JTIP
Jurnal Teknologi Informasi dan Pendidikan
http://tip.ppj.unp.ac.id
Vol. 12, No. 2, 2019

P-ISSN: 2086 – 4981
E-ISSN: 2620 – 6390

NAÏVE BAYES CLASSIFIER DAN *SUPPORT VECTOR MACHINE* SEBAGAI ALTERNATIF SOLUSI UNTUK *TEXT MINING*

IIN ERNAWATI
Prodi Teknik Informatika, Fakultas Ilmu Komputer, UPN "Veteran" Jakarta
E-mail : iin_ernawati@yahoo.com

ABSTRACT

This study was conducted to text-based data mining or often called text mining, classification methods commonly used method Naïve bayes classifier (NBC) and support vector machine (SVM). This classification is emphasized for Indonesian language documents, while the relationship between documents is measured by the probability that can be proven with other classification algorithms. This evident from the conclusion that the probability result Naïve Bayes Classifier (NBC) word "party" at least in the economic document and political. Then the result of the algorithm support vector machine (svm) with the word "price" and "kpk" contains in both economic and politic document.

***Keyword** : Classification, Probability, Naïve Bayes Classifier, Support Vector Machine*

INTISARI

Penelitian ini dilakukan untuk penambangan data berbasis teks atau sering disebut penambangan teks, metode klasifikasi yang biasa digunakan metode Naïve bayes classifier (NBC) dan support vector machine (SVM). Klasifikasi ini ditekankan untuk dokumen berbahasa Indonesia, sedangkan hubungan antar dokumen diukur dengan probabilitas yang dapat dibuktikan dengan algoritma klasifikasi lainnya. Ini terbukti dari kesimpulan bahwa probabilitas hasil Naïve Bayes Classifier (NBC) kata "partai" setidaknya dalam dokumen ekonomi dan politik. Kemudian hasil dari algoritma dukungan mesin vektor (svm) dengan kata "harga" dan "kpk" berisi dokumen ekonomi dan politik.

Kata kunci : klasifikasi, probabilitas, Naïve Bayes Classifier, Support Vector Machine.

PENDAHULUAN

Perkembangan informasi global menuntut penyediaan informasi tersebut dapat dinikmati/dirasakan secara cepat dan tepat. Informasi berupa teks yang diinginkan dapat diakomodasi oleh teknologi komputer khususnya internet, karena internetlah yang menjadi acuan utama beberapa penelitian mengenai penambangan data berbasis teks dilakukan atau yang sering disebut denga teks dilakukan atau yang sering disebut dengan text mining. Oleh karena itu, diperlukan teknik penyaringan informasi secara berkala supaya menghasilkan informasi yang baik. Informasi yang diolah berupa klasifikasi dokumen teks Bahasa Indonesia yang diambil dari media elektronik. Salah satunya adalah penelitian terhadap dokumen Ekonomi dan Politik, bertujuan untuk mengetahui sejauh mana tingkat akurasi dan perkembangan konten-konten tersebut disediakan.

Salah satu metode klasifikasi dokumen yang dapat digunakan adalah khusus penggalian teks untuk mendeteksi informasi yang relevan dengan cara metode *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM).

Klasifikasi ditekankan untuk dokumen berbahasa Indonesia, sementara keterkaitan antar dokumen diukur berdasarkan probabilitas kemudian dibuktikan dengan algoritma klasifikasi lainnya.

Berita Bahasa Indonesia

Secara sosiologis, berita adalah semua hal yang terjadi di dunia. Dalam gambaran yang sederhana, seperti dilukiskan dengan baik oleh para pakar jurnalistik, berita adalah apa yang ditulis surat kabar, apa yang disiarkan radio, dan apa yang ditayangkan televisi. Berita menampilkan fakta, tetapi tidak setiap fakta merupakan berita. Berita pun banyak macamnya, ada yang berupa berita

politik, berita ekonomi, berita edukasi, berita kesehatan, berita sosial, berita hukum, berita budaya, bahkan berita hiburan. Pada tugas akhir ini akan membahas tentang berita Ekonomi dan Politik.

