ANALISIS SENTIMEN REVIEW KEBIJAKAN ZONASI SEKOLAH PADA PENERIMAAN SISWA BARU DENGAN TEKS BAHASA INDONESIA MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACHINE*

SENTIMENT ANALYSIS OF SCHOOL ZONATION POLICY REVIEW ON ACCEPTANCE OF NEW STUDENTS WITH INDONESIAN TEXT USING SUPPORT VECTOR MACHINE ALGORITHM

Oryza Idzaa Mahendra[1], Budhi Irawan[2], Casi Setianingsih[3]
[123]Prodi S1 Teknik Komputer, Fakultas Teknik Elektro, Universitas Telkom
1oryzaidzaamahendra@student.telkomuniversity.ac.id, [2]**budhiirawan@telkomuniversity@ac.id**, [3]**Setiacasie@telkomuniversity.ac.id**

Abstrak

Setiap tahunnya, dunia pendidikan Indonesia berusaha untuk memperbaiki sistem pendidikannya. Tetapi, tidak selalu usaha untuk memperbaiki pendidikan di Indonesia akan menghasilkan hasil baik. Salah satu contoh yang tidak memiliki hasil yang baik adalah sistem zonasi yang sudah diterapkan beberapa saat lalu. Dimana, kebijakan ini hanya menggunakan domisili siswa dalam proses penerimaan siswa baru yang pastinya akan bermasalah bagi siswa yang memiliki domisili yang cukup jauh dari sekolahnya. Hal ini membuat kebanyakan orang tua siswa resah akan sistem yang sudah diterapkan oleh pemerintah ini. Oleh karena itu, dibuatlah sistem analisis sentimen pada sosial media twitter tentang sistem zonasi sekolah dengan menggunakan metode *support vector machine.* Sistem ini akan mengklasifikasikan opini opini yang beredar di twitter, yang nantinya akan di kategorikan di beberapa kategori yaitu, positif, negatif, dan netral. Pada penelitian ini didapat Sistem ini menghasilkan akurasi sebesar 90.41%, nilai presisi sebesar 90.39%, nilai recall sebesar 90.42%, dan nilai f1 score sebesar 90.40% .

Kata kunci : Klasfikasi Teks, Support Vector Machine.

Abstract

Every year, Indonesia's education world tries to improve the education system. However, not always efforts to improve education in Indonesia will produce good results. One example that does not have good results is the zoning system that was implemented a while ago. Where, this policy only uses student domicile in the process of admitting new students which will certainly be problematic for students who have domicile quite far from their school. This makes most parents worried about the system that has been implemented by the government. Therefore, a sentiment analysis system was created from Twitter social media about the school zoning system using the support vector machine method. This system will classify opinions circulating on Twitter, which will be categorized in several categories which is, positive, negative, and neutral. The result of this final task research is 90.41% accuracy, 90.39% precission, 90.42% recall, and 90.40% f1 score.

Keywords: text classification, Support Vector Machine

1. Pendahuluan

1.1 Latar Belakang

Beberapa saat lalu, Kementrian Pendidikan dan Kebudayaan Indonesia menerapkan sistem zonasi pada PPDB (Penerimaan Peserta Didik Baru) 2019. Di mana peserta didik di haruskan untuk menempuh Pendidikan di sekolah yang memiliki radius terdekat dari domisilinya. Sistem seleksi zonasi tersebut di lakukan dengan cara pemeringkatan yang berbeda di setiap provinsinya. Akan

tetapi, faktor umum yang digunakan pada penyeleksian peserata adalah dengan jarak, nilai UN (Ujian Nasional), usia peserta didik, dan waktu mendaftar[1].

Akan tetapi, sistem tersebut tidak terlihat lancar seperti yang di harapkan. Menurut media metro.tempo.co pada tahun 2019 dilaporkan ada beberapa kasus pemalsuan data kependudukan dan prestasi sekolah yang fiktif[2]. Karena Masalah tersebut Beberapa orang mengutarakan pendapatnya tentang sistem zonasi tersebut, tetapi tidak sedikit juga orang hanya mengungkapkan rasa kesalnya pada sistem ini. Walau begitu ada juga beberapa orang yang mempunyai komentar positif pada sistem zonasi yang di terapkan oleh kemendikbud ini.

