ANALISIS SENTIMEN DAN PERSEPSI PUBLIK PADA KOMENTAR YOUTUBE STUDI KASUS TENTANG KAITAN ANTARA PENERIMA BEASISWA LUAR NEGERI DENGAN ISU PERPINDAHAN KEWARGANEGARAAN (ANAK JADI WNA) MENGGUNAKAN MACHINE LEARNING

Teguh Muryanto
Program Studi Teknik Informatika, FTI, Institut Teknologi Budi Utomo Jakarta
teguhmuryanto@gmail.com

Abstrak
Penelitian ini dilatarbelakangi maraknya polemik digital di platform YouTube terkait pernyataan seorang alumni penerima beasiswa negara (LPDP) mengenai anaknya yang menjadi Warga Negara Asing (WNA). Isu sensitif tersebut memicu perdebatan publik yang luas di ruang siber mengenai nilai nasionalisme, identitas kebangsaan, serta tanggung jawab moral awardee terhadap tanah air. Guna memetakan polarisasi opini masyarakat yang masif dan tidak terstruktur secara otomatis, penelitian ini menerapkan pendekatan kuantitatif dengan metode komparatif eksperimental berbasis machine learning. Data komentar dikumpulkan dari platform YouTube menggunakan YouTube Data API v3 untuk kemudian diklasifikasikan ke dalam tiga kategori sentimen utama, yaitu positif, negatif, dan netral. Dalam proses pengerjaannya, data teks mentah yang bersifat informal terlebih dahulu dilewatkan pada rangkaian tahapan text preprocessing yang ketat, meliputi cleaning, Case Folding, normalisasi kata tidak baku (slangwords), Tokenizing, stopword removal, hingga Stemming dengan bantuan pustaka Sastrawi. Ekstraksi fitur dilakukan menggunakan metode pembobotan Term Frequency-Inverse Document Frequency (TF-IDF). Tahap pelabelan data awal dijalankan dengan pendekatan leksikon (lexicon-based matching), sebelum akhirnya data diumpankan pada tiga algoritma klasifikasi yang dibandingkan, yaitu Naive Bayes, Random Forest, dan Support Vector Machine (SVM). Kinerja dari ketiga model tersebut dievaluasi secara objektif menggunakan instrumen Confusion Matrix berdasarkan metrik akurasi, presisi, recall, dan F1-Score tanpa melibatkan implementasi sistem aplikasi.
Kata Kunci : Analisis Sentimen, YouTube, LPDP, Support Vector Machine, Random Forest, Naive Bayes, TF-IDF.

1. PENDAHULUAN

Kemajuan teknologi informasi telah menjadikan media sosial, khususnya *YouTube*, sebagai ruang publik baru bagi masyarakat untuk mengekspresikan pendapat dan merespons berbagai isu aktual. Setiap unggahan video yang menyentuh isu sensitif kebangsaan berpotensi memicu ribuan bahkan jutaan komentar yang mencerminkan keberagaman opini publik. Salah satu fenomena yang baru-baru ini memicu perdebatan sengit di ruang digital adalah kasus seorang alumni penerima beasiswa Lembaga Pengelola Dana Pendidikan (LPDP) yang secara terbuka menyatakan kebanggaannya karena anaknya telah menjadi Warga Negara Asing (WNA). Pernyataan kontroversial ini tidak hanya menyoroti etika penerima beasiswa negara, tetapi juga membuka diskusi lebih luas tentang nasionalisme, identitas kebangsaan, serta hubungan kontraktual antara negara dan para *awardee* di era globalisasi.

Seiring dengan meningkatnya jumlah komentar yang membahas isu tersebut, analisis secara manual menjadi kurang efektif karena keterbatasan waktu, subjektivitas, serta potensi inkonsistensi hasil. Oleh karena itu, pendekatan berbasis *machine learning* semakin banyak digunakan dalam analisis sentimen karena mampu mengolah data teks dalam jumlah besar secara otomatis, objektif, dan terukur. Media sosial menjadi wadah yang krusial karena melaluinya data berukuran besar mengalir setiap harinya tanpa intervensi struktural yang membatasi hak berpendapat.

