



Penerapan Teknik SMOTE untuk Mengatasi *Imbalance Class* dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN

Anis Nikmatul Kasanah¹, Muladi², Utomo Pujianto³

^{1,2,3}Teknik Elektro, Fakultas Teknik, Universitas Negeri Malang

¹kurniaanis73@gmail.com, ²muladi@um.ac.id, ³utomo.pujianto.ft@um.ac.id

Abstract

Amount of information in the form of online news needs to be balanced with the ability of readers to sort or classify subjective or objective news. So that a special system is needed that can be used for online news objectivity classification so that it can help readers to pick up subjective or objective news. This research proposes the development of techniques in machine learning to help sort out news objectivity automatically based on the content of the news. The algorithm proposed is *K-Nearest Neighbor (KNN)* algorithm. News samples obtained from *kompas.com* by *scrapping* occur imbalance classes where the number of objective news and subjective news are not balanced. So that it can affect the performance of the classification algorithm. One technique to overcome the imbalance class is to apply the *Synthetic Minority Over-sampling Technique (SMOTE)* technique.. *SMOTE* is the generation of minority data as much as the majority data. This study compares the performance of *KNN* algorithm without *SMOTE* and the performance of *KNN* algorithm with *SMOTE*. Based on the results of the study by applying a variety of neighboring *k* values, namely 1, 3, 5, 7 and 9, it was found that the application of *SMOTE* could improve the accuracy of the *KNN* algorithm at values $k = 1$ and $k = 3$ with an average increase of 3.36. At values $k = 5, 7$ and 9 the algorithm experiences an average decrease in accuracy of 6.67.

Keywords: classification, KNN, SMOTE, subjective, objective

Abstrak

Berlimpahnya informasi dalam bentuk berita online perlu diimbangi dengan kemampuan pembaca untuk memilah atau mengklasifikasikan berita yang subjektif ataupun objektif. Sehingga diperlukan sistem khusus yang dapat digunakan untuk klasifikasi objektivitas berita *online* sehingga dapat membantu pembaca dalam memilah berita yang subjektif ataupun berita objektif. Penelitian ini mengusulkan pengembangan teknik dalam *machine learning* untuk membantu pemilahan objektivitas berita secara otomatis berdasarkan isi dari berita tersebut. Algoritma yang diusulkan adalah algoritma *K-Nearest Neighbor (KNN)*. Sampel berita yang didapatkan dari *kompas.com* dengan cara *scrapping* terjadi *imbalance class* dimana jumlah berita objektif dan berita subjektif tidak seimbang yang. Sehingga dapat mempengaruhi performa algoritma klasifikasi. Salah satu teknik untuk mengatasinya adalah menerapkan teknik *Synthetic Minority Over-sampling Technique (SMOTE)*. *SMOTE* adalah pembangkitan data minoritas sebanyak data mayoritas. Penelitian ini membandingkan performa algoritma *KNN* tanpa *SMOTE* dan performa algoritma *KNN* dengan *SMOTE*. Berdasarkan hasil penelitian dengan menerapkan nilai *k* tetangga yang bervariasi yaitu 1, 3, 5, 7 dan 9 diperoleh bahwa penerapan *SMOTE* dapat meningkatkan performa akurasi algoritma *KNN* pada nilai $k=1$ dan $k=3$ dengan rata-rata peningkatan akurasi sebesar 3,36. Pada nilai $k = 5, 7$ dan 9 algoritma mengalami rata-rata penurunan akurasi sebesar 6,67.

