KLASIFIKASI EMOSI TEKS BERBAHASA INDONESIA MENGGUNAKAN METODE MAXIMUM ENTROPY

Tedy Agastya Dwi Permana[*], Firdaus Sholihin[**], Fika Hastarita[***]
Jurusan Teknik Informatika, Fakutas Teknik, Universitas Trunojoyo Madura
Jl. Raya Telang PO. BOX 2 Kamal, Bangkalan, Madura, 691962
E-Mail: [*]Tedy.infor@gmail.com, [**]fsolihin@if.trunojoyo.ac.id,
[***]hastarita.fika@gmail.com

ABSTRAK

Seiring dengan berkembangnya teknologi, banyak sekali cara seseorang untuk melakukan interaksi terhadap orang lain dengan cara pertukaran informasi, dalam bentuk teks. Di dalam informasi berbentuk teks tersebut tidak hanya terdapat suatu keterangan pesan namun juga terdapat suatu keterangan yang menyatakan emosi dari informasi tersebut. Oleh karena itu, perlu adanya perancangan dan pembuatan suatu aplikasi yang dapat melakukan klasifikasi emosi dari suatu informasi sehingga para pengguna tidak hanya mengetahui keterangan pesan namun juga dapat mengetahui keterangan emosi yang terdapat dalam informasi tersebut. Maximum Entropy (MaxEnt) merupakan salah satu metode klasifikasi dokumen dengan menggunakan nilai distribusi probabilitas dalam proses pengklasifikasiannya. Dengan aplikasi ini diharapkan mampu untuk membantu pengguna untuk mengetahui keterangan emosi dari suatu informasi yang berbentuk teks. Dari uji coba sistem klasifikasi dengan menggunakan data query diperoleh hasil akurasi sebesar 93%, dan dengan menggunakan data crawler twitter diperoleh hasil akurasi sebesar 63%, kemudian dengan menggunakan data sample diperoleh hasil rata-tata akurasi sebesar 64,6%.

Kata Kunci: Klasifikasi emosi, Maximum Entropy (MaxEnt), distribusi probabilitas.

1. Pendahuluan

Analisis komputasi emosi telah dianggap sebagai penelitian yang menantang dan menarik. Para peneliti mengandalkan berbagai isyarat seperti sensor fisiologis dan ekspresi wajah untuk mengidentifikasi emosi manusia. Ada beberapa karya sebelumnya yang bekerja dengan input tekstual untuk menganalisis emosi ini.

Analisis emosi (EA) dari teks adalah suatu pekerjaan untuk memprediksi emosi dalam sepotong teks [1]. Proliferasi pendekatan analisis emosi telah dimotivasi oleh munculnya Web 2.0. Karena populernya media sosial, orang mengekspresikan emosi di web akhir-akhir ini. Selain itu, weblog, forum diskusi, dan komentar mudah diakses. Semua bentuk teks ini tersedia untuk orang yang tertarik dalam penelitian analisis emosi, yang dapat menerapkan algoritma pemrosesan bahasa alami variabel pada data teks tersebut.

Penelitian tentang klasifikasi emosi dalam sebuah teks banyak dilakukan dalam banyak bahasa. Danisman & Alpkocak menciptakan aplikasi yang diberinama Feeler untuk mengklasifikasikan emosi dalam teks bahasa Inggris menggunakan metode Vector Space Model. Metode Vector Space Model sendiri sudah banyak

digunakan dalam banyak penelitian di bidang temu kembali informasi berbasis teks [2]–[5]. Masih menggunakan teks berbahasa Inggris, Inkpen dkk [6] mengusulkan sebuah pendekatan hirarki untuk mengklasifikasikan emosi dalam sebuah blog.

Sementara itu Wen & Wan[7] melakukan klasifikasi emosi teks dalam sebuah microblog berbahasa China menggunakan aturan pengurutan kelas (class sequential rule).

Penelitian klasifikasi emosi dalam teks berbasaha Indonesia dilakukan oleh Sumpeno dkk [8] yang mengevaluasi kinerja dari metode Maximum Entropy (MaxEnt), Support Vector Machine (SVM) dan Na¨ıve Bayes (NB). Dalam penelitian tersebut disimpulkan bahwa metode MaxEnt memiliki performa terbaik dibandingkan dua metode pembandingnya.

