Peningkatan Performa Klasifikasi *Machine Learning* Melalui Perbandingan Metode *Machine Learning* dan Peningkatan *Dataset*

Fikri Baharuddin[1]*, Aris Tjahyanto[2]
Departemen Sistem Informasi[1], [2]
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
fikribaharuddin@gmail.com[1], aristj@its.ac.id [2]

***Abstract*—Classification using machine learning is an alternative that is widely used to classify data. There are various classification methods or also known as machine learning classification algorithms that can be used. However, to get the best classification results, we need a classifier that fits the dataset type to provide the best classification performance. In addition, the quality and quantity of data contained in a dataset also has an influence on the classification performance. In this study, several attempts were made to improve the classification performance of the dataset of Indonesian language exam questions at the elementary school level based on the category of difficulty level. The efforts made consist of improving the quality of the dataset and using the StringToWordVector filter algorithm to manage textual data, as well as the use of several classification algorithms such as the nave Bayes algorithm, Random Forest, and REPTree. Classification is done by using WEKA Tools. The results of the experiments carried out showed the highest performance increase of 15% after improving the quality of the dataset and using the right classification method.**

Keywords*— *Classification, StringToWordVector, Machine Learning, Exam Classification

***Abstrak*—Klasifikasi menggunakan *machine learning* menjadi alternatif yang banyak dilakukan untuk melakukan klasifikasi data. Terdapat berbagai metode klasifikasi atau juga dikenal dengan istilah algoritma pengklasifikasi *machine learning* yang dapat digunakan. Namun untuk mendapatkan hasil klasifikasi terbaik, diperlukan sebuah pengklasifikasi yang sesuai dengan tipe *dataset* agar dapat memberikan performa klasifikasi terbaik. Selain itu, kualitas dan kuantitas data yang terdapat dalam sebuah *dataset* juga memberikan pengaruh terhadap performa klasifikasi. Pada penelitian ini, dilakukan beberapa upaya untuk meningkatkan performa klasifikasi yang dilakukan terhadap *dataset* soal ujian Bahasa Indonesia tingkat sekolah dasar berdasarkan kategori tingkat kesulitannya. Upaya yang dilakukan terdiri dari peningkatan kualitas *dataset* dan penggunaan algoritma filter *StringToWordVector* untuk mengelola data tekstual, serta penggunaan beberapa algoritma pengklasifikasi seperti algoritma *naïve bayes, Random Forest,* dan *REPTree*. Klasifikasi dilakukan dengan memanfaatkan WEKA *Tools*. Hasil percobaan yang dilakukan menunjukkan adanya peningkatan performa tertinggi sebesar 15% setelah dilakukan peningkatan kualitas *dataset* dan penggunaan metode klasifikasi yang tepat.**

Kata Kunci*—*Klasifikasi, StringToWordVector, Machine Learning, Klasifikasi Ujian

I. PENDAHULUAN

Machine Learning (ML) adalah studi ilmiah tentang algoritma dan model statistik yang digunakan sistem komputer untuk melakukan tugas tertentu tanpa diprogram secara eksplisit yang bertujuan untuk mengajarkan mesin mengenai bagaimana cara mengelola data secara lebih efisien[1]. Dalam implementasinya, *machine learning* memiliki banyak metode yang dapat digunakan untuk menangani klasifikasi, *clustering*, dan pengelolaan data lainnya. Dalam penelitian [2], [3] digunakan beberapa model *machine learning* yang diantaranya terdapat algoritma *Naïve Bayes* dan *Random Forest*. Beberapa penelitian terdahulu menggunakan algoritma *Naïve Bayes* untuk mempelajari hubungan antar konsep dari keterkaitan yang diekstrak[4] dan menganalisis beberapa atribut yang menyebabkan pelajar melepaskan diri sehingga tidak mengikuti kegiatan belajar di lingkungan belajar *online*[5]. Algoritma *Random Forest* banyak memberikan kontribusi terhadap berbagai penelitian seperti yang dijabarkan pada [6]–[8].