Dari permasalahan tersebut, analisis sentimen mengimplementasikan metode klasifikasi SVM (*Support Vector Machine*) atas kebijakan zonasi merupakan hal yang tepat dilakukan mengingat masyarakat mempunyai opini yang beragam tentang kebijakan tersebut. Dengan adanya tweets yang mewakilkan ekspresi dari masyarakat, hal tersebut dapat menyatakan berapa banyak orang yang memiliki opini yang positif atau negatif terhadap kebijakan ini yang diharapkan menjadi saran untuk dunia pendidikan Indonesia.

1.2 Tujuan

Adapun Tujuan penelitian ini adalah merancang sistem analisis sentiment review tentang kebijakan zonasi sekolah dan mengetahui performansi sistem analisis sentiment ulasan yang dibangun dengan menggunakan metode klasifikasi SVM.

1.3 Rumusan Masalah

Adapun perumusan masalah pada penelitian ini adalah Bagaimana sistem analisis mengetahui opini pengguna media sosial tentang kebijakan zonasi sekolah di Indonesia dan mengetahui bagaimana performansi metode klasifikasi SVM pada sistem analisis yang dibuat.

2. Dasar Teori /Material dan Metodologi/perancangan

Pada Bagian ini berisi tentang dasar teori yang digunakan untuk pembuatan sistem analisis sentiment review terhadap kebijakan zonasi sekolah. Adapun beberapa teori yang digunakan adalah sebagai berikut:

2.1 Sosial Media

Sosial Media didefiniskan sebagai aplikasi berbasis internet yang terbuat dari fondasi ideologis dan teknologi Web 2.0, dan yang memungkinkan pembuatan dan pertukaran konten yang dilakukan oleh pengguna[3]. Secara umum Media social adalah sebuah media online, dengan para penggunanya dapat berinteraksi satu sama lain, dan bahkan menciptakan suatu karya dalam suatu konten tertentu.

2.2 Text Mining

Text Mining adalah suatu proses penganalisis-an kumpulan dokumen yang mempunyai tujuan untuk memahami konten dan makna dari informasi di dalamnya. Text mining juga bertujuan untuk meningkatkan kemampuan manusia dalam memproses informasi berskala besar, dan bernilai tinggi. Text Mining dapat menyelesaikan beberapa masalah lainnya diantaranya adalah ketidakpastian sesuatu dan ambiguitas nya suatu teks dokumen [4].

2.3 Analisis Sentimen

Analisis Sentimen adalah studi yang mempelajari komputasi yang berisi pendapat, evaluasi, sikap, pandangan dan emosi yang diungkapkan dalam teks yang mewakili dari suatu Entitas atau individu. . Hal ini merujuk pada masalah klasifikasi itu sendiri dimana kita akan selalu meramalkan suatu pendapat yang di dapat lalu mengklasifikasikannya menjadi bagian positif dan negatif. Analisis Sentimen akan mengidentifikasi sentimen dari seseorang ungkapkan dan kemudian mengklasifikasikan polaritasnya[5].

2.4 Pre-processing

Pre-Processing merupakan salah satu langkah yang cukup penting dalam machine learning. Tujuan dari *Pre-Processing* tersebut adalah untuk menghilangkan dataset yang akan mengganggu sehingga memperkecil akurasi dari sebuah prediksi.

2.5 TF-IDF

TF-IDF dikenal sebagai suatu ekstraksi fitur yang digunakan untuk memberikan suatu bobot, dan performansinya bahkan masih bisa dibandingkan dengan teknik teknik baru. Dokumen akan digunakan sebagai faktor dalam pembobotan[6].

2.6 Chi-Square

Seleksi fitur dilakukan untuk mengurangi fitur fitur yang tidak dibutuhkan atau tidak relevan dalam proses klasifikasi. Chi square adalah jenis seleksi fitur yang menggunakan teori statistik untuk menguji sebuah independensi suatu term. Seleksi fitur mempunyai tujuan yaitu menghilangkan fitur pengganggu dalam proses klasifikasi.

2.7 Support vector machine

Support Vector Machine adalah suatu machine learning untuk memecahkan masalah klasifikasi dari dua kelompok. Secara konsep, SVM mengimplementasikan beberapa ide, diantaranya: vektor yang di inputkan dipetakan secara non-linear ke ruang fitur yang sangat tinggi. Dalam ruang fitur ini permukaan linear dari suatu keputusan di bangun. Properti khusus dari permukaannya memastikan kemampuan generalisasi yang tinggi dari *machine learning*[7]**.**

3. Perancangan

3.1 Data Scrapping

Dataset yang digunakan diambil dari media social twitter yang berhubungan dengan kata kunci zonasi, zonasi SD, zonasi sekolah. Dataset yang diambil berjumlah 1456 data, yang berupa 485 data berlabel positif, 485 data berlabel negative, dan 485 data berlabel netral dengan bentuk file berupa *Comma Separated Value* (CSV).

