PEMANFAATAN TEKNIK *SUPERVISED* UNTUK KLASIFIKASI TEKS BAHASA INDONESIA

Disusun guna memenuhi tugas mata kuliah data mining

Dosen Pengampu:

Charles Eferaim Mongi S.Si, M.Si

Oleh:

Alfa Sean Kalapadang Lonteng 20101106067

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SAM RATULANGI MANADO

2023

KATA PENGANTAR

Puji syukur saya panjatkan kepada Tuhan Yang Maha Esa, karena atas limpahan rahmatnya penyusun dapat menyelesaikan makalah ini tepat waktu tanpa ada halangan yang berarti dan sesuai dengan harapan.

Ucapan terima kasih saya sampaikan kepada bapak Charles Eferaim Mongi S.Si, M.Si sebagai dosen pengampu mata kuliah Data Mining yang telah membantu memberikan arahan dan pemahaman dalam penyusunan makalah ini.

Saya menyadari bahwa dalam penyusunan makalah ini masih banyak kekurangan karena keterbatasan. Maka dari itu penyusun sangat mengharapkan kritik dan saran untuk menyempurnakan makalah ini. Semoga apa yang ditulis dapat bermanfaat bagi semua pihak yang membutuhkan.

Manado, 20 Juni 2023

Alfa Sean

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan informasi global menuntut penyediaan informasi tersebut dapat dinikmati/dirasakan secara cepat dan tepat. Informasi yang diinginkan dapat diakomodasi oleh teknologi komputer khususnya internet. Karena internet-lah yang menjadi acuan utama beberapa penelitian mengenai penambangan data berbasis teks dilakukan atau yang sering disebut dengan text mining. Seringkali pada web, dimana kita mencari suatu informasi tertentu, banyak hal yang penting justru terlewatkan, malah yang tidak penting banyak terserap. Untuk mengatasi gap tersebut, salah satu teknik text mining adalah dengan mengklasifikasikan teks tersebut sesuai dengan karakteristik, fitur, maupun kelasnya berdasarkan aturan baku bahasa yang akan diolah, dalam penelitian ini bahasa Indonesia yang digunakan sebagai sumber acuan.

Riset mengenai pemrosesan teks sebenarnya telah lama dilakukan, untuk peringkasan teks misalnya, telah mulai diteliti sejak tahun 1958 oleh peneliti dari IBM. Meredup di tahun 70-80 dan kembali bergairah di akhir tahun 90-an sampai sekarang. Internet menjadikan pemrosesan teks kembali bangkit. Jumlah dokumen teks yang ada di internet tumbuh dengan sangat pesat. Menurut riset dari Barkeley, ukuran internet di tahun 2002 mencapai 532,897 Terabytes dengan sekitar 41.7%-nya adalah teks (dan ini berupa teks bukan multimedia). Dokumen teks ini dapat berupa static page, dynamic page, file dokumen, email, forum online dan blog. Dokumen teks juga semakin

Berperan sejalan munculnya web 2.0 yang mendorong pengguna internet untuk membuat dan berbagi content (dua yang paling terkenal: blog dan social network). Aliran content segar dengan volume besar per harinya membanjiri internet. Volume yang besar membuat pengguna internet semakin sulit memperoleh informasi yang

sesuai dengan apa yang diinginkan. Oleh karenanya dibutuhkan teknik tertentu untuk mengolah dokumen teks. Inilah fungsi dari pengolahan teks (text processing). Hasil pencarian yang dilakukan oleh mesin pencari didasarkan pada algoritma tertentu yang membaca isi atau deskripsi tentang sumber informasi. Dengan demikian, penentuan keabsahan suatu sumber merupakan keahlian tersendiri yang harus dimiliki oleh pengguna. Di pihak lain, perpustakaan yang juga merupakan penyedia sumber informasi senantiasa mengelola sumber informasi dengan melakukan klasifikasi. Klasifikasi ini membantu pengguna untuk mengalokasi sumber informasi secara fisik dan mendapatkan informasi tentang sumber informasi tersebut secara sederhana.