Kata Kunci : klasifikasi, KNN, SMOTE, subjektif, objektif

1. Pendahuluan

Berita adalah salah satu sumber informasi mengenai kejadian terkini yang dapat ditemukan pada media massa seperti surat kabar, televisi dan media lainnya [14]. Perkembangan teknologi informasi turut membantu proses penyebaran berita yang disajikan melalui media online. Media *online* adalah salah satu jenis media massa yang memanfaatkan jaringan internet yang bersifat *realtime* dan praktis. Pembaca tidak perlu menunggu besok pagi untuk mendapatkan berita terbaru, hanya dengan menggunakan internet berita sudah dapat dibaca dengan cepat. Penyebaran berita *online* yang cepat didukung oleh banyaknya portal berita *online* seperti *kompas.com*, *detik.com*, *viva news* dan masih banyak lagi. Banyaknya berita *online* belum diimbangi dengan kualitas objektivitas berita dengan baik. Situasi ini membuat pembaca harus memilah sendiri berita *online* yang dibaca sehingga membutuhkan waktu yang lama. Oleh karena itu diperlukan sebuah sistem yang dapat memilah atau mengklasifikasikan berita tersebut berdasarkan sifat objektivitas berita.

Penelitian tentang klasifikasi subjektivitas berita pernah dilakukan oleh Yara dari American University of Beirut [1], dengan penelitian yang berjudul *Syntactic Generative Algorithm for a Subjectivity Analysis of Sport Articles*. Penelitian ini membahas tentang subjektivitas berita olahraga online yang berbahasa Inggris menggunakan algoritma *Syntactic Genetic*. Berdasarkan fungsi evaluasi yang diusulkan dengan validasi sebanyak tiga kali, *Syntactic Genetic* menghasilkan rata-rata akurasi 96,2% pada data latihan dan 94,5% pada data uji.

Berdasarkan penjelasan di atas, penelitian ini akan membahas tentang pengklasifikasian berita *online* berdasarkan objektivitasnya. Berita online yang digunakan untuk klasifikasi diperoleh dari salah satu portal berita yang populer di Indonesia yaitu *kompas.com*. Terdapat beberapa algoritma untuk mengklasifikasikan teks diantaranya algoritma *Support Vector Machine (SVM)*, *Naive Bayes* dan *K-Nearest Neighbour (KNN)* [2]. Algoritma klasifikasi yang diusulkan dalam penelitian ini adalah algoritma KNN. Algoritma KNN adalah sebuah metode yang digunakan untuk melakukan klasifikasi terhadap objek berdasarkan data latihan yang memiliki jarak paling dekat dengan objek tersebut [3].

Jumlah sampel berita yang didapatkan dari *kompas.com* mengalami *imbalance class* dimana jumlah berita objektif lebih banyak dari berita subjektif. Klasifikasi pada *imbalance class* akan cenderung mengabaikan kelas yang memiliki jumlah sampel yang sedikit sehingga dapat berpengaruh buruk terhadap performa dari algoritma klasifikasi [4]. Dikutip dari penelitian [5]. Salah satu cara untuk menangani *imbalance class* yaitu menggunakan teknik *Synthetic*

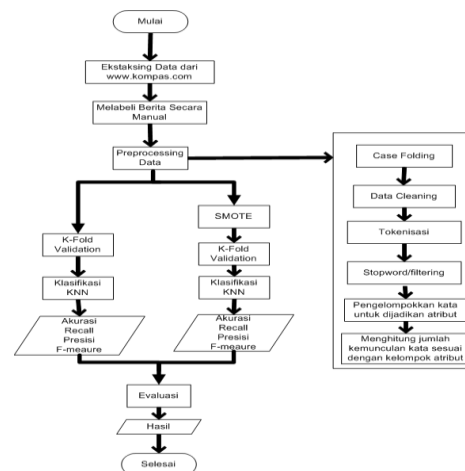
Minority Over-sampling Technique (SMOTE). SMOTE merupakan metode pembangkitan data minoritas sebanyak data mayoritas. Penelitian mengenai algoritma KNN dan SMOTE pernah diterapkan oleh Rimbun Siringoringo. Penelitian ini berjudul *Klasifikasi data tidak seimbang menggunakan algoritma SMOTE dan KNN pada tahun 2018*. Penelitian ini menghasilkan bahwa teknik SMOTE dapat menyelesaikan masalah ketidakseimbangan kelas pada dataset *Credit Card Fraud*. Dengan menerapkan skema evaluasi *10-cross validation* diperoleh kesimpulan bahwa SMOTE dapat meningkatkan rata-rata G-Mean dari 53,4% ke 81,0% dan rata-rata *F-measure* dari 38,7% ke 81,8% [6].