Berdasarkan latar belakang diatas maka peneliti tertarik melakukan penelitian tentang klasifikasi emosi berdasarkan teks bahasa Indonesia dengan menggunakan metode Maximum Entropy (MaxEnt). Dalam penelitian ini dipilih 7 kategori yang akan dijadikan kelas klasifikasi teks, yaitu senang, sedih, marah, takut, jijik, malu, dan menyesal karena kategori tersebut telah umum digunakan dalam berkomunikasi.

2. Tinjauan Pustaka

2.2. *Maximum Entropy*

Metode MaxEnt telah banyak digunakan untuk berbagai tugas NLP karena terbukti sebagai algoritma yang praktis dan kompetitif dalam domain ini [8], [9]. Gagasan yang memotivasi di belakang entropi maksimum adalah bahwa seseorang harus memilih model yang paling seragam yang juga memenuhi kendala yang diberikan. Sebagai contoh, pertimbangkan tugas klasifikasi teks empat arah di mana kita hanya diberi tahu bahwa rata-rata 40% dokumen dengan kata "profesor" di dalamnya ada di kelas fakultas. Secara intuitif, ketika diberikan dokumen dengan "profesor" di dalamnya, kita akan mengatakan itu memiliki peluang 40% untuk menjadi dokumen fakultas, dan peluang 20% untuk masing-masing dari tiga kelas lainnya. Jika sebuah dokumen tidak memiliki "profesor" kita akan menebak distribusi kelas seragam, masing-masing 25%. Model ini adalah model entropi maksimum yang sesuai dengan batasan yang diketahui. Menghitung model itu mudah dalam contoh ini, tetapi ketika ada banyak kendala untuk memuaskan, teknik yang ketat diperlukan untuk menemukan solusi yang optimal.

Dalam formulasi yang paling umum, entropi maksimum dapat digunakan untuk memperkirakan distribusi probabilitas. Dalam makalah ini kami tertarik pada klasifikasi; jadi kami membatasi diskusi lebih lanjut untuk mempelajari distribusi bersyarat dari data pelatihan berlabel. Secara khusus, kami mempelajari distribusi bersyarat label kelas yang diberikan dokumen.

2.2.Model *Maximum Entropy*

Pemodelan dengan menggunakan Maximum Entropy digunakan untuk mencari distribusi yang seragam dari suatu kumpulan probabilitas. Untuk menerapkan entropi maksimum ke domain, kita perlu memilih serangkaian fitur yang akan digunakan untuk mengatur kendala. Untuk klasifikasi teks dengan entropi maksimum, kita menggunakan jumlah kata sebagai fitur.

Fakta dari data *training* dapat dinyatakan sebagai fungsi fitur f_j:(a,b)→{0,1} yang dipelajari dari kumpulan dokumen B, dengan ketentuan $F_j(a,b)$

- 1, jika F_j muncul di dokumen b pada kelas a
- 0, jika F_j tidak muncul di dokumen b pada kelas a

Sebagai contoh, misalkan B adalah kumpulan artikel berita olahraga {b_1, b_2, … ,b_n} dan A adalah himpunan cabang olahraga {basket, sepak bola, tenis}. Probabilitas p(basket|b_1) adalah probabilitas kemungkinan cabang olahraga basket dibahas pada artikel berita b_1. Fitur $F_j(a|b)$ dapat juga dilihat sebagai probabilitas kemunculan sebuah kata pada dokumen b untuk cabang olahraga a, sebagai contoh $F_j(a,b)$ =

- 1, jika kata liga muncul di berita b pada cabang olahraga a
- 0, jika kata liga tidak muncul di berita b pada cabang olahraga a

Batasan-batasan atau fakta-fakta yang telah diketahui dalam proses pembelajaran dengan data *training*, dimasukkan dalam penghitungan sebagai nilai ekspektasi dari suatu nilai fitur [9], sebagai berikut:

$$Epfj = \sum P(a,b) \,.\, fj(a,b) \qquad (1)$$

dengan $P(a,b)$, dengan merupakan probabilitas kemunculan bersama (*joint probability*) pasangan a dan b. Probabilitas ini dapat dihitung dengan persamaan [10] sebagai berikut:

$$P(a|b) = \frac{f(b.a)+1}{f(a)+|W|} \qquad (2)$$

dengan $P(a|b)$ merupakan nilai kemunculan dokumen b pada kategori a dan $f(a)$ merupakan jumlah keseluruhan kata pada kategori a dan $|W|$ merupakan jumlah keseluruhan kata/fitur.

$$P(a) = \frac{f(a)}{|D|} \qquad (3)$$

dengan $P(a)$ merupakan probabilistic kategori a dan $|D|$ merupakan jumlah data *training* dan $f(a)$ merupakan jumlah dokumen yang dimiliki kategori a.