Didasari alternatif tersebut, maka dalam penelitian ini akan dibangun suatu aplikasi perangkat lunak yang dapat melakukan klasifikasi data teks terhadap sumber informasi teks elektronik yang diunggah secara terpandu dan selektif. Metode yang digunakan untuk mendukung proses klasifikasi ini adalah Naïve-Bayes, dan TF-IDF. Klasifikasi yang dilakukan berdasarkan 3 (tiga) kelas yang ditentukan, yaitu komputer teknologi, kesehatan dan olahraga.

1.2 Rumusan Masalah

Rumusan masalah sebagai berikut :

a. Bagaimana kerangka kerja penelitian dari pemanfaatan teknik supervised untuk klasifikasi teks bahasa Indonesia?
b. Bagaimana desain ekperimen dan analisa percobaan pada penelitian ini?

1.3 Tujuan Penulisan

Tujuan penulisan sebagai berikut :

a. Mengetahui kerangka kerja penelitian dari pemanfaatan teknik supervised untuk klasifikasi teks bahasa Indonesia

b. Memahami desain ekperimen dan analisa percobaan yang terdapat pada penelitian ini

BAB II
PEMBAHASAN

Penggolongan Teks

Pengklasifikasian teks sangat dibutuhkan dalam berbagai macam aplikasi, terutama aplikasi yang jumlah dokumennya bertambah dengan cepat. Ada dua cara dalam penggolongan teks, yaitu clustering teks dan klasifikasi teks. Clustering teks berhubungan dengan menemukan sebuah struktur kelompok yang belum kelihatan (tak terpandu atau unsupervised) dari sekumpulan dokumen. Sedangkan pengklasifikasian teks dapat dianggap sebagai proses untuk membentuk golongan-golongan (kelas-kelas) dari dokumen berdasarkan pada kelas kelompok yang sudah diketahui sebelumnya (terpandu atau supervised). Beberapa cara pada pengolahan teks antara lain :

- Information retrieval : pencarian dokumen
- Klasifikasi dokumen : membagi dokumen ke dalam kelas-kelas yang telah ditentukan sebelumnya. Misalnya secara otomatis dapat menentukan apakah dokumen ini masuk ke dalam kategori politik, ekonomi, militer dan lain sebagainya.
- Document Clustering : mirip dengan klasifikasi dokumen, hanya saja kelas dokumen tidak ditentukan sebelumnya. Misalnya berita tentang lalulintas dapat menjadi satu kelas dengan berita tentang kriminal karena didalamnya banyak memuat tentang orang yang tewas, cedera, rumah sakit.
- Peringkasan teks : Menghasilkan ringkasan suatu dokumen secara otomatis.
- Ekstraksi informasi. Mengekstrak informasi yang dianggap penting dari suatu dokumen. Misalnya pada dokumen lowongan, walaupun memiliki format beragam dapat diekstrak secara otomatis job title, tingkat pendidikan, penguasaan Bahasa.

Pengklasifikasian Teks

Banyak metode yang dapat digunakan untuk pengklasifikasian teks [Yang, 1999], antara lain adalah Na¨ıve Bayes [Lewis, 1998], k-nearest neighbor [Yavuz, 1998], Support Vector Machines (SVM), boosting, algoritma pembelajaran aturan (rule learning algorithms) dan Maximum Entropy (MaxEnt). Dalam makalah ini mengggunakan dua metode yaitu : Naïve Bayes dan k-Nearest Neighbor. Metode Naïve Bayes dikenal dengan algoritma klasifikasi simple Bayesian [Dai, 1997]. Algoritma ini banyak digunakan karena terbukti efektif untuk kategorisasi teks, sederhana, cepat dan akurasi tinggi. Klasifikasi atau kategorisasi teks merupakan suatu proses penempatan suatu dokumen ke suatu kategori atau kelas sesuai dengan karakteristik dari dokumen tersebut. Dalam text mining, klasifikasi mengacu kepada aktifitas menganalisis atau mempelajari himpunan dokumen teks pre-classified untuk memperoleh suatu model atau fungsi yang dapat digunakan untuk mengelompokkan dokumen teks lain yang belum diketahui kelasnya ke dalam satu atau lebih kelas kelas pre-defined tersebut [28, 29, 30].

