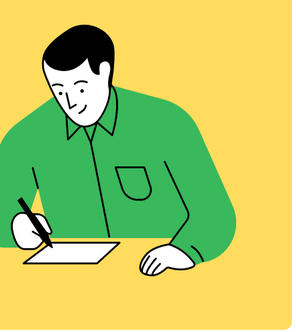
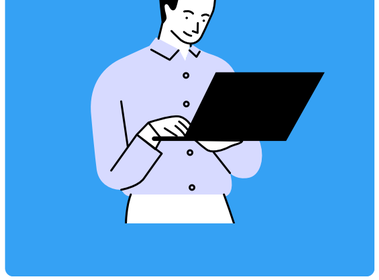
Presidential Election Sentiment Analysis

NLP PROJECT BY TEAM A

Indonesia AI
AI for Everyone, AI for Indonesia

Spärck





Meet the Team

Alfandy Surya

Dhanny Aryanda

Kusumo Jati Utomo

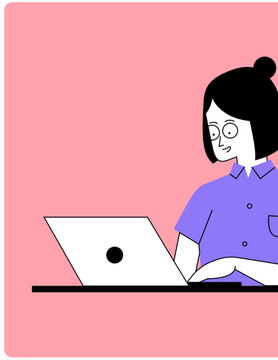
Alfian Ali Murtadho

Efrad Galio

Muhammad Habibullah

Anton Pranowo Medianto

Hendra Susanto



Spärck

Background &
Problem Statement

Twitter menjadi wadah besar diskusi publik saat Pilpres 2019 karena ada sekitar 5,7 juta tweet per hari terkait pilpres 2019. Jumlah data yang relatif besar tersebut dapat digunakan untuk berbagai analisis, salah satunya adalah **analisis sentimen.**

Analisis sentimen tweet penting untuk memahami opini publik terhadap para kandidat dan isu-isu terkait. Ini dapat membantu tim kampanye, media, peneliti, dan masyarakat. Menganalisis sentimen bisa mengungkap isu disorot publik, persepsi terhadap kandidat, dukungan, dan dampak misinformasi.

Spärck

Spärck

Objectives & Scope

Objectives:

Mengembangkan model machine learning dan deep learning untuk mengklasifikasikan tiga sentimen (positif, netral, dan negatif) dari tweet yang berkaitan dengan pilpres 2019 di Indonesia.

Scope:

- Menggunakan data tweet beserta label sentimen yang telah disediakan
- Model yang digunakan adalah Random Forest Classifier (machine learning) dan LSTM (deep learning)

Data Inspection

Dataset used: tweet.csv

Check null values:

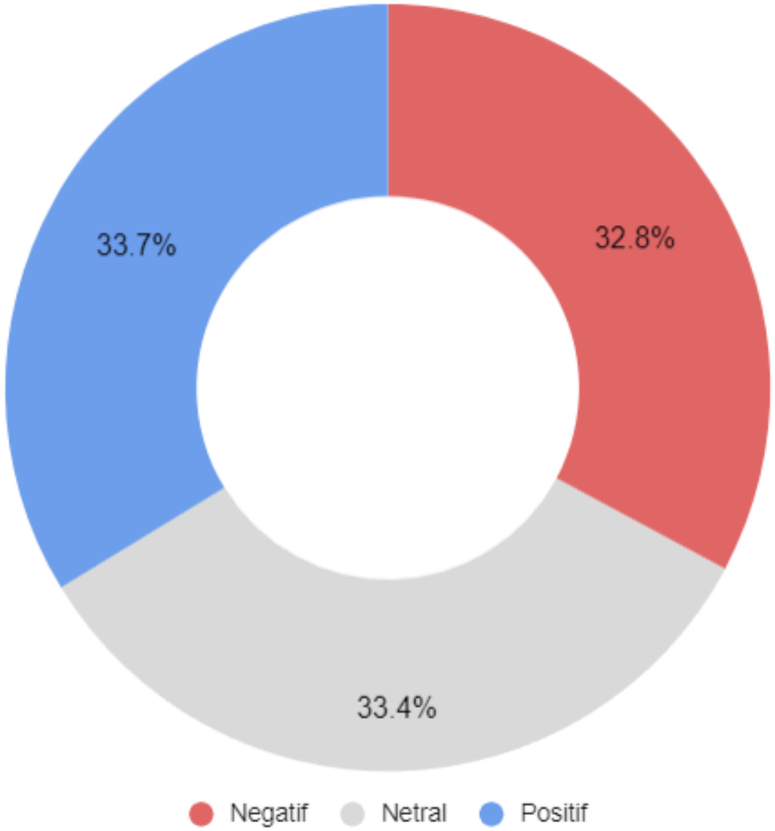
```
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   sentimen  1815 non-null   object
 1   tweet     1815 non-null   object
```

