



**SKRIPSI**

**PENDEKATAN *MULTITOPIC STACKING ENSEMBLE* UNTUK KLASIFIKASI ULASAN EMOSI PADA MEDIA SOSIAL BERBASIS LSA (STUDI KASUS : PT. JAWAPOS MEDIA TELEVISI)**

**JASMINE AULIA**  
NPM 22083010074

**DOSEN PEMBIMBING**  
Amri Muhaimin, S.Stat., M.Stat., M.S.  
Andri Fauzan Adziima, S.Si., M.Si.

**KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI  
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR  
FAKULTAS ILMU KOMPUTER  
PROGRAM STUDI SAINS DATA  
SURABAYA  
2026**

**LEMBAR PENGESAHAN**

**PENDEKATAN *MULTITOPIC STACKING ENSEMBLE* UNTUK  
KLASIFIKASI ULASAN EMOSI PADA MEDIA SOSIAL BERBASIS LSA  
(STUDI KASUS : PT. JAWAPOS MEDIA TELEVISI)**

Oleh:  
JASMINE AULIA  
NPM. 22083010074

Telah dipertahankan di hadapan dan diterima oleh Tim Penguji Sidang Skripsi Program Studi Sains Data Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jawa Timur pada Tanggal 09 Juni 2026:

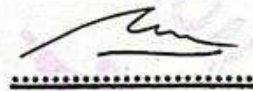
Menyetujui,

Amri Muhaimin, S.Stat., M.Stat., M.S.  
NIP. 19950723 202406 1 002



(Pembimbing I)

Andri Fauzan Adzilma, M.Si.  
NIP. 19950512 202406 1 001



(Pembimbing II)

Kartika Maulida Hindrayani, S.Kom., M.Kom.  
NIP. 19920909 202203 2 009



(Ketua Penguji)

Alfan Rizaldy Pratama, S.Tr.T., M.Tr.Kom.  
NIP. 19990606 202406 1 001



(Penguji I)

Mengetahui,

**Dekan Fakultas Ilmu Komputer**



Prof. Dr. Ir. Novirina Hendrasarie, MT.

NIP. 19681126 199403 2 001

**LEMBAR PERSETUJUAN**

**PENDEKATAN *MULTITOPIC STACKING ENSEMBLE* UNTUK  
KLASIFIKASI ULASAN EMOSI PADA MEDIA SOSIAL BERBASIS LSA  
(STUDI KASUS : PT. JAWAPOS MEDIA TELEVISI)**

Oleh:  
JASMINE AULIA  
NPM. 22083010074

Telah disetujui untuk mengikuti Ujian Skripsi



**Menyetujui,**

**PLT Koordinator Program Studi Sains Data  
Fakultas Ilmu Komputer**

**Dr.Ir. I Gede Susrama Mas Diyasa, S.T., M.T., IPU.**

**NIP. 19700619 2021211 009**

## SURAT PERNYATAAN BEBAS PLAGIASI

Saya yang bertanda tangan di bawah ini:

Nama Mahasiswa : Jasmine Aulia  
NPM : 22083010074  
Program : Sarjana (S1)  
Program Studi : Sains Data  
Fakultas : Fakultas Ilmu Komputer

Menyatakan bahwa dalam dokumen ilmiah Skripsi ini tidak terdapat bagian dari karya ilmiah lain yang telah diajukan untuk memperoleh gelar akademik di suatu lembaga Pendidikan Tinggi, dan juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang/lembaga lain, kecuali yang secara tertulis disitasi dalam dokumen ini dan disebutkan secara lengkap dalam daftar pustaka.

Dan saya menyatakan bahwa dokumen ilmiah ini bebas dari unsur-unsur plagiasi. Apabila di kemudian hari ditemukan indikasi plagiat pada Skripsi ini, saya bersedia menerima sanksi sesuai dengan peraturan perundang-undangan yang berlaku.

Demikian surat pernyataan ini saya buat dengan sesungguhnya tanpa ada paksaan dari siapapun juga dan untuk dipergunakan sebagaimana mestinya.