Text Mining

Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan pada text mining adalah sekumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari text mining antara lain yaitu pengkategorisasian teks (*text categorization*) dan pengelompokan teks (*text clustering*). Proses text mining meliputi proses *tokenizing, filtering, stemming, dan tagging.*

Naïve Bayes Classifier

Pada NBC setiap dokumen berita direpresentasikan dalam pasangan atribut <a1,a2 …. an> di mana a1 adalah katapertama, a2 kata kedua dan seterusnya. Sedangkan V adalah himpunan kategori berita. Pada saat klasifikasi, pendekatan Bayes akan menghasilkan label kategori yang paling tinggi probabilitasnya (vMAP) dengan masuka atribut <a1,a2,….an>.

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2 ... a_n) \quad (1)$$

Teorema Bayes menyatakan:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} \quad (2)$$

Menggunakan teorema Bayes ini, persamaan (2.1) ini dapat ditulis:

$$v_{MAP} = \arg\max_{v_j \in V} \frac{P(a_1, a_2 ... a_n \mid v_j)P(v_j)}{P(a_1, a_2 ... a_n)} \quad (3)$$

P(a1,a2 … an) nilainya konstan untuk semua vj sehingga persamaan ini dapat ditulis sebagai berikut:

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2 ... a_n \mid v_j)P(v_j) \quad (4)$$

Tingkat kesulitan menghitung P(a1, a2, …, an | vj) menjadi tinggi karena jumlah term P(a1, a2, …, an |vj) bisa jadi akan sangat besar. Ini disebabkan jumlah term tersebut sama dengan jumlah semua kombinasi posisi kata dikali dengan jumlah kategori. Naïve Bayes Classifier menyederhanakan hal ini dengan mengasumsikan bahwa di dalam setiap kategori, setiap kata independen satu sama lain. Dengan kata lain:

$$P(a_1, a_2 ... a_n \mid v_j) = \prod_i P(a_i \mid v_j) \quad (5)$$

P(vj) dan probabilitas kata wk untuk setiap kategori P(wk | vj) dihitung pada saat pelatihan.

$$P(v_j) = \frac{|docs_j|}{|Contoh|} \quad (6)$$

$$P(w_k \mid v_j) = \frac{n_k + 1}{n + |kosakata|} \quad (7)$$

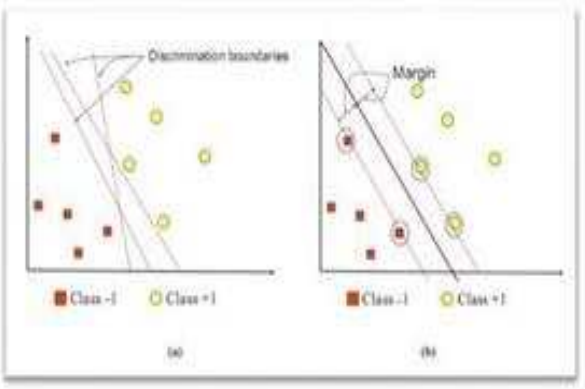
di mana | docsj | adalah jumlah kata pada kategori j dan |Contoh| adalah jumlah dokumen yang digunakan dalam pelatihan. Sedangkan nk adalah jumlah kemunculan kata wk pada kategori vj , n adalah jumlah semua kata pada kategori vj dan |kosakata| adalah jumlah kata yang unik (distinc) pada semua data latihan. Ringkasan algoritma untuk Naïve Bayes Clasifier adalah sebagai berikut:

a. Proses pelatihan. Input adalah dokumen-dokumen contoh yang telah diketahui kategorinya:
   1) Kosakata → Himpunan semua kata yang unik dari dokumen-dokumen contoh
   2) Untuk setiap kategori vj lakukan:
      a) Docsj → Himpunan dokumen-dokumen yang berada pada kategori vj
      b) Hitung P(vj) dengan persamaan 2.7
b. Untuk setiap kata wk pada kosakata lakukan:
   1) Hitung P(wk | vj) dengan persamaan 2.8
   2) Proses klasifikasi. Input adalah dokumen yang belum diketahui kategorinya:
      a) Hasilkan vmap sesuai dengan persamaan 2.6 dengan menggunakan P(vj) dan P(wk | vj) yang telah diperoleh dari pelatihan.