Soal Ujian Bahasa Indonesia tingkat sekolah dasar yang digunakan dalam penelitian ini merupakan data dengan tipe data tekstual. Performa yang ditunjukkan oleh sebuah algoritma pengklasifikasi dapat berbeda-beda bergantung pada objek yang diklasifikasi. Untuk memberikan performa klasifikasi terbaik, perlu digunakan sebuah algoritma pengklasifikasi yang sesuai. Selain penggunaan metode pengklasifikasi yang tepat, faktor lain yang mempengaruhi performa klasifikasi adalah kualitas yang dimiliki oleh obyek yang akan diklasifikasi yaitu *dataset*. Secara garis besar, diperlukan sebuah upaya untuk meningkatkan performa klasifikasi yang akan dilakukan terhadap *dataset* soal ujian Bahasa Indonesia tingkat sekolah dasar.

Mengacu pada artikel[9], performa klasifikasi dapat

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i1.1337, Copyright ©2020
Submitted : December 21, 2021; Revised : February 9, 2022; Accepted : February 12, 2022; Published : March 7, 2022

ditingkatkan dengan cara menambahkan jumlah data, melengkapi data yang tidak lengkap, rekayasa fitur(*feature engineering*), penyaringan fitur(*feature selection*), penggunaan berbagai algoritma, penyesuaian algoritma, dan *ensemble method.* Beberapa peneliti terdahulu menggunakan berbagai metode untuk meningkatkan performa klasifikasi. Penelitian [10], [11] menggunakan model *hybrid parameterization* dan RHSBoost *ensemble method* untuk meningkatkan performa klasifikasi yang dilakukan terhadap data yang tidak seimbang. Pujari dan Gupta[12] menerapkan *feature selection* dan *ensemble model* untuk meningkatkan akurasi klasifikasi dalam penelitiannya. Penelitian yang dilakukan oleh [13] menggunakan *hybrid artificial intelligence* dan model *support vector mechine*(SVM) untuk meningkatkan akurasi pada penelitian terkait penyelesaian sengketa proyek. Penelitian [14] mengintegrasikan pra-proses data jaringan pengklasifikasi *Naïve Bayes* dengan algoritma pencarian *tree augmented Naïve Bayes.* Dan pada penelitian [15] menggunakan kombinasi beberapa algoritma pengklasifikasi untuk menghasilkan performa yang baik untuk diterapkan terhadap teks dokumen yang tidak terstruktur. Dalam beberapa penelitian terdahulu disarankan bahwa pada penelitian sejenis di masa mendatang dapat menerapkan penggunaan model *feature selection* untuk mendapatkan performa klasifikasi yang lebih baik.

Dalam penelitian ini, terdapat dua upaya yang dilakukan untuk meningkatkan performa klasifikasi data soal ujian Bahasa Indonesia tingkat sekolah dasar. Upaya yang pertama adalah upaya peningkatan kualitas *dataset* yang dilakukan dengan melakukan penambahan jumlah data dalam *dataset*, Upaya yang kedua adalah dengan melakukan komparasi atau perbandingan terhadap performa klasifikasi dari beberapa algoritma pengklasifikasi. Algoritma pengklasifikasi yang digunakan selama percobaan terdiri dari algoritma *Naïve Bayes*, *Random Forest*, dan *REPTree*. Beberapa peneliti menggunakan algoritma REPTree untuk mengklasifikasikan data terkait susunan kalimat dan Bahasa seperti pengelolaan *spam* pada *e-mail*[16] dan klasifikasi teks Bahasa Arab[17]. Penelitian Percobaan klasifikasi dilakukan dengan memanfaatkan WEKA(*Waikato Environment for Knowledge Analysis*) *Tools* yang dikembangkan oleh Universitas Waikato.

II. METODOLOGI

A. Pengumpulan Data

Pengumpulan data yang dilakukan dalam penelitian ini dilakukan dalam tiga tahap. Tahap pertama dilakukan pada saat membangun *dataset* awal yang terdiri dari 183 data soal. Pada pengumpulan data tahap kedua terdapat penambahan jumlah data menuju 273 soal. Dan pengumpulan data tahap ketiga terdapat penambahan jumlah data hingga menjadi 418 data soal. Dalam *dataset* terdapat dua atribut yang dapat digunakan untuk mengidentifikasi soal ujian. Atribut tersebut adalah jenis soal dan kategori soal. Jenis soal terdiri dari 11 jenis. Sedangkan untuk kategori soal terdiri dari tiga jenis yang mengidentifikasikan tingkat kesulitan soal. Jenis dan kategori soal dapat dilihat pada *TABEL 1*.