Number of null rows: 0

Check null values:

Number of duplicated rows: 0

Label Proportion:



Evaluation metrics: Accuracy

Text Preprocessing

Text Cleaning Process:

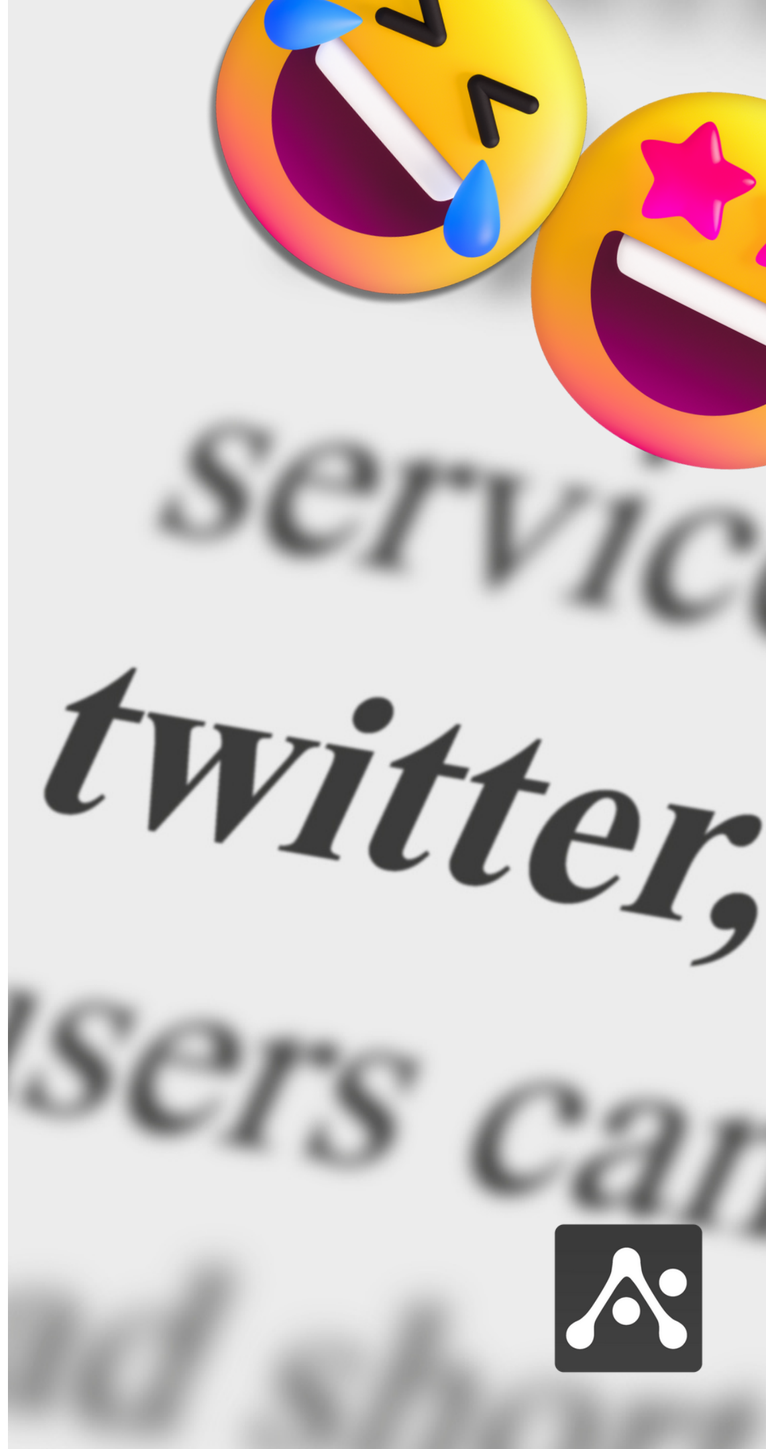
1. Emoji conversion
2. Remove hashtag, URL & HTML
3. Remove punctuation
4. Remove special character
5. Case folding
6. Processing slang words
7. Processing short words
8. Remove stopwords
9. Process number
10. Remove extra spaces