Dalam perkembangan teknologi informasi yang semakin pesat, analisis sentimen menjadi salah satu metode yang banyak digunakan untuk memahami opini publik terhadap suatu topik tertentu. Meningkatnya penggunaan internet dan media sosial menyebabkan bertambahnya volume data teks yang dihasilkan pengguna, seperti komentar dan diskusi daring. [1]

Meskipun penelitian analisis sentimen pada komentar *YouTube* telah banyak dilakukan, sebagian besar masih berfokus pada topik umum seperti produk, hiburan, atau peristiwa politik berskala luas. Pemanfaatan komentar *YouTube* untuk menganalisis sentimen publik terhadap isu yang mengaitkan "beasiswa dari negara" dengan fenomena "anak jadi WNA" yang memiliki sensitivitas dan dampak sosial tinggi masih relatif terbatas.

Dalam penelitian berbasis media sosial, analisis sentimen memiliki peranan penting karena mampu memberikan gambaran mengenai respons masyarakat terhadap suatu fenomena secara luas dan real-time. Media sosial menjadi sumber data utama karena menyediakan opini pengguna secara langsung dalam jumlah besar, termasuk melalui komentar pada platform YouTube [2].

Komentar *YouTube* umumnya bersifat tidak terstruktur, mengandung bahasa informal, singkatan, sarkasme, serta ekspresi emosional yang beragam, sehingga menimbulkan tantangan dalam proses klasifikasi sentimen.

Proses analisis sentimen umumnya melibatkan beberapa tahapan, yaitu pengumpulan data, *preprocessing* teks, ekstraksi fitur, serta proses klasifikasi menggunakan algoritma machine learning. Tahapan tersebut dilakukan untuk menghasilkan data yang lebih bersih dan terstruktur sehingga dapat meningkatkan akurasi proses klasifikasi [3].

Berdasarkan penelitian terdahulu, Algoritma *Naive Bayes, Random Forest,* dan *Support Vector Machine* (SVM) telah banyak digunakan dalam analisis sentimen komentar *YouTube*, namun umumnya diterapkan pada topik yang berbeda dan jarang dibandingkan dalam satu kerangka evaluasi yang seragam. Kondisi ini menunjukkan adanya celah riset berupa kurangnya studi komparatif yang secara khusus mengevaluasi kinerja ketiga algoritma tersebut pada kasus polemik yang melibatkan penerima beasiswa negara dan isu perpindahan kewarganegaraan. Kebaruan penelitian ini terletak pada penggabungan isu sensitif kebangsaan dengan analisis sentimen melalui pendekatan komparatif tiga algoritma *machine learning.*

1. METODOLOGI

1.1 Jenis Penelitian

Metode penelitian yang digunakan dalam studi ini adalah metode kuantitatif dengan pendekatan komparatif. Pendekatan komparatif dipilih untuk membandingkan kinerja tiga algoritma klasifikasi, yaitu *Naive Bayes, Random Forest,* dan *Support Vector Machine* (SVM) dalam menganalisis sentimen komentar pada video *YouTube* yang membahas isu penerima beasiswa luar negeri yang dikaitkan dengan perpindahan kewarganegaraan.

Oleh karena itu, algoritma *machine learning* seperti *Naive Bayes*, *Random Forest*, dan *Support Vector Machine* (SVM) sering digunakan dalam analisis sentimen karena terbukti mampu mengklasifikasikan teks ke dalam kategori sentimen positif, negatif, dan netral dengan tingkat akurasi yang baik [4].

Data penelitian berupa komentar pengguna yang diperoleh dari platform YouTube. Data tersebut kemudian diproses untuk mengidentifikasi tiga kategori sentimen, yaitu positif, negatif, dan netral. Melalui pendekatan ini, penelitian bertujuan untuk menentukan algoritma dengan performa terbaik dalam mengklasifikasikan sentimen, sehingga mampu menghasilkan analisis yang lebih akurat terhadap persepsi publik di media sosial.