Support Vector Machine

Support Vektor Machine (SVM) dikembangkan oleh Boser, Guyon dan Vapnik. SVM pertama kali dipresentasikan pada tahun 1992 di A*nnual Workshop on Computational Learning Theory*. SVM merupakan *supervised learning* yang merupakan sebuah kombinasi harmonis dari teori *margin hyperplane* (Duda&Hart,1973; Cover, 1965; Vapnik, 1964) dan kernel yang diper

kenalkan oleh Aronszanjn pada tahun 1950 serta beberapa konsep pendukung yang lain. Prinsip dasasr SVM adalah linier classifier. Sedangkan pengembangan untuk masalah yang non linier dapat menambahkan kernel trick pada ruang kerja berdimensi tinggi. SVM berusaha mencari hyperplane terbaik pada input space. Hyperplane merupakan garis tengah yang memisahkan antara kelas satu dengan kelas yang lain dalam sebuah klasifikasi. Garis tengah terbaik didapatkan dengan mencari margin terbesar anatar kelas yang berbeda. Pencarian margin terbesar dapat diilustrasikan pada gambar 3 berikut. (a) menunjukkan banyak pilihan garis yang dapat memisahkan kelas -1 dengan kelas +1. Sedangkan (b) menunjukkan pilihan terbaik dengan margin terbesar.



Gambar 1. Pemisahan dua kelas (class-1 dan class+1)

Hyperplane terbaik merupakan garis tengah antara garis luar kelas-1 dan garis luar kelas+1. Sedangkan untuk kelas +1 dapat dihitung dengan rumus:

$$(W.Xi) + b \geq 1 \quad (8)$$

Sedangkan hyperplane dapat dihitung dengan rumus:

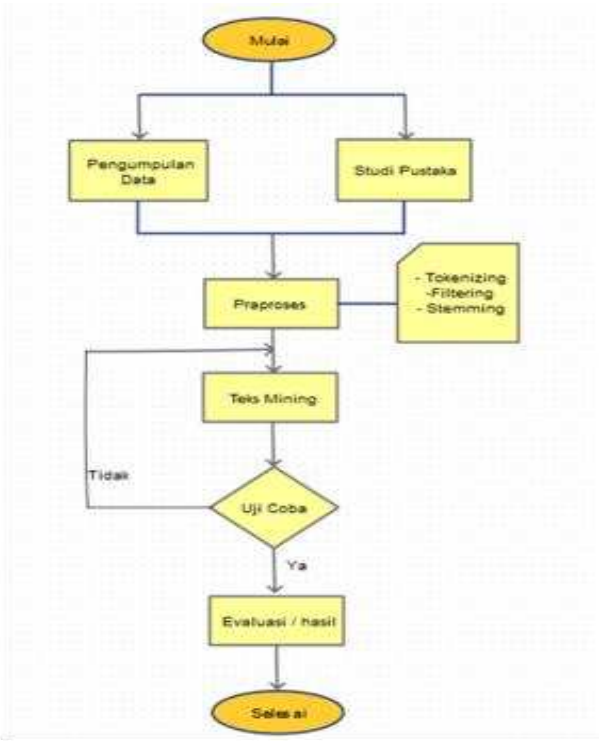
$$(W.Xi) + b = 0 \quad (9)$$

Keterangan:
W : Bobot dari sebuah atribut
Xi : Atribut ke-i
b : Bias

PENDEKATAN PEMECAHAN MASALAH

Desain Penelitian

Dalam melakukan penelitian, dibutuhkan desain penelitian agar penelitian yang dilakukan dapat berjalan dengan baik. Berikut ini merupakan desain penelitian yang digunakan pada proses pengklasifikasian dokumen teks berbahasa Indonesia dengan metode Naïve Bayes Classifer dan Support Vector Machine (svm).