Dokumen yang digunakan untuk pembelajaran dinamakan contoh (sample atau training data set) yang dideskripsikan oleh himpunan atribut atau variabel. Salah satu atribut mendeskripsikan kelas yang diikuti oleh suatu contoh, hingga disebut atribut kelas. Atribut lain dinamakan atribut independen atau predictor. Klasifikasi termasuk pembelajaran jenis supervised learning. Jenis lain adalah unsupervised learning atau dikenal sebagai clustering. Pada supervised learning, data latihan mengandung pasangan data input (biasanya vektor) dan output yang diharapkan, sedangkan pada unsupervised learning belum terdapat target output yang harus diperoleh. Proses klasifikasi teks dapat dibagi ke dalam dua fase, yaitu [31] :

- se information retrieval (IR) untuk mendapatkan data numerik dari dokumen teks. Langkah pertama yang dilakukan pada fase ini adalah feature extraction. Pendekatan yang umum digunakan adalah distribusi frekuensi kata. Nilai numerik yang diperoleh dapat berupa berapa kali suatu kata muncul di dalam

dokumen, 1 jika kata ada di dalam dokumen atau 0 jika tidak ada (biner), atau jumlah kemunculan kata pada awal dokumen. Feature yang diperoleh dapat direduksi agar dimensi vektor menjadi lebih kecil. Beberapa pendakatan feature reduction dapat diterapkan seperti menghapus stop-words, stemming, statistical filtering. Teknik lebih lanjut seperti SVD dan genetic algoritm akan menghasilkan vektor berdimensi lebih rendah.

- se klasifikasi utama ketika suatu algoritma memroses data numerik tersebut untuk memutuskan ke kategori mana teks ditempatkan. Terdapat beberapa algoritma klasifikasi yang merupakan kajian di bidang statistika dan machine learning yang dapat diterapkan pada fase ini, di antaranya adalah Naive Bayesian, Rocchio, Decision Tree, k- Nearest Neighbor, Neural Network, dan Support Vector Machines. Teknik-teknik tersebut berbeda dalam mekanisme pembelajaran dan representasi model yang dipelajari [29].

Klasifikasi–klasifikasi Bayes adalah klasifikasi statistik yang dapat memprediksi kelas suatu anggota probabilitas. Untuk klasifikasi Bayes sederhana yang lebih dikenal sebagai naïve Bayesian Classifier dapat diasumsikan bahwa efek dari suatu nilai atribut sebuah kelas yang diberikan adalah bebas dari atribut-atribut lain. Asumsi ini disebut class conditional independence yang dibuat untuk memudahkan perhitungan-perhitungan pengertian ini dianggap "naive", dalam bahasa lebih sederhana naïve itu mengasumsikan bahwa kemunculan suatu term kata dalam suatu kalimat tidak dipengaruhi kemungkinan kata-kata yang lain dalam kalimat padahal dalam kenyataanya bahwa kemungkinan kata dalam kalimat sangat dipengaruhi kemungkinan keberadaan kata-kata yang dalam kalimat [Surya, 2009]. Dalam Naïve Bayes di asumsikan prediksi atribut adalah tidak tergantung pada kelas atau tidak dipengaruhi atribut laten.

Algoritma Naive Bayes sangat berbeda dengan algoritma rule-based learning di atas. Naive Bayes adalah sebuah algoritma analisa statistik, yang bekerja dengan mengolah

data numerik [CLA05]. Metode ini menggunakan probabilitas Bayesian untuk menentukan sebuah e-mail tergolong spam atau tidak. Secara garis besar, cara kerja metode ini dapat direpresentasikan sebagai berikut:

- Ambil Probabilitas Spam dan Ham dari tiap kata,
- Hitung rata-rata Probabilitas keduanya,
- Tentukan klasifikasi berdasarkan nilai probabilitas di atas.

Tentunya untuk mendapatkan probabilitas dari tiap kata, filter harus terlebih dahulu melakukan pembelajaran terhadap setiap kata-kata dan probabilitasnya. Dalam proses pembelajaran ini, diperlukan sebuah training set, yang merupakan sekumpulan ham dan spam yang telah diklasifikasikan. Naive Bayes merupakan teknik klasifikasi yang sederhana dan cepat. Teknik ini bekerja dengan baik dengan representasi statistik. Berbeda dengan metode rule-based, Naive Bayesian dapat belajar secara incremental. Namun kekurangan dari Naive Bayesian adalah ukuran dari vektor fitur yang dihasilkan cukup besar dan butuh teknik untuk memperkecil ukuran vektor tersebut.