Surabaya, 09 Juni 2026  
Yang Membuat Pernyataan,



JASMINE AULIA  
NPM. 22083010074

## ABSTRAK

Nama Mahasiswa / NPM : Jasmine Aulia / 22083010074  
Judul Skripsi : Pendekatan NLP Pendekatan *Multitopic Stacking Ensemble* Untuk Klasifikasi Ulasan Emosi Pada Sosial Media Berbasis LSA (Studi Kasus : PT. Jawapos Media Televisi)  
Dosen Pembimbing : 1. Amri Muhaimin, S.Stat., M.Stat., M.S.  
2. Andri Fauzan Adziima, S.Si., M.Si.

Perkembangan media sosial telah mendorong perubahan pola interaksi audiens terhadap konten televisi, termasuk pada PT Jawapos Media Televisi (JTV) sebagai stasiun televisi lokal di Jawa Timur. Komentar audiens pada media sosial JTV dapat menjadi sumber informasi penting untuk memahami respons dan persepsi audiens, namun data tersebut bersifat tidak terstruktur, menggunakan bahasa informal, serta mengandung variasi ekspresi emosi sehingga sulit dianalisis secara manual. Penelitian ini bertujuan untuk mengklasifikasikan emosi komentar audiens media sosial JTV menggunakan metode Stacking Ensemble serta menganalisis topik dominan menggunakan *Latent Semantic Analysis* (LSA). Data komentar dikonsolidasikan ke dalam tiga kelas emosi, yaitu *joy*, *netral*, dan *negative*. Tahapan penelitian meliputi pra-pemrosesan teks, pengujian kombinasi pra-pemrosesan dan n-gram, ekstraksi fitur menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF), penerapan SMOTE pada data latih, pemodelan klasifikasi emosi, evaluasi model, serta analisis topik menggunakan LSA. Model *Stacking Ensemble* dibangun dengan mengombinasikan Random Forest, *Support Vector Machine* (SVM), dan XGBoost sebagai base learners, serta *Logistic Regression* sebagai *meta learner*. Berdasarkan hasil evaluasi, model terbaik memperoleh nilai *accuracy* sebesar 73,95% dan F1-score sebesar 73,41%. Selain itu, LSA digunakan untuk mengidentifikasi pola topik dominan pada komentar audiens. Sebagai luaran, penelitian ini juga mengembangkan aplikasi berbasis Streamlit untuk menampilkan hasil prediksi, visualisasi, dan analisis komentar secara interaktif. Dengan demikian, penelitian ini memberikan kontribusi dalam penerapan *Natural Language Processing* (NLP) untuk membantu memahami persepsi audiens media sosial secara lebih efisien.

**Kata Kunci:** JTV, *Latent Semantic Analysis*, *Natural Language Processing*, Sosial Media, *Stacking Ensemble*.

## **ABSTRACT**

*Student Name / NPM* : Jasmine Aulia / 22083010074  
*Undergraduate thesis title* : *Multitopic Stacking Ensemble Approach For Emotional Review Classification on LSA-Based Social Media (Case Study : PT. Jawapos Media Televisi)*  
*Advisors* : 1. Amri Muhaimin, S.Stat., M.Stat., M.S.  
2. Andri Fauzan Adziima, S.Si., M.Si.

*The development of social media has encouraged changes in audience interaction patterns toward television content, including at PT Jawapos Media Televisi (JTV), a local television station in East Java. Audience comments on JTV's social media platforms can serve as an important source of information for understanding audience responses and perceptions. However, these data are unstructured, use informal language, and contain various emotional expressions, making them difficult to analyze manually. This study aims to classify the emotions of JTV social media audience comments using the Stacking Ensemble method and to analyze dominant topics using Latent Semantic Analysis (LSA). The comment data were consolidated into three emotion classes, namely joy, neutral, and negative. The research stages include text preprocessing, testing preprocessing and n-gram combinations, feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF), applying SMOTE to the training data, emotion classification modeling, model evaluation, and topic analysis using LSA. The Stacking Ensemble model was built by combining Random Forest, Support Vector Machine (SVM), and XGBoost as base learners, with Logistic Regression as the meta learner. Based on the evaluation results, the best model achieved an accuracy of 73.95% and an F1-score of 73.41%. In addition, LSA was used to identify dominant topic patterns in audience comments. As an output, this study also developed a Streamlit-based application to display prediction results, visualizations, and comment analysis interactively. Thus, this study contributes to the application of Natural Language Processing (NLP) in helping to understand social media audience perceptions more efficiently.*