Gambar 2. Kerangka Berfikir

Tahapan Penelitian

1. Studi Pustaka Studi pustaka merupakan langkah awal dalam penelitian ini, studi pustaka ini dilakukan untuk melengkapi pengetahuan dasar dan teoriteori yang digunakan dalam penelitian ini.
2. Pengumpulan Data Pada tahap pengumpulan data ini, data-data diperoleh dari Kompas.com, didapat dataset yang berupa dokumen teks atau artikel bahasa Indonesia.
3. Praproses Tahap praproses terdiri dari 3 tahapan proses, yaitu text preprocessing, yaitu pemotongan string input bedasarkan tiap kata yang menyusunnya. Kedua adalah filtering, yaitu mengurangi jumlah kata-kata yang ada, yaitu dengan penghilangan stopword. Tahap ketiga yaitu tahap pencarian kata dasar atau stemming.
4. Text mining. Pada tahapdata yang telah dikumpulkan dan diperoleh berupa dokumen teks. kemudian dilakukan pembobotan dari jumlah kemunculan didalam tiap dokumen.
5. Uji coba. Pada tahapan ini, yang pertama yaitu akan dilakukan proses Tokenizing yaitu memecahkan suatu kalimat menjadi sekumpulan kata dan pengubahan karakter huruf besar menjadi huruf kecil. Kemudian

P-ISSN: 2086 – 4981

dilakukan penghilangan stopword (kata penghubung, misalnya dan, atau, ke, di, pada, dengan, serta, dan sebagainya). Kemudian dilakukan pencarian query yang akan ditentukan dalam artikel Ekonomi dan Politik sebagai kata kunci yang akan digunakan, kemudian mencari tingkat akurasi dari proses pencarian query tersebut. hasil dari proses akan diujicobakan, apakah sesuai dengan yang ingin dicapai. Jika masih belum sesuai, maka akan diulang kembali proses tersebut.

6. Evaluasi. Hasil Pada tahap ini dilakukan pengambilan kesimpulan (berupa pola terbaik) terhadap informasi yang telah diperoleh berdasarkan hasil dari proses yang telah diujicobakan terlebih dahulu.

Tata Laksana

Kegiatan diawali dengan pengelompokkan data, untuk memperoleh metode terbaik untuk pengolahan data atau artikel berbahasa Indonesia. Proses dimulai dengan pendefinisian masalah serta mempelajari bisnis proses dari system yang sedang berjalan. Pada tahap selanjutnya melakukan proses tokenizing, penghilang stopwords, pencarian query, serta tingkat akurasinya. Metodologi penelitian ini didasarkan pada empat tahapan yang dilakukan untuk memperoleh metode mana yang terbaik untuk digunakan pada pengolahan data artikel berbahasa Indonesia. Keempat tahapan tersebut adalah a) seleksi atribut dataset, b) menangani data yang tidak konsisten, redundant dan missing value, c) proses klasifikasi dokumen d) penentuan metode terbaik yang sesuai dengan data artikel berbahasa Indonesia tentang Ekonomi dan Politik.

Pada tahap pertama, seleksi artibut dalam dataset untuk mendapatkan artibut dengan record yang relevan terhadap keluaran yang diinginkan. Pada tahap kedua pemrosesan awal data artikel berbahasa Indonesia dilakukan untuk menghapus record yang tidak konsisten, redundant dan missing value. Pada tahap ketiga adalah mengekstrak data yang akan digunakan. Dan yang kempat adalah melakukan penarikan kesimpulan untuk menentukan metode mana yang terbaik untuk data artikel berbahasa Indonesia tersebut.

HASIL DAN PEMBAHASAN

Analisis

Analisis adalah kajian yang dilakukan terhadap sebuah bahasa guna meneliti struktur bahasa tersebut secara mendalam. Sedangkan pada kegiatan, analisis merupakan analisa yang dilakukan untuk meneliti suatu penelitian.