Dari probabilitas tersebut, langkah selanjutnya adalah mencari nilai maximum dengan memanfaatkan perhitungan [10] berikut :

$$P(a|b) = \frac{P(a).P(Wkj\,|a)}{\sum P(a).P(Wkj\,|a)} \qquad (4)$$

dimana Wkj merupakan fitur atau kata dari dokumen b dan $P(a)$ merupakan nilai probabilitas suatu kategori.

3. Metode

3.1. Data

Percobaan pada penelitian ini menggunakan data yang merupakan hasil kuisioner tentang ekspresi dan reaksi. Data yang digunakan adalah dokumen dari ISEAR *(International Survey on Emotion Antecedents and Reactions),* seperti yang terangkum dalam Tabel 1.

Tabel 1. Daftar Kategori data ISEAR

Emosi	Jumlah
Angry	1.087
Disgust	1.061
Fear	1.083
Joy	1.086
Sad	1.082
Shame	1.054
Guilt	1.068
Total	**7.521**

ISEAR terdiri dari 7.521 kalimat

yang merupakan hasil kuisioner tentang ekspresi dan reaksi dari 7 emosi dalam kehidupan sehari – hari, seperti senang, sedih, marah, takut, jijik, bersalah, dan malu.

3.3. *Filtering*

Filtering dilakukan untuk menghilangkan kata-kata yang tidak dipakai atau tidak penting yang disebut dengan *stopwords* serta menghilangkan tanda baca. Proses penghilangan stopwords dapat dilihat pada makalah Mastur dkk [11]. Daftar stopwords bahasa Indonesia yang digunakan dalam penelitian ini di dapat dari http://fpmipa.upi.edu/staff/yudi/stop_words_list.txt.

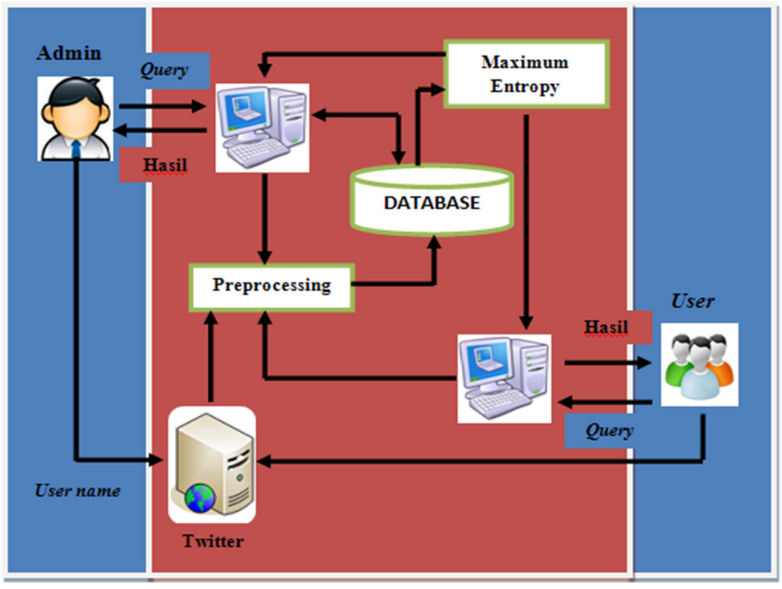
3.4. *Stemming*

Proses persiapan dokumen terakhir adalah *stemming*, yaitu proses pencarian kata dasar dari teks yang telah dipecah menjadi kata. Metode *stemming* yang digunakan pada penelitian ini menggunakan *Enhanced Confix Stripping* (ECS). Karena *Enhanced Confix Stripping* (ECS) memiliki kinerja yang paling baik dalam proses *stemming* pada Bahasa Indonesia.

3.5. Perancangan Sistem

Secara garis besar gambaran dari sistem yang dibuat dapat dilihat pada Gambar 1.

Database, yang didefinisikan sebagai kumpulan dokumen teks yang digunakan sebagai data *training*. Dalam proses pengklasifikasian penelitian penelitianini akan menggunakan data *training* dari ISEAR *(International Survey on Emotion Antecedents and Reactions).* ISEAR terdiri dari 7.521 kalimat yang merupakan hasil kuisioner tentang ekspresi dan reaksi dari 7 emosi dalam kehidupan sehari – hari, seperti senang, sedih, marah, takut, jijik, bersalah, dan malu.