**Keywords:** *JTV, Latent Semantic Analysis, Natural Language Processing, Social Media, Stacking Ensemble.*

## KATA PENGANTAR

Puji syukur kehadirat Allah SWT atas segala rahmat, hidayah dan karunia-Nya kepada penulis sehingga skripsi dengan judul **“Pendekatan *Multitopic Stacking Ensemble* untuk Klasifikasi Ulasan Emosi Pada Sosial Media Berbasis LSA (Studi Kasus : PT. Jawapos Media Televisi)”** dapat terselesaikan dengan baik. Penulis mengucapkan terima kasih kepada Bapak Amri Muhaimin, S.Stat., M.Stat., M.S. selaku Dosen Pembimbing pertama dan Bapak Andri Fauzan Adziima, S.Si., M.Si selaku Dosen Pembimbing kedua yang telah meluangkan waktunya untuk memberikan bimbingan, nasehat serta motivasi kepada penulis. Selain itu, selama penyusunan skripsi penulis juga banyak menerima bantuan dari berbagai pihak. Untuk itu penulis mengucapkan terima kasih kepada:

1. Allah SWT yang telah memberikan rahmat dan hidayah-Nya sehingga skripsi ini dapat diselesaikan dengan baik dan tepat waktu.
2. Orang Tua penulis, Alm. Ayahanda tercinta, Pudjio. Beliau memang tidak sempat menemani penulis dalam perjalanan menyusun skripsi, namun selama hidupnya telah menjadi sosok panutan, sumber semangat, dan inspirasi penuh kepada penulis. Dan teruntuk ibunda tercinta, Andi Nurmalasari, yang telah senantiasa hadir dalam doa dan dukungannya. Terima kasih atas cinta dan kasihnya turut membersamai penulis hingga titik ini.
3. Ibu Prof. Dr. Ir. Novirina Hendrasarie, M.T. selaku Dekan Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur.
4. Bapak Dr. Eng. Ir. Dwi Arman Prasetya, S.T., M.T., IPU., ASEAN. Eng selaku Ketua Program Studi Sains Data Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur.
5. Dosen-dosen Program Studi Sains Data yang telah membimbing dan memberikan ilmu yang bermanfaat bagi penulis selama masa perkuliahan.
6. Teman-teman semasa kuliah, Melati Harapan yang telah membersamai penulis dari awal perkuliahan hingga saat ini. Dan seluruh teman-teman seperjuangan dari Program Studi Sains Data Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur.

7. Ucapan paling khusus yaitu terima kasih kepada diri sendiri yang telah menguatkan diri selama pengerjaan skripsi dengan segala tantangan yang dihadapinya, tetap bertahan meskipun rasa lelah dan ragu sering datang menghampiri. Terima kasih telah memilih untuk tidak menyerah, terus berjuang, dan berusaha memahami setiap proses dengan sabar. Terima kasih telah berani menghadapi tekanan, menjaga komitmen, dan tetap percaya pada kemampuan diri sendiri. Semua perjuangan, waktu, dan air mata yang telah tercurah menjadi bukti keteguhan dan dedikasi yang luar biasa. Semoga rasa syukur dan kebanggaan ini menjadi pengingat bahwa setiap usaha tidak pernah sia-sia, dan setiap langkah yang diambil merupakan bagian dari perjalanan menuju versi terbaik dari diri sendiri.

Penulis menyadari bahwa di dalam penyusunan skripsi ini banyak terdapat kekurangan. Untuk itu kritik dan saran yang membangun dari semua pihak sangat diharapkan demi kesempurnaan penulisan skripsi ini. Akhirnya, dengan segala keterbatasan yang penulis miliki semoga laporan ini dapat bermanfaat bagi semua pihak umumnya dan penulis pada khususnya.