Text Normalization Process:

11. Lemmatization + Finalization

Text Tokenization Process:

12. Word tokenization



Tweet Example (Before & After) – 1

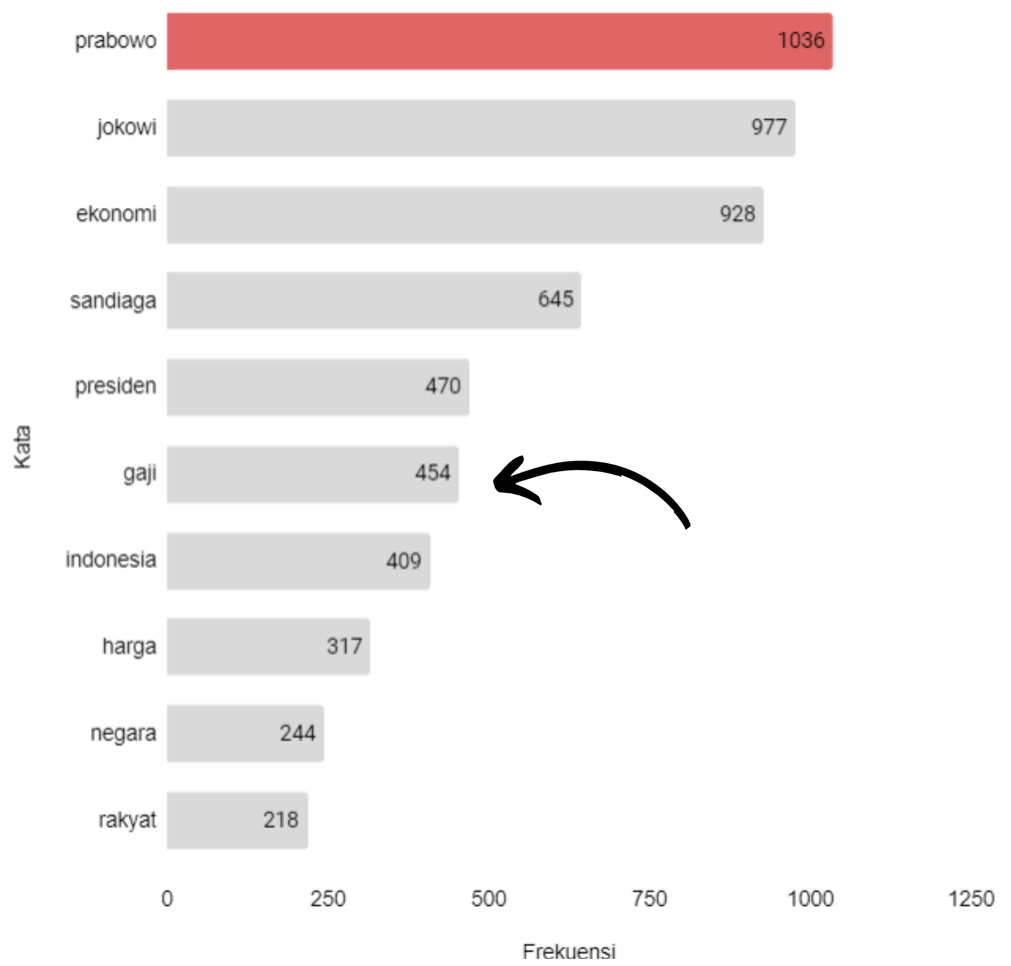
Text Preprocessing Stage	TWEET EXAMPLE (INDEX 323)
ORIGINAL TWEET	Kalo prabowo terpilih , ya benar ekonomi makin baik. Karna infrastruktur udah dibangun besar2an sama jokowi. Jadii dia tinggal duduk manis aja..🤣🤣
TEXT CLEANING	prabowo terpilih ekonomi baik infrastruktur dibangun besaran jokowi tinggal duduk manis
TEXT NORMALIZATION	prabowo pilih ekonomi baik infrastruktur bangun besar jokowi tinggal duduk manis 🤣🤣
TEXT TOKENIZATION	['prabowo', 'pilih', 'ekonomi', 'baik', 'infrastruktur', 'bangun', 'besar', 'jokowi', 'tinggal', 'duduk', 'manis', '🤣🤣']