3.2 Pre-Processing

Data yang diambil sebelumnya akan diolah pada proses ini. Dalam proses ini ada beberapa Teknik yang digunakan untuk mengolah data, diantaranya:

3.2.1 Case Folding

Pada tahapan ini, program akan menyeragamkan bentuk huruf menjadi huruf kecil atau lowercase, karena data yang diambil tidak memiliki konsistensi yang sama sehingga harus diubah menjadi huruf kecil agar sama.

3.2.2 Menghapus URL

Pada tahapan ini program akan menghapus URL bertujuan agar data yang yang diambil bersih dari hypertext yang mengarah ke sebuah web, karena teks tersebut akan dianggap sebuah noise dalam data yang tidak memiliki arti apapun.

3.2.3 Menghapus Angka

Pada tahapan ini dilakukan untuk menghapus angka yang berada di data, untuk meminimalisir fitur yang tidak relevan pada data.

3.2.4 Menghapus Spesial Karakter

Setelah menghapus angka, selanjutnya adalah menghapus karakter spesial. Karakter spesial yang dimaksud adalah @, #, dan spesial karakter lainnya.

3.2.5 Menghapus Tanda Baca

Setelah menghapus karakter spesial, selanjutnya adalah menghapus tanda baca. Seperti titik(.), koma(,), tanda seru(!), dan tanda baca lainnya.

3.2.6 Melakukan Stemming

Pada proses ini berfungsi untuk mengubah kata berimbuhan menjadi suatu kata dasar atau suku kata. Proses ini bertujuan untuk meminimalisir jumlah fitur karena dengan kata yang berbeda imbuhan akan membuat jumlah fitur membengkak.

3.2.7 Menghapus stopword

Pada tahapan ini merupakan proses untuk menghilangkan kata-kata henti yang tidak terlalu berpengaruh pada data.

3.2.8 Melakukan Tokenisasi

Pada tahapan ini merupakan pemotongan Panjang string input menjadi lebih pendek, sehingga kalimat dapat di tokenisasi, bisa disimpulkan tokenisasi merupakan kumpulan atau penggalan beberapa kata yang asalnya berupa kalimat.

3.3 Ekstraksi Fitur dengan TF-IDF

Pada tahap ini, dokumen hasil dari *text pre-processing* akan diambil kembali untuk memberikan bobot setiap data. Bobot dari setiap data dihasilkan dari perkalian antara TF dan IDF dimana, TF adalah jumlah term pada dokumen dan IDF adalah jumlah term pada seluruh dokumen yang ada pada data. Selanjutnya program akan menghitung skor dari setiap kalimatnya. Lalu, data akan dilanjutkan pada seleksi fitur, yang akan memilih kembali fitur fitur yang akan digunakan. Berikut adalah rumus dari TF-IDF:

$$w_{t,d} = tf_{t,d} \times idf_t \quad (1)$$

Dimana :

$$idf_t = log \frac{N}{df_t} \quad (2)$$

Dan

$$idf_t = \log\left(\frac{N}{df_t}\right) + 1 \quad (3)$$

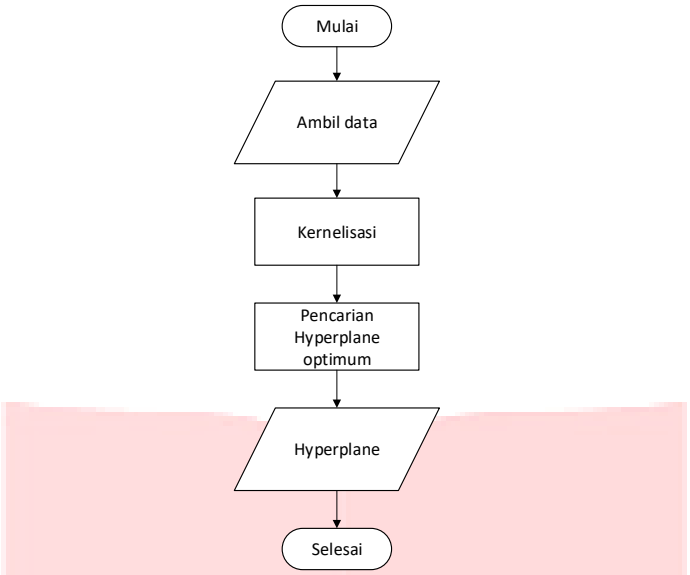
3.4 Seleksi Fitur dengan Chi Square

Pada langkah awal, data yang sudah di beri bobot oleh TF-IDF akan diberikan kelas dan atribut yang nantinya akan membentuk suatu tabel kontingensi. Tabel tersebut berguna untuk menentukan nilai chi-square. Setelah tabel tersebut terbentuk, nilai Chi square akan terlihat dengan menggunakan rumus berikut :