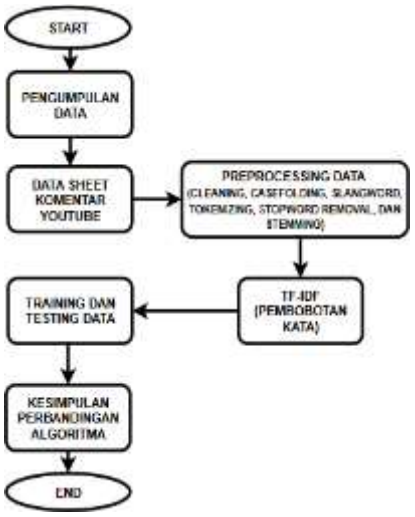
Untuk menghasilkan analisis yang akurat, penelitian ini menggabungkan metode pelabelan menggunakan kamus bahasa (*Lexicon-Based* dengan *InSet Lexicon*), serta membandingkan performa tiga algoritma sekaligus, yaitu *Naive Bayes*, *Support Vector Machine* (SVM), dan *Random Forest*.

Naive Bayes Classifier merupakan salah satu algoritma klasifikasi berbasis probabilitas yang menggunakan Teorema Bayes sebagai dasar perhitungannya. Algoritma ini bekerja dengan menghitung peluang suatu data termasuk ke dalam kelas tertentu berdasarkan karakteristik atau fitur yang dimiliki. Naive Bayes memiliki asumsi bahwa setiap fitur bersifat independen satu sama lain atau disebut sebagai asumsi naive [5].

Random Forest merupakan salah satu algoritma machine learning berbasis ensemble learning yang digunakan untuk meningkatkan performa klasifikasi dengan menggabungkan banyak pohon keputusan (decision tree). Algoritma ini bekerja dengan membangun sejumlah decision tree dari subset data yang berbeda, kemudian menggabungkan hasil prediksi dari seluruh pohon untuk menentukan hasil akhir [6]

Dengan pendekatan ini, penelitian diharapkan tidak hanya mampu mengelompokkan komentar dengan baik, tetapi juga bisa mengatasi masalah ketidakseimbangan jumlah data agar hasil akurasinya menjadi lebih maksimal.

1.2 Kerangka Pemikiran



Gambar 1: Kerangka Berpikir
Sumber: Penelitian mandiri

Berikut adalah uraian alur kerangka penelitian yang berfokus pada analisis sentimen komentar YouTube menggunakan algoritma *Naïve Bayes*, *Support Vector Machine* (SVM), dan *Random Forest*:

a. Pengumpulan Data (*Start* Pengumpulan data)Penelitian dimulai dengan tahap pengambilan data, yaitu proses mengumpulkan opini atau komentar dari platform YouTube yang berkaitan dengan topik beasiswa LPDP.
b. Data Sheet Komentar YouTube Komentar-komentar yang telah berhasil diperoleh melalui proses *crawling* kemudian dikumpulkan, ditata, dan disimpan dalam bentuk lembar data (*data sheet*) mentah yang siap untuk masuk ke tahap pemrosesan.
c. *Preprocessing* Data Langkah selanjutnya adalah pra-pemrosesan data (*data preprocessing*). Tahap ini bertujuan untuk membersihkan dan mengubah data teks mentah agar menjadi lebih terstruktur. Proses *preprocessing* yang dilakukan meliputi *cleaning* (menghapus tautan, simbol, dan karakter yang tidak diperlukan), *case folding* (mengubah semua huruf menjadi huruf kecil), *slangword* (mengubah kata tidak baku/gaul menjadi kata baku), *tokenizing* (memotong string teks menjadi potongan kata), *stopword removal* (membuang kata yang tidak memiliki makna penting), serta *stemming* (mengubah kata berimbuhan menjadi kata dasar bahasa Indonesia).
d. TF-IDF (Pembobotan Kata) Setelah data teks bersih, dilakukan proses pembobotan kata menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF). Metode ini berfungsi untuk mengubah data teks (*string*) menjadi representasi vektor numerik berdasarkan bobot kepentingan suatu kata di dalam dokumen, sehingga data dapat dipahami oleh algoritma *machine learning*.
e. *Training* dan *Testing* Data Tahap berikutnya adalah proses pelatihan (*training*) dan pengujian (*testing*) model menggunakan data yang telah melalui pembobotan. Data akan dibagi menjadi data latih dan data uji untuk diterapkan pada tiga algoritma yang dibandingkan, yaitu *Naïve Bayes Classifier*, *Support Vector Machine* (SVM), dan *Random Forest*. Proses pengujian juga mencakup evaluasi performa model menggunakan metrik *accuracy, precision, recall,* dan *F1-score*.
f. Kesimpulan Perbandingan Algoritma (*End*) Tahap terakhir adalah penarikan kesimpulan berdasarkan hasil perbandingan performa ketiga algoritma yang diuji. Pada tahap ini, dilakukan analisis untuk menentukan algoritma mana yang memiliki tingkat akurasi dan kinerja terbaik dalam mengklasifikasikan sentimen positif dan negatif pada komentar YouTube terkait LPDP.