Berdasarkan penjelasan di atas, permasalahan yang dijumpai adalah penerapan algoritma klasifikasi untuk menangani *imbalance class* dalam klasifikasi objektivitas berita online. Sehingga dalam penelitian ini diusulkan proses SMOTE untuk menangani *imbalance class* pada dataset objektivitas berita *online*. Dengan demikian penelitian ini akan memaparkan bagaimana perbandingan performa klasifikasi algoritma KNN tanpa SMOTE dan klasifikasi algoritma KNN dengan SMOTE dalam mengklasifikasikan objektivitas berita *online*. Output dari penelitian ini diharapkan klasifikasi algoritma KNN dengan SMOTE mampu meningkatkan performa dari algoritma KNN dalam mengklasifikasikan objektivitas berita.

2. Metode Penelitian

2.1. Rancangan Penelitian

Penelitian ini dibantu oleh software data mining untuk mengetahui hasil dari dua teknik klasifikasi yang telah diusulkan. *Software* data mining yang digunakan adalah *weka*, spesifikasi perangkat yang digunakan adalah RAM dengan kapasitas 4GB, prosesor *intell inside* dengan menggunakan sistem operasi *windows 8.1 64 bit*. Langkah-langkah untuk melakukan penelitian ini ditunjukkan dalam bentuk gambar 1.



Gambar 1 Rancangan Penelitian

Gambar 1 menjelaskan alur dari penelitian klasifikasi objektivitas berita online. Pada penelitian ini menggunakan metode penelitian eksperimen. Penelitian ini dimulai dari mengumpulkan data berupa berita online yang didapatkan dari portal berita kompas.com. Setelah memperoleh data langkah selanjutnya adalah melakukan pelabelan data secara manual. Hasil pelabelan manual didapatkan jumlah berita objektif dan jumlah berita subjektif mengalami imbalance class, sehingga diperlukan penambahan teknik SMOTE untuk mengatasinya. Proses selanjutnya adalah *preprocessing* data. Teknik *preprocessing* dalam penelitian ini terdiri dari 6 enam tahap. Tahap pertama adalah *case folding*. *Case folding* adalah merubah semua huruf dalam dokumen menjadi huruf kecil (*lower case*). Tahap yang kedua adalah *data cleaning*. *Data cleaning* digunakan untuk menghilangkan semua simbol dan tanda baca yang ada dalam dokumen berita. Tahap yang ke tiga adalah tokenisasi. Tokenisasi adalah proses pemotongan dokumen berita online menjadi kata yang berdiri sendiri. Tahap yang keempat adalah *stopwords* atau *filtering* yang digunakan untuk mengambil kata-kata penting dari hasil token.

Pada tahap ini kata yang tidak memiliki arti akan dihapus. Tahap kelima adalah pengelompokan setiap dokumen menjadi beberapa kelompok kata. Pengelompokan kata ini merujuk pada penelitian yang dilakukan oleh Yara Rizk dalam penelitiannya yang berjudul *Syntactic Generative Algorithm for a Subjectivity Analysis of Sport Articles*[1]. Selain itu juga merujuk pada penelitian tentang kredibilitas berita yang dilakukan oleh Vience Mutiara Rumata pada tahun 2017 dengan judul [7]. Pada penelitian ini variabel kebenaran diukur melalui tiga indikator yaitu faktualitas, akurasi dan kelengkapan berita. Berdasarkan hal tersebut maka dipilihlah 13 kelompok kata dan 1 atribut yang dijadikan kelas. Daftar atribut ditunjukkan pada Tabel 1. Kemudian tahap terakhir adalah menghitung kemunculan jumlah kata sesuai dengan atribut yang telah ditentukan.