$$X^2 = \sum \frac{(O-E)^2}{E} \quad (4)$$

kemudian akan diurutkan fitur yang nilaninya terbesar hingga terkecil dan dibatasi dengan nilai fitu yang ingin digunakan. Jika hasil memenuhi dianggap selesai, apabila fitur yang rankingnya tidak termasuk dalam kategori fitur relevan makan akan dibuang.

3.5 Klasifikasi dengan Support Vector Machine



Gambar 3.1 Diagram alir proses klasifikasi dengan algoritma *SVM*

Pada Gambar 3.1 dijelaskan, setelah fitur sudah diseleksi selanjutnya, lalu data akan diklasifikasi dengan algoritma SVM. Penilaian score SVM akan dibuat dengan menilai score apakah positif, netral, dan negatif yang merepresentasikan di bagian mana *hyperlane* dari data tersebut berada. Berikut adalah rumus umum dari hyperplane:

$$w^T \cdot x + b = 0 \quad (4)$$

Dikarenakan data tidak selalu terpisah secara linear maka dilakukan kernelisasi dengan kernel sigmoid seperti pada rumus berikut:

$$K(x_i, x_j) = \tanh(\gamma {x_i}^T x_j + r), r \quad (5)$$

Setelah dilakukan kernelisasi, selanjutnya dilakukan pencarian support vector dengan Quadratic Programming dengan persamaan Lagrange Multipliers seperti berikut:

$$\min Ld = \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \, x_i \cdot x_j - \sum_{i=1}^{N} \alpha_i \quad (6)$$

Setelah dilakukan perhitungan dengan persamaan diatas, didapat nilai α_i sehingga dapat dicari *w* (weight) dan *b* (bias) nya sehingga didapat persamaan hyperplane :

$$w^T \cdot x + b = a^T y \, K(X_{training}, X_{testing}) + b \quad (7)$$

Sehingga membentuk persamaan f(x) sebagai berikut

$$f(x) = \boldsymbol{Sgn}(\, a^T y \; tanh(\gamma(X_{training}, X_{testing}) + r) + b) \quad (8)$$

4. Pengujian dan implementasi

4.1 Pengujian Kernel

Pengujian dilakukan untuk mengetahui hasil akurasi dengan 4 macam kernel yang berbeda dengan partisi data 50% latih dan 50% data test, kernel yang diuji antara lain adalah linear, Polynomial, Sigmoid, dan Radial Basis Function.

Tabel 4.1 Pengujian Kernel

pengujian	Kernel	Akurasi (%)
1	Linear	70.74
2	polynomial	36.26
3	Radial Basis Function	70.88

pengujian	Kernel	Akurasi (%)
4	sigmoid	74.03

Pada tabel 4.1 dapat terlihat kernel yang terbaik dengan partisi data 50% data test dan 50% data training adalah kernel sigmoid dengan hasil akurasi 74.03%.

4.2 Pengujian C dan Gamma

Setelah Kernel terbaik dipilih, pengujian selanjutnya adalah pengujian parameter C dan Gamma terbaik untuk digunakan pada pengujian selanjutnya. Pengujian dilakukan dengan menguji parameter Gamma lebih dahulu dengan interval 0.001, 0.01, 0.1, 1.0, dan 10.0 , setelah parameter gamma didapat selanjutnya dilakukan pengujian parameter C dengan interval 1,10,20,30,40,50,60,70,80,90,100. Di dapat hasil pengujian dengan gamma 0.006 dan C = 1 dengan hasil akurasi 74.31%

4.3 Pengujian Partisi Data

Pada pengujian ini memiliki tujuan untuk mengetahui kinerja algoritma SVM dalam mengklasifikasian suatu data test sesuai label yang sudah di tentukan pada dataset, pada dataset akan dibagi dengan beberapa partisi yang salah satu bagian partisinya akan menjadi data testing dan lainnya akan menjadi data training.

