

Optimasi TextBlob Menggunakan Support Vector Machine untuk Analisis Sentimen (Studi Kasus Layanan Telkomsel)

Kevin Perdana¹, Titania Pricillia², Zulfachmi³

^{1,2,3} STT Indonesia Tanjungpinang

Jln. Poma Air No. 28 Tanjungpinang Kepulauan Riau Indonesia

¹kevin@sttindonesia.ac.id

²titania@sttindonesia.ac.id

³zulfachmi@sttindonesia.ac.id

Intisari— Analisis sentimen mengacu kepada teknik Natural Language Processing yang tergolong kedalam *Unsupervised Learning* untuk mengidentifikasi opini positif, negatif, atau netral. Opini-opini tersebut banyak datang melalui Twitter, dikarenakan sosial media ini cukup efektif dan efisien dalam berkomentar karena hanya dapat menulis maksimal 140 karakter. Dari penelitian sebelumnya, nilai dari keakuratan analisis sentimen yang dilakukan oleh salah satu library NLP yaitu TextBlob, telah diketahui bahwa *Unsupervised Learning* tidak menghasilkan nilai yang begitu baik. Dengan studi kasus layanan Telkomsel penulis mengambil dataset dari Twitter dan hasil analisis dengan TextBlob hanya menunjukkan nilai akurasi sebesar 58,59%. Optimasi dilakukan dengan menambah metode Support Vector Machine yang termasuk dalam kategori *Supervised Learning*. Hasil terbaik yang didapat dari penelitian ini adalah nilai akurasi menunjukkan 75%.

Kata kunci— Machine Learning, Support Vector Machine, TextBlob, Twitter, Telkomsel.

Abstract— *Sentiment analysis refers to Natural Language Processing techniques that are classified as Unsupervised Learning to identify positive, negative, or neutral opinions. Many of these opinions come through Twitter, because social media is quite effective and efficient in commenting because it can only write a maximum of 140 characters. From previous research, the value of the accuracy of the sentiment analysis carried out by one of the NLP libraries, namely TextBlob, has shown that Unsupervised Learning does not produce such good scores. With the Telkomsel service case study the writer took the dataset from Twitter and the results of the analysis with TextBlob only showed a value of 58.59%. Optimization is done by adding the Support Vector Machine method which is included in the Supervised Learning category. The best results obtained from this study are values that show 75%.*

Keywords— Machine Learning, Support Vector Machine, TextBlob, Twitter, Telkomsel.

I. PENDAHULUAN

Analisis sentimen mengacu kepada teknik Natural Language Processing dan tergolong kedalam *Unsupervised Learning*. Digunakan untuk mengidentifikasi opini dan sentimen dan mengklasifikasikannya menjadi label positif, negatif, atau netral. Library yang paling umum digunakan untuk analisis sentimen adalah TextBlob [1].

Twitter adalah salah satu sosial media yang membuat pengguna mengekspresikan opini tentang produk, layanan atau hal lain. Twitter termasuk yang efektif dan efisien karena pengguna hanya dapat menulis hingga 140 karakter [2]. Dataset didapatkan melalui Twitter API menggunakan library Python yaitu Tweepy. Tweet dipantau dan ditangkap dengan memanfaatkan kelas StreamListener [3].

Pada penelitian sebelumnya dengan judul "Twitter Sentiment Analysis Using Lexical or Rule Based Approach: A Case Study" yang hanya menggunakan cara *Unsupervised Learning*, ditemukan hasil yang tidak begitu akurat dikarenakan mengandung bahasa gaul. Dengan demikian data mentah setelah diklasifikasikan menjadi positif, negatif atau netral perlu dianalisis menggunakan algoritma *supervised machine learning* agar diperoleh hasil dan akurasi yang tepat [1].

Salah satu metode *Supervised Learning* adalah Support Vector Machine yang mengklasifikasikan kelas berdasarkan pola dari proses training. Klasifikasi dilakukan dengan garis pembatas (*hyperlane*) yang memisahkan antar kelas opini. [4]. SVM banyak digunakan di beberapa penelitian untuk masalah klasifikasi sentimen. Hasilnya menunjukkan SVM menghasilkan yang terbaik ketika dibandingkan dengan pendekatan lain [5].

