	terV..	Ter-V.. | te-rV..
7	terCerV..	Ter-CerV.. dimana C!='r'
8	terCP…	Ter-CP.. dimana C!='r' dan P!='er'
9	teC$_1$erC$_2$ ..	Te-C$_1$erC$_2$… dimana C$_1$!='r'
10	Me{l|r|w|y}V..	Me-{l|r|w|y}V…
11	Mem{b|f|v}…	Mem-{b|f|v}…
12	Mempe{r|l}…	Mem-pe..
13	Mem{rV|V}…	Me-m{rV|V}… | me-p {rV|V}…
14	Men{c|d|j|z}…	Men-{c|d|j|z}…
15	menV…	Me-nV.. |me-tV..
16	Meng{g|h|q}…	Meng-{g|h|q}…
17	mengV…	Meng-V… |meng-kV…
18	menyV…	Meny-sV…
19	mempV…	mempV… dimana V!='e'
20	Pe{w|y}V…	Pe-{w|y}V…
21	perV…	Per-V…|pe-rV…
22	perCAP..	Per-CAP.. dimana C!='r' dan P!='er'
23	perCAerV…	Per-CAerV…dimana C!='r'
24	Pem{b|f|V}..	Pem-{b|f|V}..
25	Pem{rV|V}…	Pe-m{rV|V}… | pe-p{rV|V}…
26	Pen{c|d|j|z}…	Pen-{c|d|j|z}…
27	penV…	Pe-nV…|pe-lV..
28	Peng{g|h|q}…	Peng-{g|h|q}…

- Melakukan pengukuran performa, berupa: akurasi, recall, serta precision.

$$akurasi = \frac{jumlah\ klasifikasi\ benar}{jumlah\ dokumen\ uji\ coba} \times 100\%$$

$$recall = \frac{\left|\{relevant\ doc\} \cap \{retrieved\ doc\}\right|}{\left\{\left|relevant\ doc\right|\right\}}$$

$$precision = \frac{\left|\{relevant\ doc\} \cap \{retrieved\ doc\}\right|}{\left\{\left|retrieved\ doc\right|\right\}}$$

$$F = \frac{2 \times recall \times precision}{recall + precision}$$

- Melakukan tahapan analisis yang meliputi klasifikasi pada teks menggunakan SVM.

2 Data Understanding

2.1 Collect Initial Data

Tahapan Collect Initial data merupakan tahapan yang dilakukan terhadap data penelitian yang biasanya berada pada akhir pengumpulan data ataupun pada saat memasukkan (entry) data. dimulai dari melakukan analisis statistic yang membahas terkait pertanyaan penelitian. Tahapan inisialisasi (pengenalan) data, dilakukan dengan terlebih dahulu melakukan pengumpulan data. Pengumpulan data, dilakukan dengan mengumpulkan data berupa artikel bahasa Indonesia yang dapat diperoleh dari internet. Namun, jumlah yang banyak serta kesulitan yang dihadapi dalam memproses data berbentuk sequence (teks), kemudian pada artikel tersebut akan dilakukan tahapan pra-pemrosesan data, yang dilakukan dengan: case folding, tokenizing, stopwords, dan stemming. Data yang dipilih merupakan data teks, dikarenakan data berbentuk teks memiliki bentuk yang lebih terstruktur.

2.2 Describe Data

Pada tahapan ini akan dijabarkan sedikit bahasan mengenai data yang akan digunakan dan diproses pada pengerjaan proyek ini. Data yang akan dikklasifikasi berupa artikel berita, dapat diperoleh tidak hanya melalui internet, namun dapat diperoleh dari media lain yang memungkinkan untuk mengakses berita. Juga dapat berupa dokumen, yang kemudian akan diunggah secara keseluruhan untuk mengkategorikan berdasarkan beberapa kategori tertentu.

2.3 Explore Data

Pada proses ini dilakukan dengan menggunakan fungsi statistik, matematik, dan divisualisasikan dalam bentuk grafik. Hal ini dilakukan untuk mempermudah dalam pemahaman dan pola dasar data. Selanjut nya dilakukan explore data untuk menemukan data yang bermanfaat untuk data penelitian, dan dapat digunakan untuk proses penambangan data. Pada tahap ini juga dilakukan pra-proses data misalnya pembersihan data outlier dan penanganan data yang kosong, sehingga siap untuk diolah dan dianalisa.