Tweet Example (Before & After) - 2

Text Preprocessing Stage	TWEET EXAMPLE (INDEX 1326)
ORIGINAL TWEET	Siapa yg mau bayar gajinya? Wong #jokowi presidennya 2019 -2024💪💪
TEXT CLEANING	bayar gajinya orang presidennya 2019
TEXT NORMALIZATION	bayar gaji orang presiden 2019 💪💪
TEXT TOKENIZATION	['bayar', 'gaji', 'orang', 'presiden', '2019', '💪💪']

Exploratory Data Analysis – 1

Most Used Word:



Most Used bi-gram :

Bigram	Frekuensi
('prabowo', 'sandiaga')	496
('tidak', 'ambil')	146
('ambil', 'gaji')	127
('ekonomi', 'indonesia')	116
('sandiaga', 'tidak')	89
('presiden', 'jokowi')	88
('ekonomi', 'makro')	78
('pertumbuhan', 'ekonomi')	78
('susilo', 'bambang')	75

Exploratory Data Analysis – 2

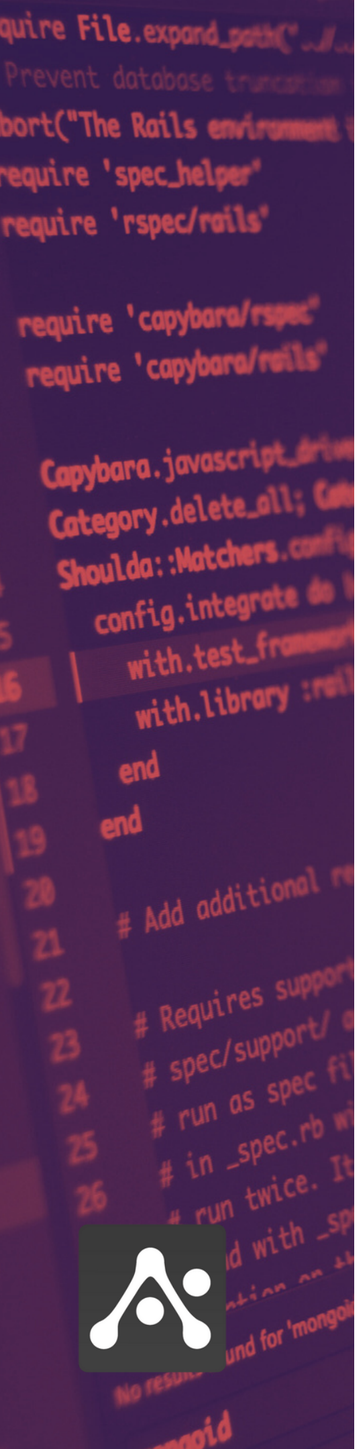
True Label Analysis (Top 10 most used words):