Tabel 1 Atribut dalam Dataset

Atribut	Tipe Data
Kata Hubung	Numerik
Kata Kerja Aktif	Numerik
Kata Kerja Pasif	Numerik
Kata Kerja yang dibendakan	Numerik
Kata Asing	Numerik
Kata Ganti Orang	Numerik
Kata Jumlah	Numerik
Kata Perbandingan	Numerik
Kata tidak Langsung	Numerik
Kata Langsung	Numerik
Kata Opini	Numerik
Kata Keterangan Waktu	Numerik
Jumlah Kata dalam Berita	Numerik
Label	Nominal

Setelah *preprocessing* selesai tahap selanjutnya adalah proses klasifikasi yang terdiri dari dua model klasifikasi yaitu klasifikasi algoritma KNN tanpa SMOTE dan klasifikasi algoritma KNN dengan SMOTE. Untuk evaluasi dari model klasifikasi menggunakan *K-fold Cross Validation* dengan jumlah k default yaitu 10. Setelah proses klasifikasi akan dilakukan evaluasi terhadap hasil klasifikasi dari kedua model klasifikasi tersebut dengan membandingkan nilai akurasi, presisi, *recall* dan *F-measure* dari kedua model tersebut. Dari hasil perbandingan kinerja ke duanya, diharapkan klasifikasi algoritma KNN dengan SMOTE mempunyai hasil kinerja lebih tinggi, sehingga klasifikasi algoritma KNN dengan SMOTE dapat meningkatkan kinerja dari klasifikasi berita *online*.

2.2. Sumber Data

Data adalah sesuatu yang belum mempunyai arti bagi penerimanya sehingga diperlukan pengolahan khusus. Data dapat berwujud angka, gambar, teks ataupun suara. Data yang digunakan dalam penelitian ini adalah data teks yang berisi teks berita Bahasa Indonesia yang didapatkan dari salah satu portal berita yang populer di Indonesia yaitu kompas.com. Data yang digunakan dalam penelitian ini berjumlah 200 berita. Perolehan berita menggunakan teknik *scraping*. Teknik *scraping* adalah sebuah teknik yang digunakan untuk mengambil atau mengekstrak sebuah data atau informasi tertentu dari suatu website secara spesifik [8]. Tools yang digunakan untuk teknik *scraping* pada penelitian ini adalah *web scraper* yang dapat diunduh melalui *extension* dari *Google Chrome*. Data yang telah diperoleh dari proses *scraping* merupakan kumpulan berita dari beberapa kategori berita, diantaranya berita ekonomi, hiburan, bisnis, sains olahraga, otomotif dan teknologi. Rentang waktu berita yang berhasil di-*scraping* dimulai dari berita tanggal 6 Oktober 2018 hingga berita pada tanggal 26 oktober 2018.

Dokumen berita yang digunakan dari hasil *scraping* adalah dokumen berita lengkap yang terdiri dari judul dan isi berita, sedangkan untuk dokumen berita yang tidak lengkap akan dihapus. Selain itu jumlah kata dalam berita berkisar antara 100 kata hingga 1000 kata. Dokumen berita selanjutnya dilakukan pelabelan objektif dan subjektif yang dilakukan secara manual oleh ahli bahasa. Setelah proses pelabelan selesai didapatkan hasil bahwa berita objektif dan berita subjektif memiliki jumlah yang berbeda dengan rincian dari jumlah berita objektif dan subjektif dapat dilihat pada Tabel 2.

Tabel 2. Rincian Jumlah Berita Subjektif dan Objektif

Kategori	Jumlah
Objektif	176
Subjektif	24

2.3. Data Preprocessing

Data Preprocessing adalah sebuah teknik yang digunakan untuk mengambil informasi dari suatu data yang masih mentah. Teknik ini digunakan untuk menghilangkan noise yang terdapat dalam data mentah, sehingga dapat diproses selanjutnya. Dalam proses ini, data mentah akan diubah menjadi kumpulan data dengan struktur, gaya penulisan dan kebenaran penulisan yang diabaikan. Penelitian yang dilakukan oleh [9] tentang tahap preprocessing teks yang terdiri dari case folding, data cleaning, tokenisasi, stopword removal.