Data

Data yang digunakan dalam penelitian ini adalah dataset yang bertipe artikel. Data artikel berbahasa Indonesia diambil dari Kompas.com. Dataset diambil dengan jumlah data yang diperoleh sebanyak 70 artikel. Adapun contoh dari salah satu artikel bahasa Indonesia yaitu :

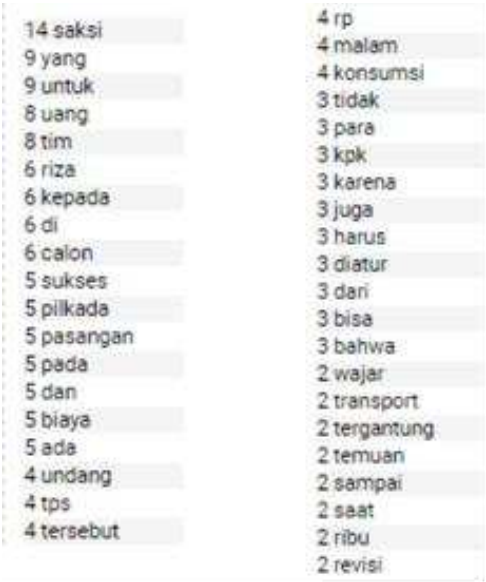
> KPK membuka temuan mereka bahwa adanya biaya untuk saksi senilai Rp 2 Miliar pada Pilkada 2015 lalu. Terkait temuan tersebut, Wakil Ketua Komisi II DPR RI Ahmad Riza Patria menganggap biaya untuk saksi pada gelaran Pilkada tergantung pada tim sukses pasangan calon. Riza menyebut ada beberapa tim pasangan calon yang memang berikan uang kepada saksi, ada juga yang lebih memilih memberikan nasi kotak kepada saksi tim pasangan calon di tiap TPS. UU," kata Riza saat dihubungi, Rabu (29/6/2016) malam. Menurut Riza, wajar bila tim pasangan calon memberi uang untuk konsumsi kepada para saksi di TPS.

Gambar 2 Contoh Artikel yang digunakan

Praproses

1. Tokenisasi

   Tokenisasi adalah proses penghilangan tanda baca pada kalimat yang ada dalam dokumen sehingga menghasilkan kata-kata yang berdiri sendiri-sendiri.



Gambar 3. Proses Tokenisasi

2. Filtering

   Tahap selanjutnya yang akan dilakukan adalah penghilangan *stopwords* yang berasal dari masing-masing artikel. *Stopwords* tersebut

berupa kata penghubung, seperti dan, dengan, ke, di, pada, serta, yang dan sebagainya.

Gambar 4. Proses Filtering

3. Penentuan pola / kata kunci

Tahap selanjutnya yang akan dilakukan adalah Penentuan pola / kata kunci. Pola /kata kunci ini didapat setelah melakukan proses tokenisasi, filtering dan stemming.

Kata	Ekonomi	Politik
kpk	6	3
biaya	1	0
miliar	1	0
pilkada	2	2
komisi	2	2
dpr	2	2
uang	2	2
konsumsi	1	0
partai	10	15

Gambar 5. Kata Kunci

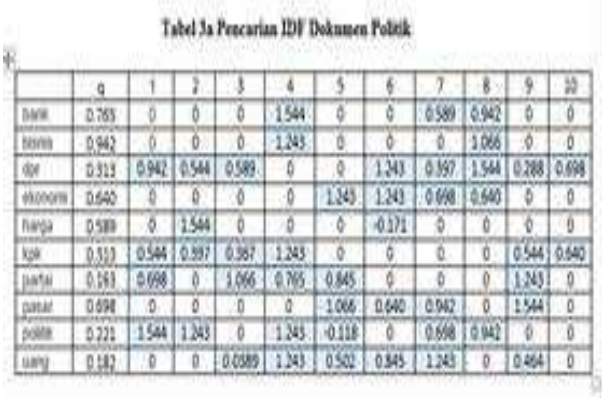
4. Perhitungan Nilai TF

Penghitungan nilai tf yaitu tahap pemilihan atau penentu query yang akan ditentukan. Kemudian dicari pada setiap artikel yang telah di kumpulkan serta dihitung, seberapa banyak query yang dihasilkan pada setiap artikel. Adapun artibut-artibut tersebut adalah kata, query, nilai tf yang terdiri dari 35 artikel berbahasa Indonesia yang bersumber dari kompas.com.