Kata	Negatif Count	Netral Count	Positif Count	Negatif %	Netral %	Positif %
ekonomi	314	333	246	35,2%	37,3%	27,5%
gaji	146	135	202	30,2%	28,0%	41,8%
harga	118	118	153	30,3%	30,3%	39,3%
indonesia	97	124	137	27,1%	34,6%	38,3%
jokowi	239	294	278	29,5%	36,3%	34,3%
negara	60	65	88	28,2%	30,5%	41,3%
prabowo	245	298	308	28,8%	35,0%	36,2%
presiden	101	123	149	27,1%	33,0%	39,9%
rakyat	80	37	89	38,8%	18,0%	43,2%
sandiaga	101	182	243	19,2%	34,6%	46,2%

Key Takeaway:

Dari tabel, kata-kata 'gaji', 'negara', 'rakyat', dan 'Sandiaga' sangat penting untuk meningkatkan performa model karena secara signifikan dapat membedakan sentimen positif.

Misalnya, 41,8% dari kemunculan kata 'gaji' terkait dengan sentimen positif. 'Gaji' digunakan oleh pasangan 02 Prabowo-Sandiaga sebagai janji kampanye untuk tidak mengambil gaji saat menjadi presiden.



Modelling

Data Splitting Strategy (Stratifed):

60% train | 20% validation | 20% test

Experimentation:

Random Forest (RF)

- Random Forest - SW Word2Vec SG (Baseline)
- Random Forest - SW Word2Vec CBOW (Baseline)
- Random Forest - SW Word2Vec SG (Tuned)
- Random Forest - SW Word2Vec CBOW (Tuned)

- Random Forest - Word2Vec SG (Baseline)
- Random Forest - Word2Vec CBOW (Baseline)
- Random Forest - Word2Vec SG (Tuned)
- Random Forest - Word2Vec CBOW (Tuned)

Long-Short Term Memory (LSTM)

- LSTM - Dengan text preprocessing (Baseline)
- LSTM - Tanpa text preprocessing (Baseline)
- LSTM - Dengan text preprocessing (Tuned)
- LSTM - Tanpa text preprocessing (Tuned)

Spärck

Result: Random Forest

Spärck

Tanpa Stopwords – Word2Vec

Hipotesis: text data tanpa stopwords akan meningkatkan performa model

SkipGram (Baseline)	SkipGram (Tuned)	CBOW (Baseline)	CBOW (Tuned)
Accuracy	**Accuracy**	**Accuracy**	**Accuracy**
Train: 0.99	Train: 0.76	Train: 0.99	Train: 0.76
Valid: 0.60	Valid: 0.61	Valid: 0.48	Valid: 0.50
Test: 0.53	Test: 0.52	Test: 0.46	Test: 0.45

Dengan Stopwords – Word2Vec

Hipotesis: text data dengan stopwords akan menurunkan performa model

SkipGram (Baseline)	SkipGram (Tuned)	CBOW (Baseline)	CBOW (Tuned)
Accuracy	**Accuracy**	**Accuracy**	**Accuracy**
Train: 1.00	Train: 071	Train: 1.00	Train: 0.87
Valid: 0.59	Valid: 0.61	Valid: 0.48	Valid: 0.54
Test: 0.57	Test: 0.55	Test: 0.51	Test: 0.49
	Best Model		

Hyperparameter Tuning (Best Model)

- Optimisasi hyperparameter menggunakan optuna (bayesian)
- Berikut ini adalah parameter hyperparameter untuk model terbaik:

{'n_estimators': 221, '
max_depth': 39,
min_samples_split': 0.041,
'min_samples_leaf': 0.0102,
'max_features': 'auto'}

Result: LSTM

Tanpa Preprocessing Data

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 198, 64)           807232

 lstm (LSTM)                 (None, 64)                33024

 dense (Dense)               (None, 3)                 195

=================================================================
Total params: 840451 (3.21 MB)
Trainable params: 840451 (3.21 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Accuracy
Train: 0.989, Valid: 0.586, Test: 0.575