Tahap pertama adalah Case folding adalah teknik untuk mengubah semua huruf yang ada menjadi huruf kecil (lower case). Pada tahap ini, hanya huruf “a” sampai “z” yang akan diterima. Tahap selanjutnya adalah data cleaning. Data cleaning adalah tahap pembersihan kata dengan menghilangkan delimiter koma(,), titik(.) dan delimiter lainnya. Tahap berikutnya adalah tokenisasi yaitu memotong dokumen menjadi kata-kata yang berdiri sendiri. Tahap terakhir adalah stopword removal / filtering yaitu tahap untuk mengambil kata-kata penting dari hasil token. Pada tahap ini kata yang tidak digunakan akan dihapus.

2.4 KNN

Algoritma KNN adalah sebuah algoritma untuk mengklasifikasikan objek berdasarkan data latih yang mempunyai jarak paling dekat dengan objek tersebut [10] KNN termasuk algoritma supervised learning. Menurut Widua Gata pada penelitiannya [11] Algoritma KNN bersifat sederhana, bekerja berdasarkan jarak terpendek dari data uji ke data latih untuk menentukan kelas dari data tersebut Setelah mengumpulkan data-data pada kelompok k tertentu, kemudian diambil kelas data mayoritas untuk dijadikan sebagai kelas prediksi dari data uji. KNN memiliki beberapa kelebihan yaitu tahan terhadap data yang memiliki noise dan efektif terhadap data latih yang berjumlah besar dan memiliki performa cukup baik. Namun waktu yang digunakan untuk komputasi sangatlah lama jika data latihnya besar dan sangat sensitive dengan ciri yang redundan atau relevan.

Ada banyak cara untuk mengukur jarak kedekatan antar data pada algoritma KNN diantaranya menggunakan Euclidean distance. Euclidean distance merupakan cara yang sering digunakan untuk menghitung jarak antar data. Jarak ini digunakan untuk menguji interpretasi kedekatan jarak antara dua objek. Berikut adalah persamaan dari Euclidean distance [10]:

$$dist = \sqrt{\sum_{i=1}^n (X_{i2} - X_{i1})^2} \quad (1)$$

dengan dist adalah Jarak, X_{i2} adalah Data uji, X_{i1} adalah Data Sampel, i adalah Atribut dan n adalah jumlah atribut.

2.5 SMOTE

SMOTE merupakan teknik menyeimbangkan jumlah distribusi data sampel pada kelas minoritas dengan cara menyeleksi data sampel tersebut hingga jumlah data sampel menjadi seimbang dengan jumlah sampel pada kelas mayoritas [6]. Penggunaan metode SMOTE memungkinkan akan adanya overfitting. Overfitting dapat terjadi karena data pada kelas minoritas diduplikasi sehingga memungkinkan adanya data latih yang sama. Tahapan dalam melakukan SMOTE dimulai dari menghitung jarak antar data pada data minoritas, selanjutnya menentukan nilai presentase SMOTE kemudian menentukan jumlah k terdekat dan yang terakhir adalah menciptakan data sintesis dengan persamaan berikut [5] :

$$x_{syn} = x_{i+} + (x_{knn} - x_i) \times \delta \quad (2)$$

Dengan x_{syn} adalah data sintesis yang akan diciptakan x_i data yang akan direplikasi, x_{knn} data yang memiliki jarak terdekat dari data yang akan direplikasi dan δ nilai random antara 0 dan 1.