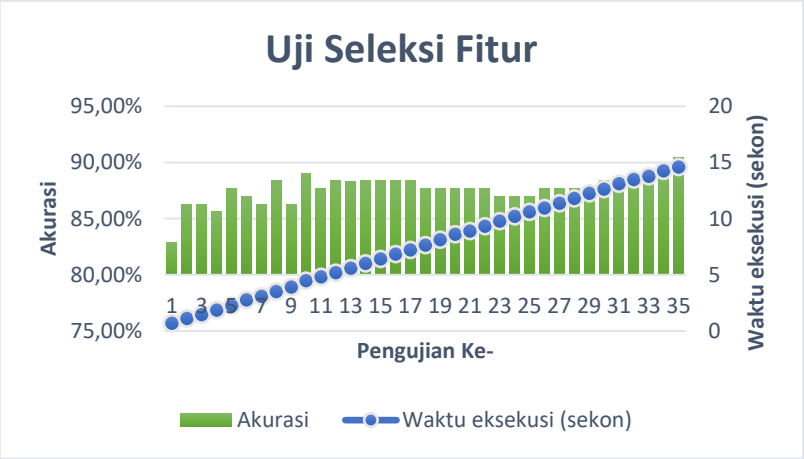
Tabel 4.2 Hasil Uji Partisi Data

Pengujian	Data Latih	Data Uji	Akurasi (%)	Presisi (%)	Recall (%)	F1 score (%)
1	50%	50%	74.3	75.5	74.3	74.2
2	60%	40%	79.0	79.8	79.0	78.9
3	70%	30%	80.5	81.7	80.5	80.4
4	80%	20%	81.4	81.7	81.4	81.5
5	90%	10%	90.4	90.5	90.4	90.4

Pada gambar 4.1 diatas, partisi yang optimal yaitu 90% data latih dan 10% data uji, dengan akurasi sebesar 90.4%, presisi 90.5%, recall 90.4%, dan f1 score 90.4%. hal ini menyatakan bahwa semakin banyak data yang dilatih maka akurasi akan semakin baik, karena hal tersebut membuat data training memiliki kata dengan variasi yang banyak untuk memprediksi suatu kelas. Hasil yang paling optimal akan digunakan pada tahap pengujian selanjutnya

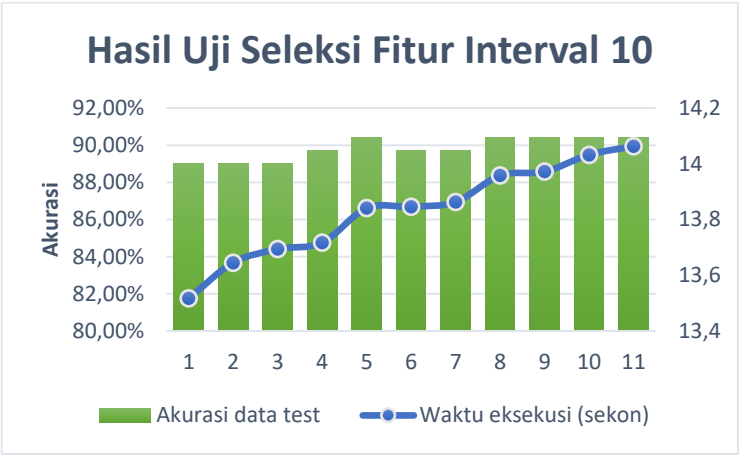
4.4 Pengujian Seleksi Fitur

Pada pengujian fitur seleksi ini, akan digunakan jumlah fitur yang berbeda setiap pengujiannya untuk mengetahui dan membandingkan akurasi dan respon waktu dari setiap pengujian, data yang digunakan adalah data dari partisi yang memiliki akurasi optimal yang diuji dalam pengujian sebelumnya yaitu 90% data training dan 10% data testing.



Gambar 4.1 Hasil Uji Fitur Seleksi

Pada gambar 4.1 tersebut terlihat pada pengujian ke 34 dengan fitur berjumlah 3400 dan akurasi sebesar 89.04% hampir mendekati akurasi terbaik pada fitur berjumlah 3491 dengan akurasi 90.41%. Pada interval tersebut juga bisa dilihat selisih waktu sebesar 0.333 detik lebih cepat diantara fitur sebesar 3400 dan 3491. Lalu diuji Kembali dengan interval diantara 3400 sampai 3491:

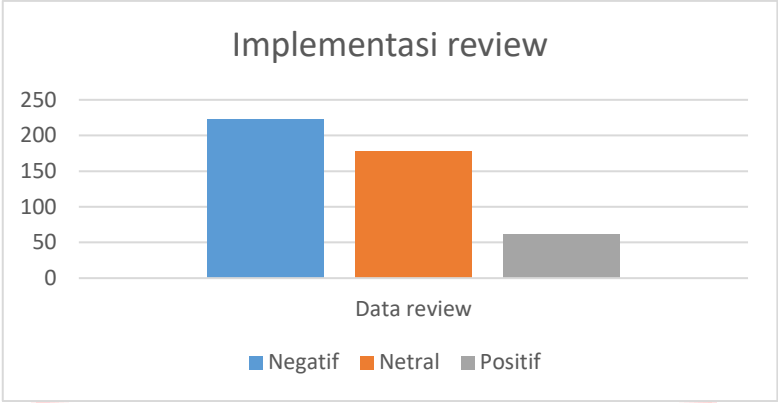


Gambar 4.2 Hasil Uji seleksi fitur interval 10

Dari gambar 4.2 terlihat pada fitur berjumlah 3440 memiliki akurasi yang sama dengan fitur yang berjumlah 3491 yaitu sebesar 90.41%, nilai presisi 90.39%, nilai recall sebesar 90.42%, dan nilai f1 score sebesar 90.40%. dan pada fitur berjumlah 3440 memiliki waktu lebih cepat dengan selisih waktu 0.222 sekon dibanding dengan fitur yang berjumlah sebanyak 3491.

4.5 Implementasi Data Uji

Untuk mengimplementasikan aplikasi yang dibuat, diambil data twitter yang bertentangan zonasi sekolah dengan parameter pencarian zonasi, zonasi sekolah, dan zonasi SD mulai dari 9 juni 2020 hingga 16 juni 2020 dengan jumlah total data sebanyak 461 data lalu kemudian dilakukan klasifikasi. Dari total 461 data yang diuji telah diklasifikasi terdapat 178 data terklasifikasi Netral, 222 data terklasifikasi Negatif, dan 61 data terklasifikasi Positif.



Gambar 4.3 Review Kebijakan Zonasi

Pada gambar 4.3 dapat diambil kesimpulan bahwa sentiment pengguna twitter terhadap zonasi sekolah dengan persentase 48.16% Negatif, 38.61 % Netral, dan 13.23% Positif, mayoritas dari pengguna twitter tidak menyukai adanya kebijakan zonasi sekolah yang diterapkan oleh Kemendikbud Indonesia.

5. Kesimpulan

Berdasarkan hasil pengujian yang sudah dilakukan, bisa didapat kesimpulan bahwa proses klasifikasi sentiment terhadap kebijakan zonasi sekolah dengan metode support vector machine dapat dilakukan dengan baik. didapat akurasi terbaik pada partisi data latih 90% dan 10% data test, fitur yang digunakan sebanyak 3440 fitur dan akurasi sebesar 90.41%, nilai presisi sebesar 90.39%, nilai recall sebesar 90.42%, dan nilai f1 score sebesar 90.40%.

Daftar Pusaka:

[1] D. Putsanra Videlia, “Memahami Sistem Zonasi Sekolah di PPDB 2019,” *tirto.id*, 2019. [Online]. Available: https://tirto.id/memahami-sistem-zonasi-sekolah-di-ppdb-2019-ecEz. [Accessed: 19-Jun-2019].

[2] J. Sugiharto, “Zonasi PPDB 2019, Kota Bogor Bongkar Pemalsuan Data Kependudukan,” *Tempo.co*, 2019. [Online]. Available: https://metro.tempo.co/read/1220408/zonasi-ppdb-2019-kota-bogor-bongkar-pemalsuan-data-kependudukan/full&view=ok. [Accessed: 02-Jul-2019].

[3] A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of Social Media,” *Bus. Horiz.*, vol. 53, no. 1, pp. 59–68, 2010.

[4] G. Shi and Y. Kong, “Advances in theories and applications of text mining,” *2009 1st Int. Conf. Inf. Sci. Eng. ICISE 2009*, pp. 4167–4170, 2009.

[5] R. Jose, “Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation,” no. November, pp. 638–641, 2015.

[6] M. Ramya and J. A. Pinakas, “Different type of feature selection for text classification,” *Int. J. Comput. Trends Technol.*, vol. 10, pp. 102–107, 2014.

[7] C. Corinna and V. Vapnik, “Support-Vector Networks,” *IEEE Expert*, vol. 7, no. 5, pp. 63–72, 1995.