Surabaya, bulan-tahun

Penulis

## DAFTAR ISI

HALAMAN SAMPUL .....	i
LEMBAR PENGESAHAN .....	ii
LEMBAR PERSETUJUAN .....	iii
SURAT PERNYATAAN BEBAS PLAGIASI .....	iv
ABSTRAK .....	v
<i>ABSTRACT</i> .....	vi
KATA PENGANTAR .....	viii
DAFTAR ISI .....	ix
DAFTAR GAMBAR .....	xi
DAFTAR TABEL .....	xii
DAFTAR LAMPIRAN .....	xiii
DAFTAR NOTASI .....	xiv
BAB I PENDAHULUAN .....	1
1.1. Latar Belakang .....	1
1.2. Rumusan Masalah .....	6
1.3. Batasan Masalah .....	6
1.4. Tujuan Penelitian .....	8
1.5. Manfaat Penelitian .....	8
BAB II TINJAUAN PUSTAKA .....	11
2.1. Penelitian Terdahulu .....	11
2.2. Kerangka Teori .....	15
2.2.1. Sosial Media .....	15
2.2.2. Media Televisi .....	16

2.2.3.	Emosi .....	17
2.2.4.	<i>Cohen's Kappa</i> .....	18
2.2.5.	<i>Natural Language Processing (NLP)</i> .....	20
2.2.6.	<i>Term Frequency-Inverse Document Frequency (TF-IDF)</i> .....	21
2.2.7.	<i>Topic Modelling</i> .....	23
2.2.8.	<i>Latent Semantic Analysis (LSA)</i> .....	24
2.2.9.	<i>Multitopic Analysis</i> .....	26
2.2.10.	SMOTE.....	27
2.2.11.	<i>Stacking Ensemble</i> .....	28
2.2.12.	Algoritma Klasifikasi Dasar ( <i>Base Learners</i> ) .....	29
2.2.13.	<i>Support Vector Machine (SVM)</i> .....	29
2.2.14.	Random Forest.....	34
2.2.15.	<i>Extreme Gradient Boosting (XGBoost)</i> .....	37
2.2.16.	<i>Meta Learner</i> .....	40
2.2.17.	<i>Logistic Regression</i> .....	42
2.2.18.	<i>Confusion Matrix</i> .....	44
2.2.19.	GUI.....	45
BAB III METODOLOGI PENELITIAN .....		48
3.1.	Variabel Penelitian dan Sumber Data.....	48
3.2.	Langkah Analisis .....	49
3.2.1.	Pengumpulan data.....	49
3.2.2.	Persiapan Data .....	58
3.2.3.	Pembagian Data .....	73
3.2.4.	Uji Kombinasi Pra-pemrosesan dan N-Gram .....	75
3.2.5.	Ekstraksi Fitur Menggunakan TF – IDF.....	78
3.2.6.	Pemodelan Klasifikasi Emosi .....	83
3.2.7.	Evaluasi Model .....	94
3.2.8.	<i>Latent Semantic Analysis (LSA)</i> .....	96
3.3.	Desain Sistem .....	102

BAB IV HASIL DAN PEMBAHASAN .....	98
4.1    Penerapan <i>Stacking Ensemble</i> dalam Klasifikasi Emosi .....	98
4.2    Evaluasi Performa Model Klasifikasi Emosi.....	101
4.2.1.    Evaluasi <i>Base Learners</i> pada Dataset Instagram .....	101
4.2.2.    Evaluasi <i>Base Learners</i> pada Dataset YouTube .....	102
4.2.3.    Evaluasi Model <i>Stacking Ensemble</i> .....	103
4.3    Penerapan <i>Latent Semantic Analysis (LSA)</i> dalam Ekstraksi Topik..	107
4.4    Implementasi <i>Graphical User Interface (GUI)</i> Berbasis Streamlit .....	112
BAB V PENUTUP .....	114
5.1    Kesimpulan .....	114
5.2    Saran Pengembangan .....	115
DAFTAR PUSTAKA .....	117
LAMPIRAN.....	119