Gambar 1. Gambaran Umum Rancangan Sistem

Twitter, merupakan proses memperoleh suatu informasi dari web tertentu. Dalam penelitian ini proses *crawling web* akan dilakukan dengan menggunakan format *API twitter* ke alamat web yang merupakan *web social networking,* yaitu *twitter.*

Username, didefinisikan sebagai pengganti *password* untuk dapat mengakses *tweet* akun twitter. *Username* dalam penelitian ini, baik *user* maupun admin dapat menggunakannya.

Query, yang merupakan suatu permintaan untuk dokumen yang menyangkut topik tertentu yang menarik bagi pengguna. *Query* di dalam penelitian ini dapat digunakan oleh *user* maupun admin itu sendiri dimana dari *query* tersebut dapat diketahui emosi yang terkandung di dalamnya.

Preprocessing, yakni proses tokenizing, filtering dan stemming sehingga dokumen tersebut menjadi set term index untuk mempermudah proses pengklasifikasian data sampel untuk setiap kategori. Setelah didapat data sampel disetiap kategorinya

Maximum Entropy, proses menganalisa *query* dengan *dataset* pada *database* untuk membentuk model probabilitas yang kemudian akan dilanjutkan dengan penentuan hasil klasifikasi.

4. Hasil dan Pembahasan

Proses uji coba pada sistem ini dilakukan 3 uji coba, yaitu uji coba sistem klasifikasi (*query*), uji coba sistem klasifikasi (*twitter*), uji coba sistem klasifikasi (Data Sampel).

4.1. Uji Coba Sistem Klasifikasi Menggunakan *Query*

Berikut adalah deskripsi data untuk uji coba *query* dari sistem ini:

- Jumlah data sampel seluruh kategori: 7.521 data.
- Jumlah data testing *query*: 14 data.

Uji coba *query* ini hanya dilakukan 1 kali dengan data testing sebanyak 14 data dari 2 data setiap kategori, data *query* yang diambil menggunakan data yang *query* yang diinputkan oleh user. Dari skenario uji coba *query* ini akan didapat hasil akurasi beserta *error rate* dari klasifikasi, seperti yang tersaji pada pada Tabel 2. Akurasi dari hasil uji coba tersebut adalah sebagai berikut:

Accuracy = 2+2+2+2+2+2+1 / 14
= 244 / 350 = 0,93 = 93%

Error Rate = 1 – Accuracy
= 1 – 0,93 = 0,07 = 7%

Tabel 2. Hasil Klasifikasi Menggunakan *Query*

		Aktual						
prediksi		sng	Sdh	Mrh	Tkt	jjk	Brlh	Ml
	Sng	2	0	0	0	0	0	0
	Sdh	0	2	0	0	0	0	0
	Mrh	0	0	2	0	0	0	0
	Tkt	0	0	0	2	0	0	0
	Jjk	0	0	0	0	2	0	0
	Brlh	0	0	0	0	0	2	1
	Ml	0	0	0	0	0	0	1

Adapun penjelasan dari setiap kata yang digunakan adalah sebagai berikut:

1. Sng = Senang
2. Sdh = Sedih
3. Mrh = Marah
4. Tkt = Takut
5. Jjk = Jijik
6. Brlh = Bersalah
7. Ml = Malu

4.2. Uji Coba Sistem Klasifikasi Menggunakan *Twitter*

Berikut adalah deskripsi data untuk uji coba *twitter* dari sistem ini:

- Jumlah data sampel seluruh kategori: 7.521 data.
- Jumlah data testing *query*: 103 data.

Uji coba *twitter* ini hanya dilakukan 1 kali dengan data testing sebanyak 103 data. data *twitter* yang diambil menggunakan fasilitas *API twitter,* dimana data testing diambil dari 5 akun twitter. Dari skenario uji coba *twitter* ini akan didapat hasil akurasi beserta *error rate* dari klasifikasi dengan hasil seperti pada Tabel 3.