Berdasarkan uraian tersebut dan penelitian sebelumnya, maka penulis tertarik untuk melakukan penelitian dengan tujuan melihat apakah nilai akurasi meningkat ketika menambah SVM (*Supervised Learning*) dari klasifikasi menggunakan TextBlob (*Unsupervised Learning*) dengan mengambil data cuitan tentang pelayanan Telkomsel.

Penulis memilih bahasa pemrograman Python karena didukung oleh library yang banyak, lebih dari 2000 [6] dan diantaranya banyak untuk kebutuhan *data science*. Dan Python dibuat untuk mengatasi kelemahan dan lebih baik dari Matlab dan juga R.

II. METODE PENELITIAN

Dalam penelitian ini, penulis membagi beberapa tahapan besar yang secara detail akan dijabarkan pada bab Pembahasan.

- Koleksi data.
- Menganalisa dengan TextBlob.
- Klasifikasi dengan Support Vector Machine.

III. HASIL DAN PEMBAHASAN

A. Koleksi Data

- Langkah awal adalah mengkoleksi data adalah melalui Twitter API, dimana untuk terhubung harus memiliki token.
- Setelah autentikasi berhasil, *tweet* dikumpulkan secara acak sebanyak 258 *tweet* menggunakan *keyword* Telkomsel.
- Data yang ditampung diubah menjadi bahasa Inggris untuk dianalisis di TextBlob.

B. Analisa dengan TextBlob

- 258 *tweet* dikumpulkan di satu variabel dengan tipe data List.
- Kemudian dianalisis menggunakan TextBlob. Dan dari masing-masing *tweet* tersebut ditetapkan polaritasnya apakah bermakna positif, negatif, atau netral.
- Dari variabel List, data ditampilkan ke Pandas DataFrame dengan field *id_tweet*, *text*, *sentimen*.
- Dan data dari Pandas DataFrame di-*export* menjadi file CSV.
- Setelah itu dilakukan pengecekan *score*.

C. Klasifikasi dengan Support Vector Machine

- File CSV di-*import* dan dibagi menjadi 80% data training dan 20% data test.
- Melakukan proses klasifikasi menggunakan kernel Linear.
- Pengecekan *score*.

Hasilnya adalah analisis menggunakan TextBlob kurang menghasilkan nilai akurasi yang baik, dengan ditunjukkan persentase dibawah ini.

```
Score Tweet Positif = 38.95348837209303%
Score Tweet Negatif = 57.89473684210527%
Score Tweet Netral = 78.9272030651341%
Average Score : 58.59180942644414%
```

Gambar 1. *Score* yang didapatkan dari TextBlob

Optimasi dilakukan menggunakan Support Vector Machine dengan kernel Linear sehingga menghasilkan keakuratan sebesar 75%.

```
BEGIN. SUPPORT VECTOR MACHINE.....
      precision    recall  f1-score   support

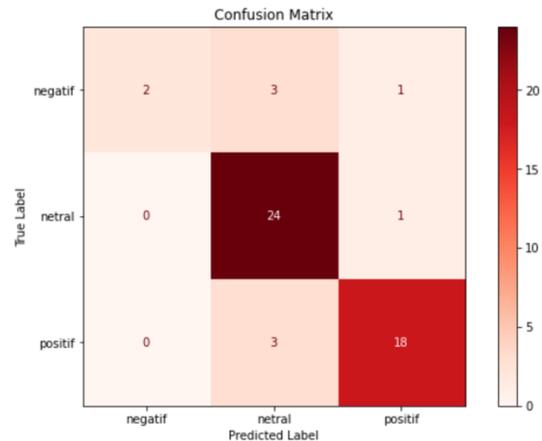
negatif      1.00      0.33      0.50         6
netral       0.80      0.96      0.87        25
positif      0.90      0.86      0.88        21

accuracy                0.85        52
macro avg           0.90      0.72      0.75        52
weighted avg       0.86      0.85      0.83        52

END. SUPPORT VECTOR MACHINE
```

Gambar 2. *Score* yang didapatkan dari Support Vector Machine

Perbandingan hasil prediksi dengan hasil sebenarnya digunakan pengujian dengan Confusion Matrix [7].



Gambar 3. Confusion Matriks

Confusion Matriks diatas adalah terdapat *Predicted Value* dan *Actual Value*. *Predicted Value* adalah hasil klasifikasi yang dilakukan oleh model dan *Actual Value* adalah nilai yang sebenarnya.