## DAFTAR GAMBAR

<b>Gambar 2.1</b> Diagram Konseptual Analisis Multitopik.....	26
<b>Gambar 2.2</b> Diagram <i>Stacking Ensemble</i> .....	28
<b>Gambar 2.3</b> Klasifikasi SVM pada data biner dan multi-kelas [49] .....	29
<b>Gambar 3.1</b> Diagram Alir Penelitian.....	49
<b>Gambar 3.2</b> Tampilan halaman <i>browser extension</i> .....	50
<b>Gambar 3.3</b> Distribusi Label Emosi Komentar Youtube .....	59
<b>Gambar 3.4</b> Distribusi Label Emosi Komentar Instagram .....	60
<b>Gambar 3.5</b> Distribusi Konsolidasi Kelas Label Emosi Komentar Youtube .....	68
<b>Gambar 3.6</b> Distribusi Konsolidasi Kelas Label Emosi Komentar Instagram....	68
<b>Gambar 3.7</b> Diagram Alir <i>Stacking Ensemble</i> .....	91
<b>Gambar 3.7</b> Tampilan <i>Homepage</i> .....	103
<b>Gambar 3.8</b> Dashboard Analisis.....	103
<b>Gambar 4.1</b> <i>Classification Report</i> Model <i>Stacking</i> pada Dataset Instagram ....	104
<b>Gambar 4.2</b> <i>Confusion Matrix</i> Model <i>Stacking</i> pada Dataset Instagram.....	105
<b>Gambar 4.3</b> <i>Classification Report</i> Model <i>Stacking</i> pada Dataset Youtube .....	106
<b>Gambar 4.4</b> <i>Confusion Matrix</i> Model <i>Stacking</i> pada Dataset Youtube .....	106
<b>Gambar 4.5</b> Distribusi Topik Dominan LSA .....	109
<b>Gambar 4.6</b> Distribusi Topik Dominan Terhadap Label Emosi .....	109
<b>Gambar 4.7</b> <i>Homepage</i> .....	112
<b>Gambar 4.8</b> Halaman Prediksi Satu Komentar .....	113
<b>Gambar 4.9</b> Hasil Prediksi Satu Komentar.....	113
<b>Gambar 4.10</b> Tampilan Prediksi Banyak Data .....	114

<b>Gambar 4.11</b> Preview Dataset pada Prediksi Banyak Data .....	115
<b>Gambar 4.12</b> Hasil <i>Preprocessing</i> Data Komentar.....	115
<b>Gambar 4.13</b> Tampilan Tahap <i>Modelling Batch</i> .....	116
<b>Gambar 4.14</b> Preview Data <i>Preprocessing</i> pada Tahap <i>Modelling Batch</i> .....	117
<b>Gambar 4.15</b> Hasil Prediksi Emosi dan Topik Dominan.....	117
<b>Gambar 4.16</b> Visualisasi Distribusi Emosi dan Topik Dominan .....	118
<b>Gambar 4.17</b> Distribusi Topik Dominan dan Ringkasan Hasil Prediksi .....	119
<b>Gambar 4.18</b> <i>Wordcloud</i> Emosi <i>Joy</i> .....	119
<b>Gambar 4.19</b> <i>Wordcloud</i> Emosi <i>Netral</i> .....	120
<b>Gambar 4.20</b> <i>Wordcloud</i> Emosi <i>Negative</i> .....	121

## DAFTAR TABEL

Tabel 2.1. Penelitian Terdahulu .....	11
Tabel 2.2 Kategori interpretasi <i>cohen's kappa</i> .....	18
Tabel 2.3 <i>Confusion Matrix</i> .....	44
Tabel 3.1 Variabel Penelitian.....	48
Tabel 3.2 Data Komentar Youtube .....	55
Tabel 3.3 Data Komentar Instagram .....	56
Tabel 3.4 Distribusi Dataset.....	57
Tabel 3.5 Distribusi Dataset Labeling Emosi .....	59
Tabel 3.6 Distribusi Dataset Labeling Emosi 2 Anotator Instagram .....	61
Tabel 3.7 Distribusi Dataset Labeling Emosi 2 Anotator Instagram .....	61
Tabel 3.8 Perhitungan <i>expected agreement (Pe)</i> .....	62
Tabel 3.9 Distribusi Dataset Labeling Emosi 2 Anotator Youtube .....	64
Tabel 3.10 Perhitungan <i>expected agreement (Pe)</i> .....	65
Tabel 3.11 Hasil <i>Case folding</i> .....	69
Tabel 3.12 Hasil <i>Emoticon Conversion</i> .....	71
Tabel 3.13 Hasil <i>Cleaning</i> .....	71
Tabel 3.14 Hasil <i>Normalization</i> .....	72
Tabel 3.15 Hasil <i>Stopword Removal</i> .....	73
Tabel 3.16 Contoh Pembentukan n-gram .....	76
Tabel 3.17 Contoh dataset.....	78
Tabel 3.18 Hasil Perhitungan <i>Term Frequency</i> .....	79
Tabel 3.19 Hasil Perhitungan <i>DF</i> dan <i>IDF</i> .....	81
Tabel 3.20 Hasil Perhitungan <i>TF-IDF</i> .....	81