2.6. Evaluasi Klasifikasi

Evaluasi klasifikasi didasarkan pengujian pada objek yang benar dan objek yang salah. Validasi ini digunakan untuk menentukan jenis model yang terbaik dari hasil klasifikasi [6]. Evaluasi pada penelitian ini menggunakan confusion matrix. Confusion matrix adalah informasi mengenai hasil klasifikasi aktual yang dapat diprediksi oleh sebuah sistem klasifikasi. Pada Tabel 3 akan ditampilkan tabel confusion matrix dua kelas.

Tabel 3 Confusion Matrix dua kelas

	Prediksi Positif	Prediksi Negatif
Aktual Positif	TP	FN
Aktual Negatif	FP	TN

Dari hasil confusion matrix akan ditentukan nilai Akurasi, presisi, recall dan F-measure [12]. Akurasi adalah ketetapan sistem dalam melakukan proses klasifikasi dengan benar.

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

Presisi adalah rasio jumlah dokumen yang relevan dengan total jumlah dokumen yang ditemukan pada sistem klasifikasi.

$$Presisi = \frac{TP}{TP+FP} \quad (3)$$

Recall adalah rasio jumlah dokumen yang ditemukan kembali oleh sistem klasifikasi dengan total jumlah dokumen yang relevan.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

F-measure adalah metrik evaluasi yang populer untuk menangani masalah *imbalance class*. *F-measure* mengkombinasi *recall*/sensitivitas dan presisi sehingga menghasilkan metrik yang efektif untuk mencari kembali informasi dalam himpunan yang mengandung ketidakseimbangan [13].

$$f - measure = \frac{2 \times (presisi \times recall)}{presisi + recall} \quad (5)$$

3. Hasil dan Pembahasan

Berdasarkan percobaan penelitian menggunakan aplikasi weka untuk klasifikasi dengan algoritma KNN menggunakan jumlah data sebanyak 200 data, dengan jumlah berita objektif sebanyak 176 dan berita subjektif sebanyak 24. Jumlah data yang digunakan dalam klasifikasi algoritma KNN tidak seimbang. Sehingga diperlukan tambahan teknik khusus untuk menyeimbangkan kelas data yaitu teknik SMOTE. Setelah menerapkan teknik SMOTE pada dataset objektivitas berita sehingga keseluruhan data untuk klasifikasi menjadi 352 data dengan jumlah berita objektif sebanyak 176 data dan jumlah berita subjektif sebanyak 176 data. Proporsi jumlah data ditunjukkan pada Tabel 4.

Tabel 4 Proporsi jumlah data

Label	KNN	SMOTE+KNN
Objektif	176	176
Subjektif	24	176

Pada Tabel 5 dan Tabel 6 akan ditampilkan data hasil dari perbandingan hasil *confusion matrix* dan hasil performa klasifikasi Algoritma KNN tanpa SMOTE dan klasifikasi algoritma KNN dengan SMOTE. Pada setiap percobaan yang dilakukan menerapkan nilai k yang bervariasi yaitu 1, 3, 5, 7 dan 9. Dimana label berita Objektif merupakan kelas Positif dan label berita Subjektif adalah kelas Negatif. Sehingga TP adalah berita berlabel Objektif yang diklasifikasikan Objektif oleh sistem, TN adalah berita Subjektif yang diklasifikasikan Subjektif oleh sistem, FP adalah berita berlabel Subjektif yang diklasifikasikan Objektif oleh sistem dan untuk FN adalah berita Objektif yang diklasifikasikan sebagai Subjektif.

Tabel 5. Hasil pengujian klasifikasi algoritma KNN

Performa	k=1	k=3	k=5	k=7	k=9
TP	159	165	171	172	176
TN	6	2	1	0	0
FP	18	22	23	24	24
FN	17	11	5	4	0
Akurasi	82,50	83,50	86,00	86,00	88,00
Presisi	0,89	0,88	0,88	0,87	0,88
Recall	0,90	0,93	0,97	0,97	1,00
F-measure	0,90	0,90	0,92	0,92	0,93