TABEL 1 JENIS DAN KATEGORI SOAL

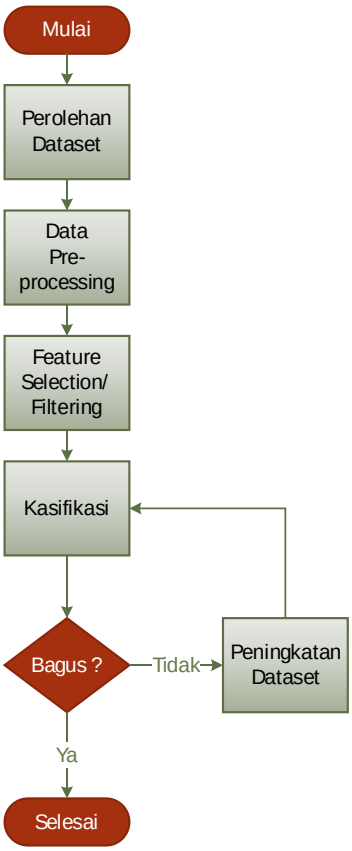
Atribut	***Value***
Jenis Soal	Memaknai
	Sinonim/Antonim
	Membaca
	Analisa
	Paragraf
	Membandingkan
	Melengkapi
	Sastra
	Mengurutkan
	Tata Penulisan
	Iklan/Laporan/Pidato
Kategori Soal	Mudah
	Sedang
	Sulit

B. Klasifikasi *Machine Learning*

Klasifikasi *machine learning* yang dilakukan dalam penelitian ini memanfaatkan WEKA *Tools*. Untuk algoritma penyaringan(*filtering*) yang digunakan adalah algoritma *StringToWordVector*. Proses klasifikasi pada penelitian ini melibatkan tiga algoritma pengklasifikasi. Algoritma pengklasifikasi yang digunakan antara lain adalah algoritma *Naïve Bayes*, *Random Forest*, dan *REPTree*. Penggunaan tiga algoritma pengklasifikasi yang berbeda tersebut bertujuan untuk membandingkan performa klasifikasi yang dihasilkan dari masing-masing algoritma.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i1.1337, Copyright ©2020
Submitted : December 21, 2021; Revised : February 9, 2022; Accepted : February 12, 2022; Published : March 7, 2022

C. Alur Penelitian



Gambar 1 Alur Penelitian

Mengacu pada *Gambar 1*, penelitian dimulai dari tahap pengumpulan data. Pengumpulan data ini bertujuan untuk membentuk *dataset* soal ujian Bahasa Indonesia. Setelah data soal ujian terkumpul, tahap selanjutnya adalah pra-proses dimana pada proses ini dilakukan pelabelan atau pengategorian soal ujian berdasarkan tingkat kesulitannya secara manual. Setelah pra-proses *dataset* selesai, *dataset* kemudian dikonversi menuju format yang didukung oleh WEKA *Tools* untuk kemudian diimpor menuju WEKA *Explorer*. Data yang telah diimpor tersebut kemudian diolah dengan menggunakan *filter* atau *tokenizer StringToWordVector*. Setelah penyaringan dilakukan, maka tahap selanjutnya adalah percobaan klasifikasi. Pada tahap percobaan klasifikasi ini, performa klasifikasi akan ditinjau. Apabila performa klasifikasi kurang bagus, maka dilakukan peningkatan terhadap *dataset* dan dilanjutkan dengan mengulang proses klasifikasi. Apabila performa mengalami peningkatan dan menunjukkan hasil yang baik, maka penelitian dinyatakan selesai.

III. HASIL DAN PEMBAHASAN

A. Perolehan *Dataset*

Dataset yang digunakan dalam penelitian ini adalah data soal Bahasa Indonesia tingkat sekolah dasar yang diperoleh melalui beberapa bank soal *online* yang dikelola oleh guru dan pengajar serta dari buku – buku pembahasan soal ujian tingkat sekolah dasar. Pada *dataset* awal, jumlah data soal yang terkumpul adalah sebanyak 183 baris data. Data soal yang dikumpulkan adalah data soal pilihan ganda. Atribut yang terdapat pada *dataset* terdiri dari atribut jenis soal, uraian soal, pertanyaan soal, opsi1 soal, opsi2 soal, opsi3 soal, opsi4 soal, kunci jawaban soal, serta kategori soal.