Dengan Preprocessing Data

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_2 (Embedding)     (None, 98, 64)            266880

 lstm_2 (LSTM)               (None, 64)                33024

 dense_2 (Dense)             (None, 3)                 195

=================================================================
Total params: 300099 (1.14 MB)
Trainable params: 300099 (1.14 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

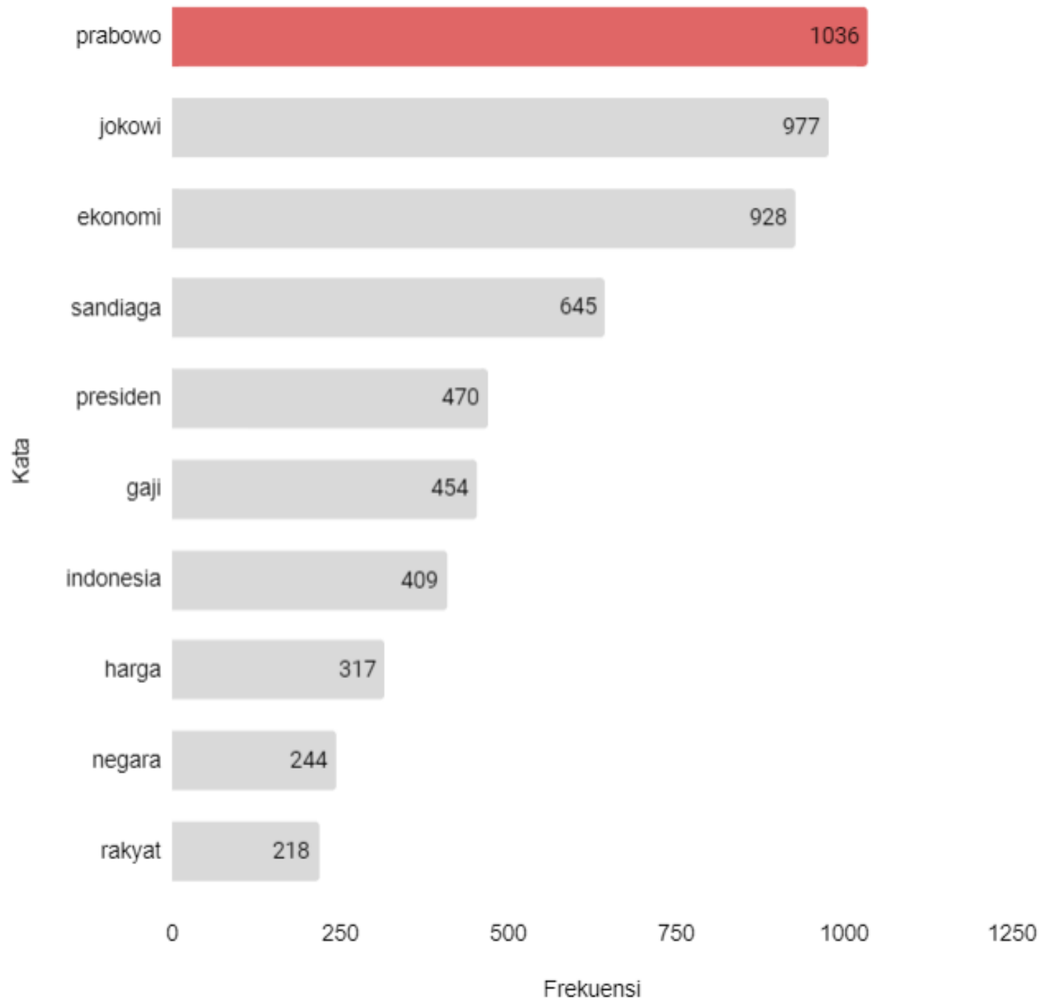
Accuracy
Train: 0.941, Valid: 0.575, Test: 0.550