Kategori	Akurasi	Precision	Recall	F-Measure
Nasional	73,30%	75,90%	73,30%	74,60%
Internasional	80,00%	72,70%	80,00%	76,20%
Olahraga	80,00%	96,00%	80,00%	87,30%
Sains	80,00%	75,00%	80,00%	77,40%
Edukasi	93,30%	77,80%	93,30%	84,80%
Ekonomi	76,70%	76,70%	76,70%	76,70%
Tekno	76,70%	100,00%	76,70%	86,80%
Entertaiment	90,00%	96,40%	90,00%	93,10%
Otomotif	83,30%	100,00%	83,30%	90,90%
Health	93,30%	87,50%	93,30%	90,30%
Properti	66,70%	83,30%	66,70%	74,10%
Travel	93,30%	65,10%	93,30%	76,70%
Rata-rata	**82,20%**	**83,90%**	**82,20%**	**82,40%**

Pada gambar diatas diamati bahwa dataset memiliki 5 kolom dan 12 baris. Dari gambar dapat dilihat niali dari Akurasi, Precission, Recall dan F-Measure terhadap baris masing masing data. Berdasarkan rata-rata *akurasi, recall, precision,* dan *F-Measure* pada Tabel memperlihatkan hasil yang cukup baik. Masing-masing nilainya adalah 82,2%, 83,9%, 82,2%, dan 82,4%. Untuk tingkat akurasi paling tinggi dihasilkan oleh kategori berita edukasi, health, dan travel dengan nilai akurasi 93,3%. Berbeda dengan tiga kategori tersebut kategori berita properti menghasilkan akurasi yang paling rendah yaitu 66,7%. Untuk ukuran *precision* kategori tekno dan otomotif bernilai 100% sebaliknya travel menjadi yang paling rendah yaitu 65,1%. *Recall* paling tinggi terdapat pada kategori edukasi, health, dan travel sedangkan untuk paling rendah adalah properti. Untuk ukuran gabungan dari *precision* dan *recall* yaitu *F-Measure* memperlihatkan bahwa kategori entertaiment adalah yang paling tinggi sedangkan yang paling rendah adalah properti. EDA juga berfungsi untuk mengoptimalkan pengetahuan mengenai data. Salah satu metode tradisional dalam EDA adalah visualisasi dalam bentuk grafik.

2.4 Verify Data Quality

Pada proses ini dilakukan untuk menemukan kualitas dari dataset yang akan digunakan apakah dataset tersebut terdapat *outlier* dan *missing value* atau tidak. Setelah mengetahui bahwa dataset yang dimiliki mengandung *outlier* ataupun *missing value*, atau dengan mengisi baris data yang mengandung *outlier* tersebut dengan suatu nilai yang konstan. Skenario yang diambil dalam penelitian ini adalah menghapus kolom yang mengandung *missing value* dan juga *outlier* dengan mempergunakan fungsi dropna() dan drop().

Kategori	Akurasi	Precision	Recall	F-Measure
Nasional	86.7%	81.3%	86.7%	83.9%
Internasional	90.0%	73.0%	90.0%	80.6%
Olahraga	86.7%	89.7%	86.7%	88.1%
Sains	80.0%	82.8%	80.0%	81.4%
Edukasi	86.7%	96.3%	86.7%	91.2%
Ekonomi	90.0%	73.0%	90.0%	80.6%
Tekno	83.3%	100.0%	83.3%	90.9%
Entertaiment	96.7%	90.6%	96.7%	93.5%
Otomotif	86.7%	100.0%	86.7%	92.9%
Health	93.3%	96.6%	93.3%	94.9%
Properti	93.3%	93.3%	93.3%	93.3%
Travel	83.3%	92.6%	83.3%	87.7%
Rata-rata	88.1%	89.1%	88.1%	88.3%

Pada gambar diatas didapati setelah melakukan *testing* data hasil klasifikasi data testing menggunakan SVM linier pada Tabel ini menunjukkan performa yang cukup baik dengan masing-masing nilai dari akurasi, *precision, recall*, dan *F-Measure* adalah 88,1%, 89,1%, 88,1%, dan 88,3%. Kategori berita entertaiment menjadi kategori dengan tingkat akurasi yang paling tinggi yaitu 96,7%, sebaliknya sains menjadi kategori dengan tingkat akurasi yang paling rendah yaitu 80,0%. Untuk precision dengan nilai paling baik adalah kategori tekno dan otomotif sebesar 100%, sedangkan kategori internasional dan ekonomi adalah kategori dengan precision terendah sebesar 73,0%. Hasil recall tertinggi adalah kategori entertaiment dengan nilai sebesar 96,7% dan terendah adalah kategori sains dengan nilai sebesar 80,0%. Nilai F-Measure menunjukkan performa yang paling baik adalah kategori health 94,9%, sedangkan yang paling rendah adalah kategori internasional dan ekonomi 80,6%. Kemudian data tersebut diverify setelah melewati *testing data*.

Referensi

[1] D. Ariadi and K. Fithriasari, "Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer," *JURNAL SAINS DAN SENI ITS,* pp. 2337-3520, 2015.

[2] R. Suman and J. Slngh, "Sentimen Analysis of Tweets Using Support Vector Machine," *International Jurnal of Computer Science and Mobile Apllications,* vol. 5, no. 10, pp. 83-91, 2017.