Atribut jenis soal sesuai dengan penjabaran yang terdapat pada Tabel 1. Atribut uraian soal digunakan untuk menyimpan uraian penjelasan dari soal yang ditanyakan. Atribut pertanyaan soal digunakan untuk menyimpan pertanyaan utama yang diberikan. Untuk atribut Opsi1 hingga Opsi4 soal digunakan untuk menyimpan pilihan jawaban yang disediakan untuk menjawab soal yang ditanyakan. Atribut kunci jawaban soal digunakan untuk menyimpan jawaban yang tepat untuk menjawab soal yang ditanyakan. Dan atribut kategori soal merupakan hasil pelabelan kategori soal yang dilakukan pada tahap selanjutnya. Data – data soal ujian yang telah terkumpul kemudian disajikan dalam format CSV. Sampel data soal dapat diamati pada Tabel 2.

TABEL 2 SAMPEL DATA SOAL

Atribut	***Value***
Jenis Soal	Membaca
Uraian Soal	Kebiasaan makan makanan bergizi dengan kadar seimbang menyebabkan tubuh anak mengalami pertumbuhan dan perkembangan yang baik. Selain itu tubuh anak yang sehat juga mempunyai daya kekebalan yang baik terhadap serangan berbagai penyakit. Orang tua harus memenuhi kebutuhan zat gizi anak-anaknya. Tubuh anak mudah terserang berbagai macam penyakit.. Pemenuhan gizi itu dimulai dengan pemberian ASI saat anak masih di usia balita. Semakin bertambah usia, maka semakin bertambah pula kebutuhan gizi seorang anak.
Pertanyaan Soal	Kalimat dalam paragraf di atas yang tidak padu adalah
Opsi1 Soal	Kebutuhan gizi anak akan semakin bertambah seiring semakin bertambah usianya.
Opsi2 Soal	Kebutuhan gizi anak baru terpenuhi saat anak beranjak dewasa.
Opsi3 Soal	Tubuh anak mudah terserang berbagai macam penyakit.
Opsi4 Soal	Pertumbuhan dan perkembangan anak akan semakin baik.
Kunci Jawaban Soal	Tubuh anak mudah terserang berbagai macam penyakit.
Kategori Soal	Sedang

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i1.1337, Copyright ©2020
Submitted : December 21, 2021; Revised : February 9, 2022; Accepted : February 12, 2022; Published : March 7, 2022

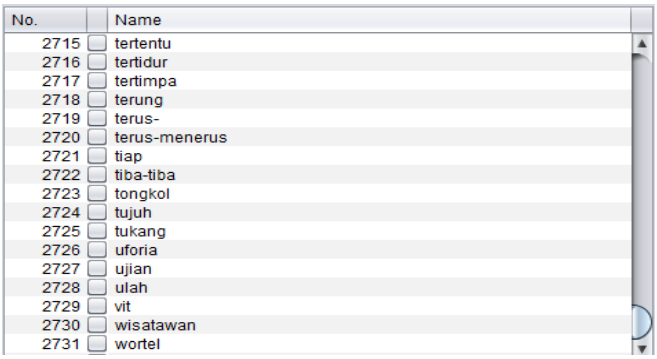
B. Pra-proses *Dataset*

Pada tahapan ini, dilakukan pelabelan terhadap data – data soal ujian yang terkumpul. Pelabelan dilakukan untuk memberikan label kategori terhadap masing–masing data soal. Kategori terbagi menjadi tiga sebagaimana dapat diamati pada Tabel 1. Pelabelan didasarkan pada hasil *interview* yang dilakukan terhadap seorang guru yang memiliki pengalaman mengajar mata pelajaran Bahasa Indonesia selama lebih dari tujuh tahun. Pada *dataset* awal, pelabelan dilakukan terhadap 183 soal.

Selanjutnya, *dataset* yang telah dilabeli sesuai dengan tingkat kesulitannya tersebut disajikan dalam bentuk CSV. *Dataset* selanjutnya dikonversi menuju format ARFF[18] agar dapat diproses oleh WEKA *Tools*. Pada WEKA *Tools*, *dataset* yang telah dikonversi tersebut diimpor untuk digunakan sebagai sumber data. Proses impor sumber data dilakukan melalui WEKA Explorer.