Tabel 3. Hasil Klasifikasi Menggunakan *Twitter*

		Aktual						
		sng	Sdh	Mrh	Tkt	jjk	Brlh	Ml
Prediksi	Sng	41	0	1	0	0	0	1
	Sdh	3	32	5	2	3	4	2
	Mrh	4	3	36	2	7	10	5
	Tkt	0	6	3	42	4	1	6
	Jjk	0	0	2	3	34	1	3
	Brlh	1	7	1	1	1	32	6
	Ml	1	2	2	0	1	2	27

Sedangkan akurasi yang diperoleh adalah sebagai berikut:

Accuracy = 8+9+10+12+10+8+12 / 103
= 69 / 103 = 0,67 = 67%

Error Rate = 1 – Accuracy
= 1 – 0,67 = 0,33 = 33%

4.3. Uji Coba Sistem Klasifikasi Menggunakan Data Sampel

Berikut adalah deskripsi data untuk uji coba dari sistem ini:

- Jumlah data sampel seluruh kategori: 7.521 data.
- Jumlah proses klasifikasi yang dilakukan: 3x.
- Jumlah data testing tiap proses: 50, 100, dan 300 data.

Uji coba ini dilakukan sebanyak 3 kali dengan data testing sebanyak 50, 100, dan 300 data dari seiap kategori, data yang diambil secara random dari masing-masing kategori. Dari skenario uji coba ini akan didapat hasil akurasi beserta *error rate* dari klasifikasi. Hasil dari ujicoba untuk setiap skenario dapat dilihat pada Tabel 4, Tabel 5 dan Tabel 6.

Akurasi dari skenario 1 yang menggunakan 50 data sampel adalah sebagai berikut :

Accuracy = 41+32+36+42+34+32+27 / 350 = 244 / 350 = 0,69 = 69%

Error Rate = 1 – Accuracy
= 1 – 0,69 = 0,31 = 31%

Untuk skenario 2 dengan 100 data hasil akurasinya adalah :

Accuracy = 70+62+60+72+58+61+56 / 700 = 439 / 700 = 0,63 = 63%

Error Rate = 1 – Accuracy
= 1 – 0,63 = 0,37 = 37%

Dan untuk skenario 3 dengan 300 data sampel diperoleh akurasi :

Accuracy = 208 +196 +185+ 217 +180 + 175 +147 / 2100
= 1308 / 2100 = 0,62 = 62%

Error Rate = 1 – Accuracy
= 1 – 0,62 = 0,38 = 38%

Tabel 4. Hasil Klasifikasi Menggunakan 50 Data Sampel

		Aktual						
		Sng	Sdh	Mrh	Tkt	Ml	Brlh	Jjk
Prediksi	Sng	8	1	0	0	2	0	0
	Sdh	0	9	0	0	0	1	0
	Mrh	2	1	10	1	1	2	1
	Tkt	0	1	0	12	0	0	2
	Ml	1	1	0	0	10	2	0
	Brlh	2	1	1	2	2	8	0
	jjk	0	1	4	0	0	2	12

Tabel 5. Hasil Klasifikasi Menggunakan 100 Data Sampel

		Aktual						
		sng	Sdh	Mrh	Tkt	Jjk	Brlh	Ml
Prediksi	Sng	208	25	20	22	5	9	15
	Sdh	23	196	15	20	18	20	19
	Mrh	20	21	185	16	42	40	42
	Tkt	14	25	22	217	28	30	23
	Jjk	7	9	19	12	180	11	17
	Brlh	15	15	25	9	13	175	37
	Ml	13	9	14	4	14	15	147

Tabel 6. Hasil Klasifikasi Menggunakan 300 Data Sampel

		Aktual						
		sng	Sdh	Mrh	Tkt	Jjk	Brlh	Ml
Prediksi	Sng	70	11	4	8	6	8	0
	Sdh	3	62	9	3	3	6	12
	Mrh	5	6	60	9	15	12	9
	Tkt	10	3	10	72	6	5	6
	Jjk	0	2	5	6	58	2	4
	Brlh	7	9	9	2	5	61	13
	Ml	5	7	3	0	7	5	56

5. Penutup

5.1. Kesimpulan

Setelah menyelesaikan penelitian tentang sistem klasifikasi dengan menerapkan metode *maximum entropy* dalam mengklasifikasikan teks berdasarkan bahasa Indonesia serta melakukan uji coba dan evaluasi sistem, maka dapat ditarik kesimpulan sebagai berikut :