1.3 Objek dan Metode Pengumpulan Data

Objek dalam penelitian ini adalah komentar pengguna *YouTube* pada video yang membahas kontroversi alumni penerima beasiswa luar negeri (LPDP) terkait perubahan kewarganegaraan anaknya menjadi WNA. Metode pengumpulan data dilakukan secara otomatis melalui metode data *crawling* memanfaatkan *YouTube* Data API v3 melalui parameter *comment Threads* dan properti video Id yang kemudian diekspor ke dalam format .csv.

2.4. Metode Analisis Data (*Text Preprocessing)*

Tahapan pra-pemrosesan teks yang dilewati secara berurutan meliputi :

a. ***Cleaning*:** Menghapus tautan URL, *mention, hashtag*, angka, dan seluruh tanda baca.
b. ***Case Folding*:** Mengubah seluruh teks alfabet menjadi huruf kecil.
c. ***Slangword Normalization*:** Mengonversi kata tidak baku/gaul menjadi kata formal.
d. ***Tokenizing*:** Memotong kalimat menjadi token kata mandiri berdasarkan spasi.
e. ***Stopword Removal:*** Menghapus kata umum penghubung menggunakan pustaka Sastrawi.
f. ***Stemming* :** Mengubah kata berimbuhan menjadi kata dasar bahasa Indonesia.

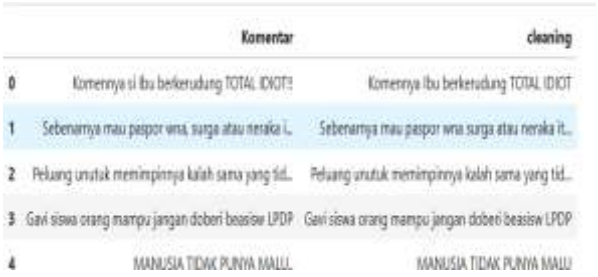
2. HASIL DAN PEMBAHASAN

Pada bagian pembahasan, peneliti tidak hanya menyajikan hasil analisis secara deskriptif, tetapi juga melakukan interpretasi terhadap pola sentimen yang muncul dalam komentar pengguna. Pembahasan ini bertujuan untuk memberikan gambaran yang lebih mendalam mengenai bagaimana masyarakat merespons isu penerima beasiswa luar negeri yang dikaitkan dengan perpindahan kewarganegaraan.

3.1. Cleaning

Cleaning merupakan tahap awal dalam proses *text preprocessing* yang bertujuan untuk membersihkan data teks dari berbagai elemen yang tidak diperlukan. Data komentar yang diambil dari media sosial biasanya masih mengandung berbagai karakter yang dapat mengganggu proses analisis, seperti tanda baca, angka, simbol khusus, URL, maupun spasi berlebih. Dengan membersihkan teks terlebih dahulu, kualitas data yang digunakan dalam proses analisis sentimen dapat menjadi lebih baik dan hasil model klasifikasi menjadi lebih akurat.

ISSN 2406-9841
e-IISN 2986-2973



	Komentar	cleaning
0	Komennya si Ibu berkerudung TOTAL IDIOT!	Komennya Ibu berkerudung TOTAL IDIOT
1	Sebenarnya mau paspor wna, surga atau neraka i...	Sebenarnya mau paspor wna surga atau neraka it...
2	Peluang untuk memimpinnya kalah sama yang tid...	Peluang untuk memimpinnya kalah sama yang tid...
3	Gavi siswa orang mampu jangan doberi beasiswa LPDP	Gavi siswa orang mampu jangan doberi beasiswa LPDP
4	MANUSIA TIDAK PUNYA MALU.	MANUSIA TIDAK PUNYA MALU

Gambar 2 : Hasil Proses *Cleaning*
Sumber: Penelitian mandiri

3.2. Case Folding

Case Folding merupakan tahap dalam *text preprocessing* yang bertujuan untuk menyeragamkan bentuk huruf dalam teks. Pada tahap ini seluruh huruf dalam teks diubah menjadi huruf kecil (*lowercase*). Proses ini dilakukan karena dalam analisis teks, perbedaan huruf besar dan huruf kecil tidak memiliki makna yang berbeda.
Sebagai contoh, kata "Pendidikan", "pendidikan", dan "PENDIDIKAN" secara semantik memiliki arti yang sama. Namun bagi sistem komputer, ketiga bentuk tersebut dianggap sebagai kata yang berbeda apabila tidak dilakukan proses case folding.
.