C. *Feature Selection* / *Filtering*

Setelah *dataset* berhasil diimpor menuju WEKA *Explorer*, Langkah selanjutnya adalah melakukan penyaringan(*filtering*) terhadap *dataset* yang akan digunakan. Pada penelitian ini, filter *tokenizer* yang akan digunakan adalah filter *unsupervised* yaitu *StringToWordVector*. Dengan menerapkan filter tersebut kepada *dataset*, maka WEKA selanjutnya akan memproses tokenisasi. Tokenisasi berfungsi untuk memecah setiap kata yang terkandung dalam data – data soal dan menjadikannya sebagai atribut. Sampel dari hasil tokenisasi dapat diamati pada Gambar 2. Setelah proses penyaringan selesai, Langkah selanjutnya adalah menentukan atribut yang akan diklasifikasikan. Pada penelitian ini, atribut yang akan diklasifikasikan adalah atribut kategori soal.



Gambar 2 Sampel Hasil Filtering Menggunakan Filter StringToWordVector

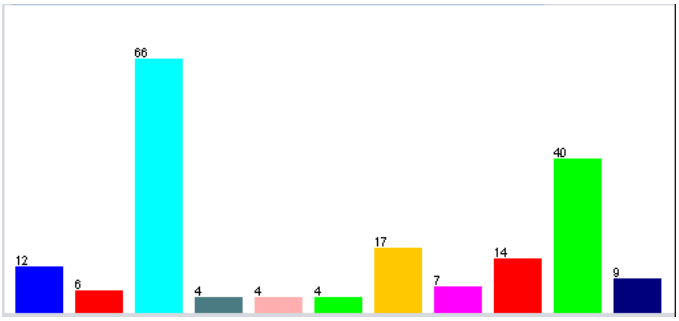
D. Klasifikasi dan Peningkatan *Dataset*

Dataset awal yang digunakan untuk memproses klasifikasi terdiri dari 183 data. Seperti yang dapat diamati pada Tabel 3, *dataset* awal terbagi atas 68 soal dengan kategori mudah, 59 soal dengan kategori sedang, dan 56 soal yang termasuk dalam kategori sulit.

TABEL 3 JUMLAH SOAL BERDASARKAN KATEGORI PADA DATASET AWAL

Kategori	Jumlah Soal
Mudah	68
Sedang	59
Sulit	56

Visualisasi persebaran data soal berdasarkan jenisnya dapat diamati pada Gambar 3. Pada hasil visualisasi tersebut, *chart* yang paling kiri mewakili jenis soal "memaknai" dan *chart* paling kanan mewakili jenis soal "iklan/laporan/pidato". Urutan jenis dari paling kiri hingga kanan diurutkan sesuai dengan urutan jenis yang terdapat pada Tabel 1.



Gambar 3 Jumlah Soal berdasarkan Jenis pada Dataset Awal

Setelah melakukan penyaringan(*filtering*) dengan memanfaatkan algoritma *StringToWordVector*, maka selanjutnya dilakukan proses klasifikasi yang terdapat pada menu "*Classifiy*" WEKA Explorer. Terdapat tiga algoritma yang digunakan pada proses klasifikasi yang terdiri dari *Naïve Bayes*, *Random Forest*, dan *REPTree*. Penggunaan tiga algoritma pengklasifikasi ini bertujuan untuk mendapatkan performa klasifikasi yang paling tinggi dalam mengklasifikasikan soal ujian Bahasa Indonesia berdasarkan tingkat kesulitannya.