Indexing

Setelah dilakukan preprocessing, maka dilakukan pengindeksan kata untuk mengubah representasi data teks menjadi numerik sehingga dapat diproses [AND06]. Teknik representasi yang paling umum digunakan adalah vector space model (VSM). Pada VSM biasanya berisi bobot dari setiap kata yang dihitung berdasarkan dua pendekatan utama yaitu :

- Semakin sering sebuah kata muncul di suatu dokumen, semakin relevan kata tersebut dalam merepresentasikan topik dokumen tersebut.
- Semakin sering sebuah kata muncul di semua dokumen dalam koleksi, semakin tidak efektif dalam membedakan satu dokumen dengan dokumen lainnya.

Pendekatan pertama kebanyakan dipakai dalam konteks klasifikasi dokumen, sedangkan pendekatan kedua biasanya dipakai dalam pemrosesan query. Setelah melihat dua tipe pendekatan pembobotan di atas, maka berikut akan dijelaskan mengenai berbagai macam teknik pembobotan teks yang sering digunakan dalam pemrosesan teks. Keterangan notasi yang digunakan pada penjelasan teknik pembobotan di bawah :

aik = bobot kata i pada dokumen k

fik = frekuensi kata i pada dokumen k

N = jumlah dokumen yang ada pada koleksi

M = jumlah seluruh kata yang ada dalam koleksi

ni = jumlah kemunculan kata i pada seluruh dokumen

Boolean Weighting

Merupakan teknik yang paling sederhana, karena hanya memperhitungkan hadir tidaknya suatu kata dalam dokumen. Nilai bobot 1 bila kata tersebut muncul pada dokumen, dan 0 bila tidak muncul.

Word Frequency Weighting

Pembobotan dengan cara menghitung frekuensi kata tersebut dalam dokumen. Kata diubah jadi huruf kecil semua atau kapital semua.

tf-idf Weighting

Berbeda dengan dua teknik sebelumnya yang tidak memperhitungkan frekuensi kemunculan kata di semua dokumen, teknik tf-idf memperhitungkan frekuensi kemunculan kata di seluruh dokumen.

ltc Weighting

Yang membedakan ltc weighting dengan tfc weighting adalah ltc weighting menggunakan nilai logaritma dari frekuensi kata, bukan nilai mentah frekuensi kata tersebut. Hal ini mengurangi perbedaan nilai yang cukup besar yang terjadi pada nilai mentah frekuensi

Entropy Weighting

Merupakan teknik pembobotan yang paling baik dibandingkan 5 teknik yang lain. Dalam penelitian yang dilakukan [DUM91], terbukti 40% lebih baik daripada 5 teknik yang lain.

2.1 Kerangka Kerja Penelitian

Pada penelitian ini, himpunan data yang akan diuji adalah kumpulan artikel-artikel yang disadur dari majalah CHIP serta dibagi menurut kelas-kelasnya. Kelas-kelas yang dimaksud adalah pengkategorian dari tiap jenis artikel yang disesuaikan dengan pengkategorian artikel di dalam majalah CHIP, sehingga bisa dibedakan menjadi 5 kelas, yaitu :

- Komputer Teknologi
- Kesehatan, dan
- Berita (news).

Jumlah total artikel yang digunakan pada penelitian ini adalah 3000 data teks yang tersebar pada tiap-tiap kelasnya.

2.1.1 Himpunan data eksperimen

Standar ukuran untuk mengevaluasi kinerja sebuah algoritma dalam pengkategorian teks antara lain adalah recall dan precision. Ukuran untuk mengevaluasi kinerja yang digunakan pada eksperimen adalah accuracy. Accuracy merupakan jumlah rata-rata dari hasil recall pada tiap kelasnya.