Gambar 6. Hasil pencarian TF

5. Perhitungan IDF

Penghitungan nilai idf adalah tahap yang melakukan penghitungan atau pencarian nilai bobot pada setiap artikel yang telah dikumpulkan. Seluruh artibut pada dataset tersebut selanjutnya akan dicari nilai idf-nya untuk mendapatkan artibut-artibut yang mempunyai nilai bobot dari masing-masing artikel berbahasa Indonesia, missing value, dan tidak redundant, dimana ketiga syarat tersebut merupakan syarat awal yang harus dikerjakan dalan data mining sehingga akan diperoleh dataset yang bersih untuk digunakan pada tahap mining data.



Gambar 7. Hasil IDF Dokumen Politik

P-ISSN: 2086 – 4981

Tabel 3b Pencarian IDF Dokumen Ekonomi

	q	1	2	3	4	5	6	7	8	9	10
bank	0.4649	0.502	1.0669	0	0	0	0.502	0.502	0	0	0
bisnis	0.5441	0	0	0	0	0.795	1.544	1.544	0	0	0
dpr	0.942	0	0	0	0	0	0	0	0.942	0	1.066
ekonomi	0.3136	0.368	1.0669	1.243	0	0	0.845	0.845	0	0	0
harga	0.3136	0.942	0.942	0	0	0.845	0	0	0	0.845	0
kpk	0.942	0	0	0.942	1.243	1.243	0	0	0	1.243	0
partai	0.699	0	0	0	0	0	0	0	1.544	0	1.243
pasar	0.2888	0.942	0	0.942	1.243	0.502	1.066	1.066	0	0.502	0
politik	0.642	0	0	1.243	0	0	0	0	0.544	0	0.699
uang	0.9579	0.645	1.5441	0	1.0669	0.942	0	0	0.096	0.942	0.288

Gambar 8. Hasil IDF Dokumen Ekonomi

6. Hasil Perhitungan Dengan *Naïve Bayes Classifier*

Tabel 5 Hasil Probabilitas Naïve Bayes

Kata	Probabilitas Politik	Persentase	Probabilitas Ekonomi	Presentase
bank	0.027	2.7%	0.047	4.7%
bisnis	0.028	2.8%	0.038	3.8%
dpr	0.027	2.7%	0.014	1.4%
ekonomi	0.035	3.7%	0.045	4.5%
harga	0.097	9.7%	0.033	3.3%
kpk	0.131	13.1%	0.019	1.9%
partai	0.042	0.042%	0.007	0.7%
Pasar	0.03	3.0%	0.065	6.5%
Politik	0.071	7.1%	0.035	3.5%
Uang	0.115	11.5%	0.125	12.5%

Gambar 9. Hasil Probabilitas dg Naïve Bayes

Dari tabel di atas dapat diketahui bahwa, kata "partai" dengan hasil 0.042 pada dokumen politik memiliki hasil yang rendah kemudian pada dokumen Ekonomi memiliki hasil probabilitas 0.007. Kata "kpk" pada dokumen politik memiliki hasil yang tinggi yaitu 0.131 dan kata "uang" pada dokumen ekonomi juga memiliki tingkat probabilitas yang tinggi yaitu 0.125. maka dapat disimpulkan kata apa saja yang memiliki tingkat akurasi yang tinggi dan rendah dengan menggunakan perhitungan probabilitas *Naïve Bayes Classifier* (NBC)

7. Implementasi dalam Perhitungan *Support Vector Machine* (SVM)

Tabel 6 Jumlah Data Ekonomi dan Politik

Kata	Jumlah Kata Politik	Jumlah Kata Ekonomi
Bank	14	26
bisnis	5	21
Dpr	14	7
ekonomi	19	25
harga	54	18
Kpk	73	10
Partai	23	3
Pasar	16	36
Politik	39	19
Uang	64	30

Gambar 10. Jumlah Data Ekonomi dan Politik

8. Hasil Perhitungan Dengan SVM

Tabel Data dan Jumlah

Gambar 11. Hasil Perhitungan dg SVM

Dari seluruh perhitungan di atas maka didapatkan hasil dari L adalah kata dari index ke 7 yaitu "pasar" kemudian dapat disimpulakan bahwa kata yang sudah dihitung probabilitanya dengan *Naive Bayes Classifier* dapat diklasifikasikan dalam 2 dokumen yaitu politik dan ekonomi dengan hasil:

Tabel Kesimpulan

Gambar 12. Kesimpulan Perhitungan NBC

Kesimpulan dari pengujian klasifikasi kata dengan *Support Vector Machine* (SVM) kata "harga" dan "kpk" setelah dihitung dengan probabilitas kemudian diuji dengan klasifikasi dokumen oleh svm berada pada kedua dokumen ekonomi dan politik dengan hasil 73.5 dan 131.5 pada dokumen Politik dan 44.5 dan 30 pada dokumen Ekonomi.

KESIMPULAN

Adapun kesimpulan dari penelitian ini adalah sebagai berikut :

1. Kata "partai" baik dalam dokumen Politik atau Ekonomi memiliki hasil probabilitas paling kecil. Kemudian kata "kpk" pada dokumen politik dan kata "uang" pada dokumen ekonomi adalah kata-kata terbanyak dalam pencarian probabilitas dengan *Naïve Bayes Classifier* (NBC) yaitu kata "partai" dengan hasil 0.042 pada dokumen politik, kemudian pada dokumen Ekonomi memiliki hasil probabilitas 0.007. Kata "kpk" pada dokumen politik

memiliki hasil yang tinggi yaitu 0.131 dan kata "uang" pada dokumen ekonomi 0.125.

2. Dari pengujian klasifikasi kata *dengan Support Vector Machine* (SVM) kata "harga" dan "kpk" setelah dihitung dengan probabilitas kemudian diuji dengan klasifikasi dokumen oleh SVM berada pada kedua dokumen ekonomi dan politik dengan hasil 73.5 dan 131.5 pada dokumen Politik dan 44.5 dan 30 pada dokumen Ekonomi.

SARAN

Adapun saran yang dapat diberikan terkait kesimpulan diatas adalah Penambahan data, karena data yang digunakan masih terbilang cukup sedikit. Sehingga untuk penelitian selanjutnya dapat ditambahkan data artikel Bahasa Indonesia lebih banyak lagi.

DAFTAR PUSTAKA

[1] Imam Kurniawan & Ajib Susanto, 2019, Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019*, Jurnal Eksplora Informatika,* Vol. 9 No.1, 2019.

[2] Andini, S 2013, Klasifikasi dokumen teks menggunakan algoritma naïve bayes dengan bahasa pemrograman java, *Jurnal Teknologi Informasi & Pendidikan*, vol.6, no.2, pp. 140-147, 2012.

[3] Darujati, C 2010, Perbandingan klasifikasi dokumen teks menggunakan metode naïve bayes dengan k-nearest, *Jurnal Link*, vol. 13, no. 1, pp. 2-1-2-9, 2010.

[4] Darujati, C & Bimo, A 2012, Pemanfaatan teknik supervised untuk klasifikasi teks bahasa indonesia, *Jurnal Link*, vol.16 no.1, pp. 1-8, 2012.

[5] Hamzah, A 2012, Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita Dan Abstract Akademis, Prosiding Seminar Aplikasi Sains & Teknologi (snast) Periode III, pp. B-269-B-277, 2012.

[6] Samodra, J, Sumpeno, S & Hariadi, M, Klasifikasi dokumen teks berbahasa Indonesia dengan menggunakan naïve bayes, Seminar Nasional Electrical, Informatic, And Its Education, hlm. B-1-71-B-1-74, 2009.

[7] Jiawei Han, Kamber, Pei, *Data Mining Concepts and Techniques*, 3rd Edition, Elsevier Inc, 2012.

[8] Hermawati, FA, *Data mining*, Yogyakarta, Andi, 2013.

[9] Kusrini & Luthfi, ET, *Algoritma data mining*, Yogyakarta, Andi, 2009.

[10] Prasetyo, E, *Data Mining Konsep Dan Aplikasi Menggunakan Matlab*, Yogyakarta, Andi, 2012.

P-ISSN: 2086 – 4981