Pada *dataset* awal yang digunakan, proses klasifikasi yang dilakukan menggunakan algoritma *Naïve Bayes* menunjukkan bahwa dari total soal yang berjumlah 183, terdapat 111 soal yang terklasifikasi dengan benar serta terdapat 72 soal yang tidak terklasifikasi dengan benar. Dengan hasil tersebut, maka diperoleh performa klasifikasi sebesar 60.66 %. Sedangkan klasifikasi yang dilakukan dengan memanfaatkan algoritma *Random Forest* menghasilkan 115 soal terklasifikasi dengan benar dan terdapat 68 soal yang tidak terklasifikasi dengan baik. Performa klasifikasi dengan menggunakan *Random Forest* adalah sebesar 62.84 %. Dan klasifikasi *dataset* awal yang memanfaatkan algoritma *REPTree* menghasilkan 139 soal terklasifikasi dengan benar dan 44 soal yang tidak terklasifikasi dengan benar. Performa klasifikasi yang diperoleh dari penggunaan algoritma *REPTree* ini adalah sebesar 75.96 %. Lebih detail mengenai hasil percobaan klasifikasi pada *dataset* awal ini dapat dilihat pada Tabel 4.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i1.1337, Copyright ©2020
Submitted : December 21, 2021; Revised : February 9, 2022; Accepted : February 12, 2022; Published : March 7, 2022

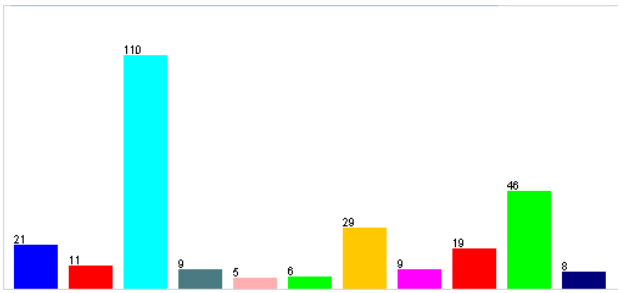
TABEL 4 PERBANDINGAN HASIL KLASIFIKASI PADA DATASET AWAL

Components	Classification Algorithms		
	Naïve Bayes	Random Forest	REPTree
Number of Data	183	183	183
Filter	StringToWord Vector	StringToWord Vector	StringToWord Vector
Correct Classification	111	115	139
Incorrect Classification	72	68	44
Correct Classification Rate	60.66%	62.84%	75.96%

Selain menggunakan algoritma pengklasifikasi yang berbeda, upaya lain yang dilakukan untuk meningkatkan performa klasifikasi adalah dengan meningkatkan jumlah data dalam *dataset*. Peningkatan *dataset* yang pertama dilakukan dengan menambahkan data yang semula hanya terdiri dari 183 data soal hingga menjadi 273 data soal. Setelah dilakukan penambahan data soal, maka perubahan jumlah soal berdasarkan kategori dan jenisnya juga mengalami perubahan. Jumlah soal berdasarkan kategori setelah peningkatan *dataset* yang pertama dapat dilihat pada Tabel 5. Dan visualisasi jumlah soal berdasarkan jenis soal setelah peningkatan *dataset* pertama dapat dilihat pada Gambar 4.

TABEL 5 JUMLAH SOAL BERDASARKAN KATEGORI PADA PENINGKATAN DATASET PERTAMA

Kategori	Jumlah Soal
Mudah	106
Sedang	74
Sulit	93



Gambar 4 Jumlah Soal berdasarkan Jenis pada Peningkatan Dataset Pertama

Proses klasifikasi dilakukan kembali dengan menggunakan *dataset* yang telah ditingkatkan. Sebagai perbandingan, klasifikasi kembali dilakukan dengan menggunakan tiga algoritma yang sama dengan percobaan klasifikasi sebelumnya. Dengan menggunakan *dataset* yang telah diberikan peningkatan pertama, performa klasifikasi yang menggunakan algoritma *Naïve Bayes* mengalami penurunan menjadi 58.97 %. Sedangkan performa klasifikasi dengan algoritma *Random Forest* mengalami peningkatan menjadi 63.004 %. Dan performa klasifikasi dengan algoritma *REPTree* mengalami peningkatan menjadi 82.78%. Selengkapnya mengenai hasil percobaan klasifikasi setelah peningkatan *dataset* pertama dapat dilihat pada Tabel 6.