Tabel 3.21 Ilustrasi Input Model dari Matriks TF-IDF .....	84
Tabel 3.22 Konfigurasi Model Default, Manual, dan <i>Hyperparameter Tuning</i> ....	85
Tabel 3.23 Konfigurasi Model Default, Manual, dan <i>Hyperparameter Tuning</i> ....	86
Tabel 3.24 Ilustrasi Prediksi <i>Base Learners</i> pada <i>Stacking Ensemble</i> .....	91
Tabel 3.25 Ilustrasi Prediksi <i>Confusion Matrix Multiclass</i> .....	94
Tabel 3.26 Contoh <i>Confusion Matrix</i> .....	95
Tabel 3.27 Contoh Nilai Singular .....	97
Tabel 3.28 Contoh bentuk representasi hasil LSA .....	99
Tabel 3.29 Contoh bobot Topik 1 .....	101
Tabel 4.1 Top 10 Kombinasi <i>Preprocessing</i> Terbaik Komentar YouTube .....	98
Tabel 4.2 Top 10 Kombinasi <i>Preprocessing</i> Terbaik Komentar Instagram.....	99
Tabel 4.3 Hasil Evaluasi <i>Base Learners</i> (Instagram) .....	101
Tabel 4.4 Hasil Evaluasi <i>Base Learners</i> (Youtube) .....	102
Tabel 4.5 Hasil Evaluasi <i>Stacking</i> pada Dataset Instagram .....	103
Tabel 4.6 Hasil Evaluasi <i>Stacking</i> pada Dataset Youtube.....	103
Tabel 4.7 Hasil <i>Explained Variance</i> Instagram.....	107
Tabel 4.8 Hasil <i>Explained Variance</i> Youtube .....	108
Tabel 4.9 Hasil Pemodelan Topik LSA Instagram.....	110
Tabel 4.10 Hasil Pemodelan Topik LSA Youtube .....	111

## DAFTAR LAMPIRAN

**Lampiran 1.** Dataset

**Lampiran 2.** Aplikasi GUI Streamlit

**Lampiran 3.** *Letter of Acceptance* (LOA)

## DAFTAR NOTASI

$D$	:	Kumpulan dokumen atau ulasan komentar audiens sosial media JTV
$d_i$	:	Dokumen atau komentar ke- $i$
$t_j$	:	Kata ( <i>term</i> ) ke- $j$ dalam korpus
$n$	:	Jumlah total kata unik
$m$	:	Jumlah total kata dokumen atau ulasan
$X$	:	Matriks <i>term-document</i> hasil representasi TF-IDF
$w_{ij}$	:	Bobot TF-IDF dari kata ke- $j$ pada dokumen ke- $i$
$U, S, V^T$	:	matriks hasil dekomposisi SVD pada metode LSA
$k$	:	Jumlah topik yang dihasilkan oleh metode LSA
$X_{train}, X_{test}$	:	Data latih dan data uji untuk proses klasifikasi
$f_i(x)$	:	Model <i>base learner</i> ke- $i$ pada <i>Stacking Ensemble</i>
$g(\bullet)$	:	Model <i>meta learner</i> yang menggabungkan hasil prediksi <i>base learners</i>
$acc$	:	Nilai <i>accuracy</i> model
$p$	:	Nilai <i>precision</i> model
$R$	:	Nilai <i>recall</i> model
$f_i$	:	Nilai <i>f1-score</i> model
$Entropy(S)$	:	Ukuran ketidakpastian pada himpunan data S
$p_i$	:	Probabilitas kemunculan kelas ke- $I$ dalam himpunan data