Definisi Confusion Matriks ditemukan beberapa variabel matriks sebagai berikut.

True Positive (TP) :

- 18 untuk *tweet* positif.
- 24 untuk *tweet* netral.
- 2 untuk *tweet* negatif.

True Negatif (TN) untuk masing-masing kelas adalah yang bukan dalam rentang *predicted* dan *actual value*.

False Positive (FP) nilai yang dikeluarkan *predicted value* per kelas tidak sesuai dengan *actual value*.

False Negative (FN) adalah dua nilai prediksi yang dikeluarkan adalah berbeda dengan per kelas *actual value*.

Pada gambar 2, nilai precision menunjukkan proporsi kasus prediksi positif yang benar-benar positif pada *actual value* [8]. Nilai precision didapat dengan formula :

$$\text{precision} = \frac{TP}{TP + FP} \quad [9]$$

Nilai recall adalah proporsi kasus *actual value* tepat dengan apa yang diprediksi [8].

Nilai recall didapat dengan formula :

$$\text{recall} = \frac{TP}{TP + FN} \text{ [9]}$$

Nilai F1 didapat dengan formula :

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \text{ [10]}$$

Macro average adalah nilai dengan bobot yang sama untuk setiap kelas [11].

IV. KESIMPULAN

Support Vector Machine berhasil melakukan klasifikasi sentimen dengan lebih baik. Nilai akurasi yang didapatkan adalah 75%. Lebih baik dari yang hanya menggunakan *Unsupervised Learning*.

Untuk penelitian selanjutnya, analisis sentimen ini dapat dibuat menjadi *realtime* (berdasarkan data *tweet* yang terbaru).

REFERENSI

- [1] S. Zahoor and R. Rohilla, "Twitter Sentiment Analysis Using Lexical or Rule Based Approach: A Case Study," *ICRITO 2020 - IEEE 8th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir.)*, pp. 537–542, 2020, doi: 10.1109/ICRITO48877.2020.9197910.
- [2] K. Sussolaikah, "Analysis of Telkomsel ' S User Internet Service User Satisfaction in Social," *2018 3rd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng.*, pp. 259–264, 2018.
- [3] C. Kariya and P. Khodke, "Twitter sentiment analysis," *2020 Int. Conf. Emerg. Technol. INCET 2020*, pp. 7–9, 2020, doi: 10.1109/INCET49848.2020.9154143.
- [4] B. W. Sari and F. F. Haranto, "Implementasi Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter Terhadap Pelayanan Telkom Dan Biznet," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 171–176, 2019, doi: 10.33480/pilar.v15i2.699.
- [5] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, "Classification with Support Vector Machines," *Math. Mach. Learn.*, pp. 335–356, 2020, doi: 10.1017/9781108679930.014.
- [6] A. Kumar and S. P. Panda, "A Survey: How Python Pitches in IT-World," *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Perspectives Prospect. Com. 2019*, pp. 248–251, 2019, doi: 10.1109/COMITCon.2019.8862251.
- [7] K. A. B. Permana, M. Sudarma, and W. G. Ariastina, "Analisis Rating Sentimen pada Video di Media Sosial Youtube Menggunakan STRUCT-SVM," *Maj. Ilm. Teknol. Elektro*, vol. 18, no. 1, p. 113, 2019, doi: 10.24843/mite.2019.v18i01.p17.
- [8] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," pp. 37–63, 2020, [Online]. Available: <http://arxiv.org/abs/2010.16061>.
- [9] P. Zhang and W. Su, "Statistical inference on recall, precision and average precision under random selection," *Proc. - 2012 9th Int. Conf. Fuzzy Syst. Knowl. Discov. FSKD 2012*, no. Fskd, pp. 1348–1352, 2012, doi: 10.1109/FSKD.2012.6234049.
- [10] J. Sepúlveda and S. A. Velastín, "F1 score assesment of Gaussian mixture background subtraction algorithms using the MuHAVi dataset," *IET Semin. Dig.*, vol. 2015, no. 5, pp. 1–6, 2015, doi: 10.1049/ic.2015.0106.
- [11] D. Zhang, J. Wang, and X. Zhao, "Estimating the uncertainty of average F1 scores," *ICTIR 2015 - Proc. 2015 ACM SIGIR Int. Conf. Theory Inf. Retr.*, pp. 317–320, 2015, doi: 10.1145/2808194.2809488.