TABEL 6 PERBANDINGAN HASIL KLASIFIKASI PADA PENINGKATAN DATASET PERTAMA

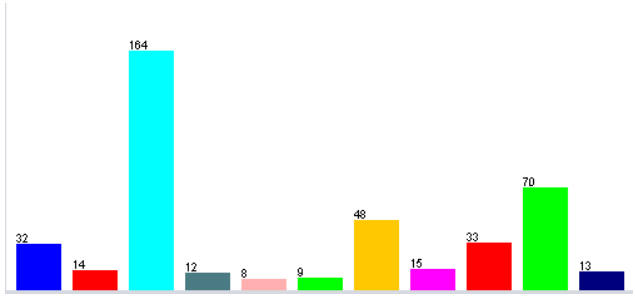
Components	Classification Algorithms		
	Naïve Bayes	Random Forest	REPTree
Number of Data	273	273	273
Filter	StringToWord Vector	StringToWord Vector	StringToWord Vector
Correct Classification	161	172	226
Incorrect Classification	112	101	47
Correct Classification Rate	58.97%	63.004%	82.78%

Hasil klasifikasi yang dilakukan setelah peningkatan *dataset* pertama memberikan peningkatan cukup signifikan terutama pada algoritma pengklasifikasi *REPTree*. Untuk mendapatkan kesimpulan percobaan, maka dilakukan peningkatan kedua terhadap *dataset* soal ujian. *Dataset* yang sebelumnya ditambah menjadi 273 data, ditingkatkan kembali jumlahnya menjadi 418 data. Detail jumlah soal berdasarkan kategori dan visualisasi persebaran soal berdasarkan jenisnya dapat diamati pada Tabel 7 dan Gambar 5.

TABEL 7 JUMLAH SOAL BERDASARKAN KATEGORI SETELAH PENINGKATAN DATASET KEDUA

Kategori	Jumlah Soal
Mudah	168
Sedang	104
Sulit	146

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i1.1337, Copyright ©2020
Submitted : December 21, 2021; Revised : February 9, 2022; Accepted : February 12, 2022; Published : March 7, 2022



Gambar 5 Jumlah Soal berdasarkan Jenis pada Peningkatan Dataset Kedua

Dengan menggunakan *dataset* yang telah ditingkatkan untuk kedua kalinya, diperoleh performa klasifikasi sebagai berikut. Klasifikasi dengan menggunakan algoritma *Naïve Bayes* meningkat menjadi 64.83 %. Untuk klasifikasi yang dilakukan dengan menggunakan algoritma *Random Forest*, terdapat peningkatan performa dibanding percobaan sebelumnya menjadi 70.57 %. Dan untuk klasifikasi dengan algoritma *REPTree* menunjukkan peningkatan performa klasifikasi menjadi 91.15 %. Hasil klasifikasi dengan menggunakan *dataset* yang telah diberikan peningkatan kedua dapat diamati pada *TABEL 8*.

TABEL 8 PERBANDINGAN HASIL KLASIFIKASI PADA PENINGKATAN DATASET KEDUA

Components	**Classification Algorithms**		
	Naïve Bayes	**Random Forest**	**REPTree**
Number of Data	418	418	418
Filter	StringToWord Vector	StringToWord Vector	StringToWord Vector
Correct Classification	271	295	361
Incorrect Classification	147	123	37
Correct Classification Rate	64.83%	70.57%	91.15%

Dari tiga percobaan klasifikasi yang dilakukan dengan menggunakan tiga algoritma pengklasifikasi yang terdiri dari algoritma *Naïve Bayes*, *Random Forest*, dan *REPTree*, dapat diketahui bahwa algoritma *REPTree* memiliki performa tertinggi yang menandakan bahwa algoritma tersebut dapat digunakan untuk mengklasifikasikan tingkat kesulitan lebih baik dari dua algoritma lain yang digunakan.

IV. KESIMPULAN

Penelitian ini membahas tentang beberapa upaya yang dilakukan untuk meningkatkan hasil klasifikasi pada *dataset* soal ujian Bahasa Indonesia tingkat sekolah dasar berdasarkan tingkat kesulitannya. Upaya yang dilakukan antara lain adalah dengan menambah jumlah data dan menggunakan algoritma klasifikasi yang berbeda. Dalam upaya yang dilakukan, terdapat penambahan jumlah data yang semula hanya berjumlah 183 data menjadi 418 data. Untuk algoritma klasifikasi yang dibandingkan adalah *Naïve Bayes*(NB), *Random Forest*(RF), dan *REPTree*. Klasifikasi dilakukan dengan memanfaatkan WEKA *Tools*. Berdasarkan hasil percobaan yang dilakukan, terdapat adanya peningkatan performa tertinggi sebesar 15% setelah dilakukan peningkatan kualitas *dataset* dan penggunaan metode klasifikasi yang tepat. Dengan demikian dapat disimpulkan bahwa algoritma klasifikasi *REPTree* yang dikombinasikan dengan penggunaan StringToWordVector menunjukkan tingkat klasifikasi tertinggi untuk mengategorikan soal ujian Bahasa Indonesia berdasarkan tingkat kesulitannya. Selain itu, jumlah data dalam *dataset* serta algoritma klasifikasi yang digunakan dapat mempengaruhi hasil klasifikasi secara signifikan. Penelitian sejenis di masa mendatang dapat dilakukan terhadap *dataset* yang jumlahnya lebih banyak serta menggunakan algoritma *tokenizer* yang lain agar mendapatkan hasil klasifikasi yang lebih baik untuk diterapkan pada tipe *dataset* yang serupa.