2.1.2 Perancangan Aplikasi

Pada perancangan aplikasi pengklasifikasian berita ini akan dijelaskan mengenai rancangan aplikasi yang akan dikerjakan serta fitur-fitur yang akan dipakai pada aplikasi tersebut. Objek dari penelitian ini yaitu teks berita dimana data latih maupun data uji terdiri dari judul berita, teras berita dan tubuh berita. Hal ini akan menjadi satu kesatuan dalam pemrosesannya. Pada proses pembentukan pengetahuan maupun klasifikasi akan melewati proses

text mining yang memiliki 3 tahapan, yaitu text preprocessing, text transformation, dan pattern discovery.

2.1.3 Perancangan Text Preprocessing

Kata menjadi elemen penting bagi pelaksanaan proses pembangunan pengetahuan dan proses klasifikasi. Pada penelitian ini akan digunakan definisi kata dari Porter Stemmer, yaitu kata sebagai kumpulan huruf alfabetik sedangkan tanda baca, angka dan karakter selain huruf dianggap sebagai delimiter atau pemisah antara kata. Pada preprocessing, langkah-langkah yang akan dilakukan adalah case folding yaitu mengubah semua huruf dalam teks menjadi huruf kecil. Kemudian dilakukan proses parsing [MUS09].

2.2 Desain Eksperimen dan Analisa Percobaan

Eksperimen yang dilakukan adalah melihat kinerja dari Algoritma klasifikasi dokumen teks yaitu algoritma Naïve Bayes. Pengujian dilakukan validasi silang (cross validation) sebanyak 10 kali (10 folds validation), yaitu dengan membagi data uji menjadi 10 sub samples, Untuk rasio data uji dimulai dari 10%, naik 10% setiap kali uji sampai dengan 90%. Tiap rasio dilakukan 10 kali pengujian dan output yang diinginkan adalah accuracy rata- ratanya.

2.2.1 Analisa Hasil Percobaan

Pada bagian ini akan diujicobakan untuk mengolah teks disertai dengan penggunaan Stopword Removal dalam algoritma serta seberapa besar pengaruh penggunaan stopword tersebut.

2.2.2 Pengujian Metode

Pengujian klasifikasi teks dengan NB akan dilatih terlebih dahulu sebelum dimasukkan stopword kedalam proses klasifikasi.

2.2.3 Implementasi Algoritma Naïve Bayes

Hasil implementasi Algoritma Naive Bayes pada dokumen teks sebagaimana adanya (tanpa stop word removal). Akurasi terbesar terjadi pada data pelatihan (training sample) mencapai 70% dengan nilai accuracy 87.45%.

2.2.4 Penggunaan Stopword Removal Dalam Algoritma Naïve Bayes

Stopword removal pada klasifikasi teks menggunakan metode Naïve Bayes, tentu saja hasil akurasi yang diperoleh berbeda jika dibandingkan tanpa menggunakan stop word.

2.2.5 Pengaruh Stopword Removal Dalam Kinerja Algoritma Naïve Bayes

Penggunaan stopword hanya berdampak sangat kecil pada kinerja/accuracy. (sehingga diagram terlihat berimpit) Terlihat accuracy terbesar sebesar 74,2% sama-sama diperoleh dengan menggunakan stopword maupun tidak. Keduanya memperoleh kinerja terbesar saat data pelatihan mencapai 90 %.

BAB III
PENUTUP

3.1 Kesimpulan

Setelah algoritma Naïve Bayes diimplementasikan dalam pengklasifikasian dokumen teks, ternyata penggunaan stopword hanya berdampak kecil. Dari algoritma tersebut kinerja terbaik diperoleh jika tanpa menggunakan stopword. Kesimpulan yang diperoleh tentang klasifikasi dokumen berita bahasa Indonesia yaitu :

- Penerapan metode naive bayes classifier dalam klasifikasi berita memiliki akurasi yang baik terbukti pada data uji yang bersumber dari situs web menghasilkan nilai akurasi dengan persentase yang tinggi yaitu lebih dari 87 % untuk data latih yang besar (100 artikel).
- Dari percobaan yang telah dilakukan, klasifikasi dapat berjalan cukup baik pada data latih lebih dari 150 dokumen. Terbukti pada data latih lebih dari 150 dokumen akurasi mencapai 90 %.
- Akurasi sistem semakin tinggi dengan meningkatnya data latih yang digunakan dalam pembelajaran.