Gambar 3 : Hasil Proses *Case Folding*
Sumber: Penelitian mandiri

3.3. Slangwords

Slangwords merupakan kata-kata tidak baku atau bahasa gaul yang sering digunakan dalam percakapan sehari-hari, terutama pada media sosial. Dalam data komentar dari platform seperti YouTube, atau Instagram, sering ditemukan berbagai bentuk kata tidak baku seperti singkatan, penulisan yang disingkat, maupun variasi kata yang tidak sesuai dengan kaidah bahasa Indonesia yang baku.

Pada tahap ini dilakukan proses normalisasi kata slang menjadi kata yang lebih baku agar dapat dipahami dengan lebih baik oleh sistem analisis teks. Contohnya seperti kata "gk" atau "ga" yang diubah menjadi "tidak", "bgt" menjadi "banget", "yg" menjadi "yang", serta "dgn" menjadi "dengan".

Gambar 4 : Hasil Proses Slangwords
Sumber: Penelitian mandiri

3.4. Tokenize

Tokenizing merupakan tahap dalam *text preprocessing* yang bertujuan untuk memecah teks menjadi bagian-bagian yang lebih kecil yang disebut *token*. Token biasanya berupa kata-kata tunggal yang dihasilkan dari pemisahan sebuah kalimat atau dokumen teks. Pada tahap ini, setiap kalimat akan dipecah berdasarkan spasi atau tanda pemisah lainnya sehingga menghasilkan daftar kata yang terpisah. Sebagai contoh, kalimat "pendidikan di indonesia perlu diperbaiki" setelah melalui proses tokenizing akan diubah menjadi daftar kata seperti ["pendidikan", "di", "indonesia", "perlu", "diperbaiki".

Gambar 5 : Hasil Proses Tokenize
Sumber: Penelitian mandiri

3.5. Stopword Removal

Stopword Removal merupakan tahap dalam *text preprocessing* yang bertujuan untuk menghapus kata-kata umum yang sering muncul dalam teks tetapi tidak memiliki makna penting dalam proses analisis. Kata-kata tersebut biasanya hanya berfungsi sebagai penghubung dalam kalimat dan tidak memberikan kontribusi signifikan terhadap makna utama dari sebuah teks.

Pada tahap ini, sistem akan menghapus kata-kata stopword dari hasil tokenizing sehingga hanya menyisakan kata-kata yang dianggap lebih penting atau memiliki makna yang lebih kuat dalam analisis. Dengan menghilangkan kata-kata yang tidak relevan, jumlah fitur dalam data teks dapat dikurangi sehingga proses analisis menjadi lebih efisien.

ISSN 2406-9841
e-IISN 2986-2973

Gambar 6 : Hasil Proses *Stopword Removal*
Sumber: Penelitian mandiri

3.6. Steming

Stemming merupakan salah satu tahap dalam *text preprocessing* yang bertujuan untuk mengubah kata yang memiliki imbuhan menjadi bentuk kata dasar (*root word*). Dalam bahasa Indonesia, banyak kata yang mengalami penambahan awalan, akhiran, atau konfiks sehingga menghasilkan berbagai variasi kata yang sebenarnya memiliki makna dasar yang sama.
Proses stemming dilakukan dengan cara menghilangkan imbuhan seperti awalan (prefix), akhiran (suffix), maupun kombinasi imbuhan pada suatu kata. Contohnya seperti kata "memakan", "dimakan", dan "makanan" yang setelah melalui proses stemming akan menjadi kata dasar "makan".

Gambar 7 : Hasil Proses *Steming*
Sumber: Penelitian mandiri

3.7. Hasil Evaluasi Model (*Confusion Matrix*)

Dataset dibagi menjadi 80% data *training* dan 20% data testing. Hasil matriks konfusi untuk pengujian masing-masing algoritma pada data uji dipaparkan di bawah ini.