DAFTAR PUSTAKA

[1] B. Mahesh, "Machine Learning Algorithms-A Review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, pp. 381–386, 2020.

[2] T. P. Carvalho, F. A. Soares, R. Vita, R. da P. Francisco, J. P. Basto, and S. G. S. Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance," *Computers & Industrial Engineering*, vol. 137, p. 106024, 2019.

[3] A. Althnian *et al.*, "Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain," *Applied Sciences*, vol. 11, no. 2, p. 796, 2021.

[4] G. Zayaraz, "Concept relation extraction using Naïve Bayes classifier for ontology-based question answering systems," *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 1, pp. 13–24, 2015.

[5] T. GopalaKrishnan and P. Sengottuvelan, "A hybrid PSO with Naïve Bayes classifier for disengagement detection in online learning," *Program*, 2016.

[6] W. G. Touw *et al.*, "Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?," *Briefings in bioinformatics*, vol. 14, no. 3, pp. 315–326, 2013.

[7] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," *Pattern recognition*, vol. 44, no. 2, pp. 330–349, 2011.

[8] D. Denisko and M. M. Hoffman, "Classification and interaction in random forests," *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, pp. 1690–1692, 2018.

[9] "How To Increase Accuracy Of Machine Learning Model." https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/ (accessed Dec. 19, 2021).

[10] M. Mohamad, A. Selamat, I. M. Subroto, and O. Krejcar, "Improving the classification performance on imbalanced data sets via new hybrid parameterisation model," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 7, pp. 787–797, 2021.

[11] J. Gong and H. Kim, "RHSBoost: Improving classification performance in imbalance data," *Computational Statistics & Data*

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i1.1337, Copyright ©2020
Submitted : December 21, 2021; Revised : February 9, 2022; Accepted : February 12, 2022; Published : March 7, 2022

Analysis, vol. 111, pp. 1–13, 2017.

[12] P. Pujari and J. B. Gupta, "Improving classification accuracy by using feature selection and ensemble model," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, no. 2, pp. 380–386, 2012.

[13] J.-S. Chou, M.-Y. Cheng, and Y.-W. Wu, "Improving classification accuracy of project dispute resolution using hybrid artificial intelligence and support vector machine models," *Expert Systems with Applications*, vol. 40, no. 6, pp. 2263–2274, 2013.

[14] W.-W. Wu, "Improving classification accuracy and causal knowledge for better credit decisions," *International Journal of Neural Systems*, vol. 21, no. 04, pp. 297–309, 2011.

[15] M. Mowafy, A. Rezk, and H. El-Bakry, "An efficient classification model for unstructured text document," *American Journal of Computer Science and Information Technology*, vol. 6, no. 1, p. 16, 2018.

[16] S. K. Trivedi and P. K. Panigrahi, "Spam classification: a comparative analysis of different boosted decision tree approaches," *Journal of Systems and Information Technology*, 2018.

[17] H. Naji and W. Ashour, "Text Classification for Arabic Words Using Rep-Tree," *International Journal of Computer Science & Information Technology (IJCSIT) Vol*, vol. 8, 2016.

[18] F. Baharuddin and A. Tjahyanto, "Dataset Soal Ujian Bahasa Indonesia Tingkat Sekolah Dasar," Dec. 2021, doi: 10.5281/ZENODO.5793377.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i1.1337, Copyright ©2020
Submitted : December 21, 2021; Revised : February 9, 2022; Accepted : February 12, 2022; Published : March 7, 2022