Tabel 6. Hasil pengujian klasifikasi SMOTE+KNN

Performa	k=1	k=3	k=5	k=7	k=9
TP	134	125	110	102	105
TN	174	175	176	176	176
FP	2	1	0	0	0
FN	42	51	66	74	21
Akurasi	87,50	85,22	80,68	78,90	79,82
Presisi	0,98	0,99	1,00	1,00	1,00
Recall	0,76	0,71	0,62	0,57	0,59
F-measure	0,89	0,86	0,83	0,82	0,81

Berdasarkan hasil penelitian pada kedua tabel diatas didapatkan bahwa pengaruh SMOTE dapat meningkatkan nilai akurasi untuk k=1 dan k=3. Dengan nilai akurasi k=1 meningkat 5,00 dan untuk nilai k=3 meningkat 1,72. Untuk nilai k=5, k=7 dan k=9 SMOTE dapat menurunkan nilai akurasi. Dengan nilai penurunan akurasi k=5 sebesar 4,80, k=7 sebesar 7,10 dan k=9 sebesar 8,10. SMOTE meningkatkan nilai presisi untuk semua nilai k dengan rincian nilai k=1 presisi meningkat 0,09, nilai k=3 meningkat 0,11, nilai k=5 meningkat 0,11, nilai k=7 meningkat 0,12 dan nilai k=9 meningkat 0,12. Untuk nilai *recall* dan *F-measure* SMOTE dapat menurunkan nilai keduanya untuk semua nilai k. Untuk k=1 *recall* menurun 0,14, dan *F-measure* menurun 0,05. Untuk k=3 *recall* menurun 0,22, dan *F-measure* menurun 0,07. Untuk k=5 *recall* menurun 0,34, dan *F-measure* menurun 0,16. Untuk k=7 *recall* menurun 0,40, dan *F-measure* menurun 0,19. Dan untuk k=9 *recall* menurun 0,41, dan *F-measure* menurun 0,18.

Performa klasifikasi algoritma KNN yang paling baik adalah saat nilai k=9 dengan nilai akurasi, presisi, recall dan *F-measure* paling tinggi dari nilai k lainnya yaitu sebesar 88,00 untuk akurasi, 0,88 untuk presisi, 1 untuk *recall* dan 0,93 untuk *F-measure*. Performa klasifikasi SMOTE+KNN yang paling baik adalah saat nilai k=1 dengan nilai akurasi, recall dan *F-measure* paling tinggi dari nilai k lainnya yaitu sebesar 87,50 untuk akurasi, 0,80 untuk presisi, 0,98 untuk *recall* dan 0,88 untuk *F-measure*.

4. Kesimpulan

Dataset yang didapatkan dari kompas.com merupakan dataset yang memiliki *imbalance class*. Sehingga pada penelitian ini diterapkan metode *Syntetic Minority Over Sampling Technique* (SMOTE) untuk mengatasi permasalahan *imbalance class* tersebut dan untuk memaksimalkan kinerja dari algoritma KNN. Hasil penelitian dengan menerapkan nilai k yang bervariasi yaitu 1, 3, 5, 7 dan 9 diperoleh bahwa penerapan SMOTE dalam menangani *imbalance class* pada klasifikasi objektivitas berita menghasilkan performa yang kurang efektif pada nilai k 5, 7 dan 9. Hal ini ditunjukkan saat performa algoritma KNN mengalami rata-rata penurunan nilai akurasi sebesar -6,67. Sedangkan untuk nilai k=1 dan k= 3 performa

algoritma KNN mengalami rata-rata peningkatan nilai akurasi sebesar 3,36. Berdasarkan penelitian ini didapatkan hasil klasifikasi algoritma KNN tanpa SMOTE lebih sensitif untuk berita objektif yang ditunjukkan dengan banyaknya nilai *true positif* (TP) atau berita objektif yang diklasifikasikan sebagai berita objektif. Dan klasifikasi algoritma KNN dengan SMOTE lebih sensitif terhadap berita subjektif yang ditunjukkan dengan banyaknya nilai *true negatif* (TN) atau berita subjektif yang diklasifikasikan menjadi berita subjektif.