3.2 Saran

Sebagai langkah perbaikan untuk penelitian yang akan datang, dapat dipertimbangkan beberapa hal antara lain :

- Perlu ada penelitian, apakah kata keterangan, sambung, depan memang dapat diganti dengan kata tugas. Tentunya berdasarkan kinerja untuk task tertentu, misalnya untuk summmarization atau information retrieval.
- Penambahan stopword list dapat lebih memuat banyak kata.
- Teknik parsing dapat lebih diperbarui untuk pemenggalan suku kata terhadap kata-kata yang baru

DAFTAR PUSTAKA

Fabrizio Sebastiani and Consiglio Nazionale Delle Ricerche. Machine learning in automated text categorization. ACM Computing Surveys, 34:1–47, 2002.

David D. Lewis. Naive (bayes) at forty: The independence assumption in Information retrieval. pages 4–15. Springer Verlag, 1998.

Yiming Yang. An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1:67–88, 1999.

Tuba Yavuz and H. Altay Guvenir. Application of k- Nearest Neighbor on Fearure Projection Classifier to Text Categorization, 1998

Wenyuan Dai, et all. Transferring Naïve Bayes Classifiers for Text Classifications, 1997

Ali Ridho Barakbah, Instance base learning (Nearest Neighbor)

Apte, C., Damerau, F.,Weiss, S. Automated learning of decession rules for text categorization. ACM Transcations on Information Systems, 12(3),233,25, 1994

Keraf, Goris (1984), "Tatabahasa Indonesia", Nusa Indah.

Kosasih, E (2004), "Kompetensi Ketatabahasaan dan Kususastraan", Yrama Widya, Cetakan 2.

Cover**, T.M. and Hart, P.E**.: Nearest neighbor pattern Classification, IEEE Trans. Inf. Theory, Vol.IT-13, No.1, pp.21–27, 1967

Fukunaga, K.: Bias of nearest neighbor error estimation, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.PAMI-9, No.1, pp.103–112, 1987

The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer)

Ian Witten, "Data Mining Practical Machine Learning Tools and Techniques".

Nils J. Nilsson, "Introduction to Machine Learning"; Ville Kyrki, Pattern Recognition

Penerapan teknik ..., **Johanes Andria**, Fasilkom UI, 2006

Bayu Distiawan Trisedya dan Hardinal Jais, Klasifikasi Dokumen Menggunakan Algoritma Naïve Bayes dengan Penambahan Parameter Probabilitas Parent Category, Laporan Fasilkom UI, 2009.

Musthafa,A., Klasifikasi Otomatis Dokumen Berita Kejadian Berbahasa Indonesia., 2009.

Surya Sumpeno, Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naïve Bayes, Seminar Nasional Pascasarjana, Institut Teknologi Sepuluh Nopember, 13 Agustus 2009.

Budiman, K. 2005. Dasar-Dasar Jurnalistik. Pelatihan Jurnalistik-info jawa 12-15 desember 2005.www.infojawa.org. Diakses tanggal 15 Juni 2009.

Pusat Bahasa Departemen Pendidikan Nasional. 2007. Kamus Besar Bahasa Indonesia. Jakarta : Pusat Bahasa.

Shaleh Qamaruddin.1985. Asbabun Nuzul.Bandung:Diponegoro.

Hearst, Marti. 2003. What Is Text Minning?.SIMS,UC Berkeley. http://www.sims.berkeley.edu/~hearst/text. mining.html .

Harlian, Milka. 2006. Machine Learning Text Kategorization. Austin : Universit y of Texas.

Budyatna, Muhammad. 2005. Jurnalistik Teori & Praktik. Bandung : PT. Remaja Rosdakarya.

Tala., Fadillah Z. 2003. A Study of Stemming Efects on Information Retrieval in Bahasa Indonesia. Institute for Logic, Language and Computation Universite itvan Amsterdam The Netherlands. www.illc.uva.nl/publications/ResearchRep ort/Mol-200302.text.pdf .

Santoso, Budi. 2007. Data Mining Teknik Pemanfaatan data Untuk Keperluan Bisnis. Yogyakarta : Graha Ilmu.

Cahyo, Agustinus. 2012 Pemanfaatan Teknik Supervised Untuk Klasifikasi Teks