1. Untuk mengetahui emosi yang terkandung dalam sebuah teks maka pada penelitian penelitianini menggunakan data *training* yang merupakan hasil kuisioner tentang ekspresi dan reaksi, yaitu dokumen dari ISEAR *(International Survey on Emotion Antecedents and Reactions)* yang telah diterjemahkan ke dalam bahasa Indonesia.
2. Dalam proses melakukan *crawling web* pada akun *twitter* menggunakan format *API twitter* yang telah disediakan oleh *twitter* untuk mendapatkan *tweet* dari *username user.* Dari *tweet* inilah yang akan diketahui emosi yang terkandung di dalam *tweet* tersebut.
3. Pada sistem klasifikasi emosi berdasarkan teks berbahasa Indonesia dengan menggunakan metode *Maximum Entropy.* Telah didapatkan beberapa hasil akurasi sebagai berikut :
   - Skenario uji coba data *query* yang ada dengan menghasilkan nilai akurasi sebesar 93% dan *error rate* sebesar 7% pada saat uji coba klasifikasi *query*.
   - Skenario uji coba data *twitter* yang ada dengan menghasilkan nilai akurasi sebesar 63% dan *error rate* sebesar 37% pada saat uji coba klasifikasi data sampel.
   - Skenario uji coba data sampel yang ada dengan menghasilkan nilai rata-rata akurasi sebesar 64,6% dan *error rate* sebesar 35,4% pada saat uji coba klasifikasi data sampel.
4. Pada penelitian ini yang menggunakan metode *maximum entropy* jumlah kategori yang ada pada data tidak memberikan pengaruh pada kinerja sistem klasifikasi. Akan tetapi, tingkat seragaman data pada tiap kategori akan menurunkan hasil akurasi sistem klasifikasi.

5.2. Saran

Dari hasil uji coba yang telah dilakukan terhadap sistem klasifikasi dengan menerapkan metode maximum

entropy dalam mengklasifikasikan teks berdasarkan bahasa Indonesia dapat diberikan beberapa saran sebagai berikut:

1. Perlu dilakukan pemilihan data sampel yang mampu mewakili ciri-ciri yang signifikan dari masing-masing kategori.
2. Pemilihan jumlah data pelatihan yang optimal agar didapatkan hasil klasifikasi yang lebih baik

6. Daftar Pustaka

[1] V. Tripathi, A. Joshi, and P. Bhattacharyya, "Emotion Analysis from Text: A Survey," *Cfilt.Iitb.Ac.in*, 2015.

[2] F. A. Hermawati and D. A. Zuhdi, "Aplikasi Sistem Temu Kembali Dokumen dengan Metode Vector Space Model," *KONVERGENSI*, vol. 5, no. 2, pp. 38–49, 2009.

[3] G. Tsatsaronis and V. Panagiotopoulou, "A generalized vector space model for text retrieval based on semantic relatedness," *EACL 2009 - 12th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc.*, no. April, pp. 70–78, 2009.

[4] J. N. Singh and S. K. Dwivedi, "Performance Analysis of Layered Vector Space Model in Web Information Retrieval," *Int. J. Appl. Inf. Syst.*, vol. 8, no. 5, pp. 7–15, 2015.

[5] P. Harcourt and R. B. Japheth, "Application of Vector Space Model to Query Ranking and Information Retrieval," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 6, no. 5, pp. 42–47, 2016.

[6] D. Inkpen, F. Keshtkar, and D. Ghazi, "Analysis and generation of emotion in texts," in *KNOWLEDGE ENGINEERING: PRINCIPLE AND TECHNIQUE, KEPT 2009, International Conference on Knowledge Engineering Principles and Techniques Selected Papers*, 2009, pp. 3–14.

[7] S. Wen and X. Wan, "Emotion classification in microblog texts using class sequential rules," *Proc. Natl. Conf. Artif. Intell.*, vol. 1, pp. 187–193, 2014.

[8] S. Sumpeno, S. Member, A. Z. Arifin, and I. M. Hariadi, "A Performance Evaluation of Classifiers Employ Language Dependent Tools for Indonesian Text," in *11TH SEMINAR ON INTELLIGENT TECHNOLOGY AND ITS APPLICATIONS (SITIA 2010)*, 2010, pp. 1–6.

[9] K. Nigam, J. Lafferty, and A. Mccallum, "Using Maximum Entropy for Text Classification," in *IJCAI-99 Workshop on Machine Learning for Information Filtering*, 1999, pp. 61–67.

[10] B. Liu, X. Li, W. S. Lee, and P. S. Yu, "Text classification by labeling words," in *Proceedings of the National Conference on Artificial Intelligence*, 2004, pp. 425–430.

[11] M. Mastur, F. H. Rachman, and F. Solihin, "Efektifikats Penggunaan Stoplist Kata Umum dari Dokumen Hasil Klasifikasi Pretopology," *KONVERGENSI*, vol. 13, no. 1, pp. 1–10, 2013.