Pada penelitian ini, secara umum menerapkan teknik SMOTE untuk mengatasi masalah *imbalance class* untuk memaksimalkan performa dari algoritma KNN dalam mengklasifikasikan berita online, namun beberapa hal dapat diterapkan untuk menyempurnakan penelitian ini, yaitu menggunakan algoritma klasifikasi lain, menggunakan fitur seleksi pada atribut yang digunakan dan penerapan teknik *sampling* yang lain untuk mengatasi masalah performa *imbalance class* dalam mengklasifikasikan berita online serta dapat menggunakan dataset dengan jumlah *instance* yang lebih banyak, dan dengan rasio ketidakseimbangan kelas yang lebih tinggi.

Daftar Rujukan

- [1] Y. Rizk and M. Awad, "Syntactic Genetic Algorithm for a Subjectivity Analysis of Sports Articles," *Resesrch Gate*, vol. 5, no. May, p. 7, 2018.
- [2] O. Sakti *et al.*, "Klasifikasi Teks Menggunakan Algoritma K-Nearest Neighbor pada kasus kinerja pemerintah di twitter," *e-Proceeding Eng.*, vol. 5, no. 3, pp. 8237–8248, 2018.
- [3] Ardiyansyah, P. A. Rahayuningsih, and Reza Maulana, "Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner," *J. Khatulistiwa Inform.*, vol. VI, no. 6, pp. 20–28, 2018.
- [4] A. Y. Triyanto and R. Kusumaningrum, "Implementasi Teknik Sampling untuk Mengatasi Imbalanced Data pada Penentuan Status Gizi Balita dengan Menggunakan Learning Vector Quantization," *IPTEK-KOM*, vol. 19, no. 6, pp. 39–50, 2017.
- [5] N. V. Chawla, K. W. Bowyer, and L. O. Hall, "Handling Imbalance Data Prediksi Churn menggunakan metode SMOTE dan KNN Based on Kernel," *e-Proceeding of Engineering.*, vol. 4, no. 117, pp. 1-15, 2017.
- [6] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan K-Nearest Neighbor," *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018.
- [7] V. M. Rumata, "Objektivitas Berita pada Media dalam Jaringan (Analisis Isi Berita Pemilihan Gubernur DKI Jakarta pada Detik news Selama Kampanye periode 1)," no. October 2016, 2017.
- [8] I. P. Sonya, "Analisis Web Scraping Untuk Data Bencana Alam Dengan Menggunakan Teknik Breadth-First," *Jurnal Inform. dan Komput.*, vol. 21, no. 12, pp. 69–77, 2016.
- [9] A. Fathan Hidayatullah, M. Rifqi Ma, and Arif, "Penerapan Text Mining dalam Klasifikasi Judul Skripsi," in *Seminar Nasional Aplikasi Teknologi Informasi (SNATI) Agustus*, 2016, pp. 1907–5022.
- [10] H. Leidiyana, "Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor," *J. Penelit. Ilmu Komputer, Syst. Embed. Log.*, vol. 1, no. 1, pp. 65–76, 2013.
- [11] W. Gata and Purnomo, "Akurasi Text Mining Menggunakan Algoritma K-Nearest Neighbour pada Data Content SMS-Gateway," *J. Format*, vol. 6, no. 5, pp. 1–5, 2017.
- [12] R. I. Pristiyanti, M. A. Fauzi, and L. Muflikhah, "Sentiment Analysis Peringkasan Review Film Menggunakan Metode Information Gain dan K-Nearest Neighbor," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 3, pp. 1179–1186, 2018.
- [13] U. Pujiyanto, "Strategi Resampling berbasis Centroid untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Perangkat Lunak," *Tekno*, vol. 25, no. Maret, pp. 1–6, 2016.
- [14] Wikipedia, 2019. "Berita" [Online] (Updated 10 Januari 2019) Tersedia di: <https://id.wikipedia.org/wiki/Berita>