Tabel 1 : *Confusion Matrix Naive Bayes*

Aktual \ Prediksi	Negatif	Netral	Positif
Negatif	252	0	0
Netral	64	0	2
Positif	140	0	7

Sumber: Penelitian mandiri

Tabel 2 : *Confusion Matrix SVM*

Aktual \ Prediksi	Negatif	Netral	Positif
Negatif	243	0	9
Netral	55	5	6
Positif	95	0	52

Tabel 3 : *Confusion Matrix Random Forest*

Aktual \ Prediksi	Negatif	Netral	Positif
Negatif	229	5	28
Netral	47	13	6
Positif	89	7	51

Sumber: Penelitian mandiri

3.8. Analisis Perbandingan Model

Tabel 4 : Hasil Perbandingan Metrik Performa Klasifikasi

Model Algoritma	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Naive Bayes	0.556989	0.545369	0.556989	0.414155
Support Vector Machine	0.645161	0.722379	0.645161	0.581966
Random Forest	0.630108	0.628783	0.630108	**0.588081**

Sumber: Penelitian mandiri

3.9. Pembahasan Performa

Berdasarkan metrik evaluasi pada Tabel 3.5, *Support Vector Machine* (SVM) dikonfirmasi memiliki tingkat *Accuracy* murni tertinggi yaitu 0.645161 dan tingkat *Precision* yang dominan mencapai 0.722379. Hal ini membuktikan bahwa algoritma SVM sangat andal dalam memisahkan pola kalimat kontroversial isu kebangsaan dan meminimalkan kesalahan tipe *False Positive*.

Meskipun demikian, model *Random Forest* memperlihatkan kestabilan nilai *F1-Score* yang sangat kompetitif (0.588081) berkat kemampuannya menangkap sampel data pada kelas netral yang minoritas jauh lebih baik daripada model *Naive Bayes*.

3.10. Hasil pembahasan

Random Forest terpilih sebagai model terbaik dalam penelitian ini karena menunjukkan performa yang paling unggul dan stabil dalam mengklasifikasikan sentimen, terutama saat mempertimbangkan distribusi kelas yang tidak seimbang (*imbalanced data*). Sebagai algoritma berbasis *ensemble learning* yang mengombinasikan banyak pohon keputusan (*decision trees*), *Random*

Forest melakukan generalisasi data dengan cara mengambil sampel acak baik dari sisi data (*bagging*) maupun dari sisi fitur (*feature randomness*). Karakteristik ini membuat *Random Forest* lebih tangguh terhadap variasi teks komentar YouTube yang tidak merata dan mampu menekan risiko *overfitting*, sehingga menghasilkan nilai *Accuracy* serta keseimbangan *F1-Score* yang optimal di seluruh kelas sentimen (Positif, Negatif, dan Netral).

3. SIMPULAN

a. Karakter opini publik di kolom komentar *YouTube* terhadap isu sensitif ini secara umum sangat didominasi oleh Sentimen Negatif (54,7%), mengekspresikan ketidakpuasan publik atas lunturnya nasionalisme *awardee*.
b. Evaluasi kinerja menunjukkan bahwa model *Support Vector Machine* (SVM) merupakan algoritma terbaik untuk tugas klasifikasi ini dengan raihan nilai *Accuracy* sebesar 64,52% dan *Precision* sebesar 72,24%.
c. *Naive Bayes* mencatatkan nilai terendah dengan tingkat *Accuracy* 55,70% dan *F1-Score* sebesar 41,42% akibat bias ekstrim terhadap data tidak seimbang.

DAFTAR PUSTAKA

[1] Liu, B. (2020). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.
[2] Jurafsky, D., & Martin, J. H. (2024). *Speech and Language Processing (3rd ed. draft)*. Prentice Hall.
[3] Manning, C. D., Raghavan, P., & Schütze, H.(2008). *Introduction to Information Retrieval*. Cambridge University Press.
[4] Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5-32.
[5] Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. Machine Learning, 20(3), 273-297.
[6] Pedregosa, F., et al. (2011). *Scikit-learn: Machine learning in Python*. Journal of *Machine learning* Research, 12.