Hasil Tuning untuk Accuracy tidak ada perubahan

Hyperparameter Tuning

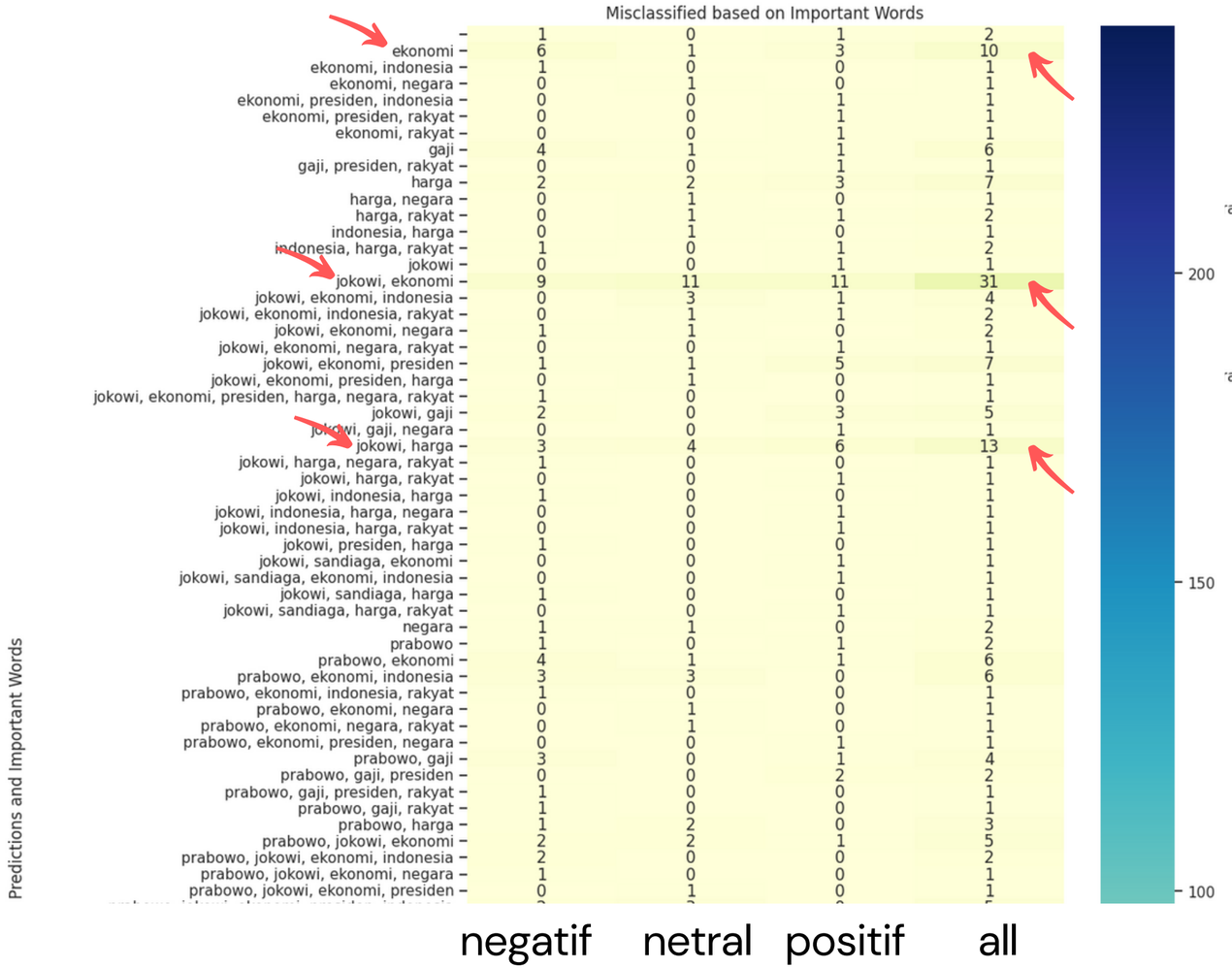
- Menambahkan dan Mengurangi LSTM Layer dan jumlah unit baik di LSTM Layer dan Embedding Layer.
- Menambahkan Dropout Layer, EarlyStopping, dan ReduceLROnPlateau.
- Mengubah EPOCH, BATCHSIZE dan OPTIMIZER.
- Mengubah MAX_WORDS dan MAX_LEN.

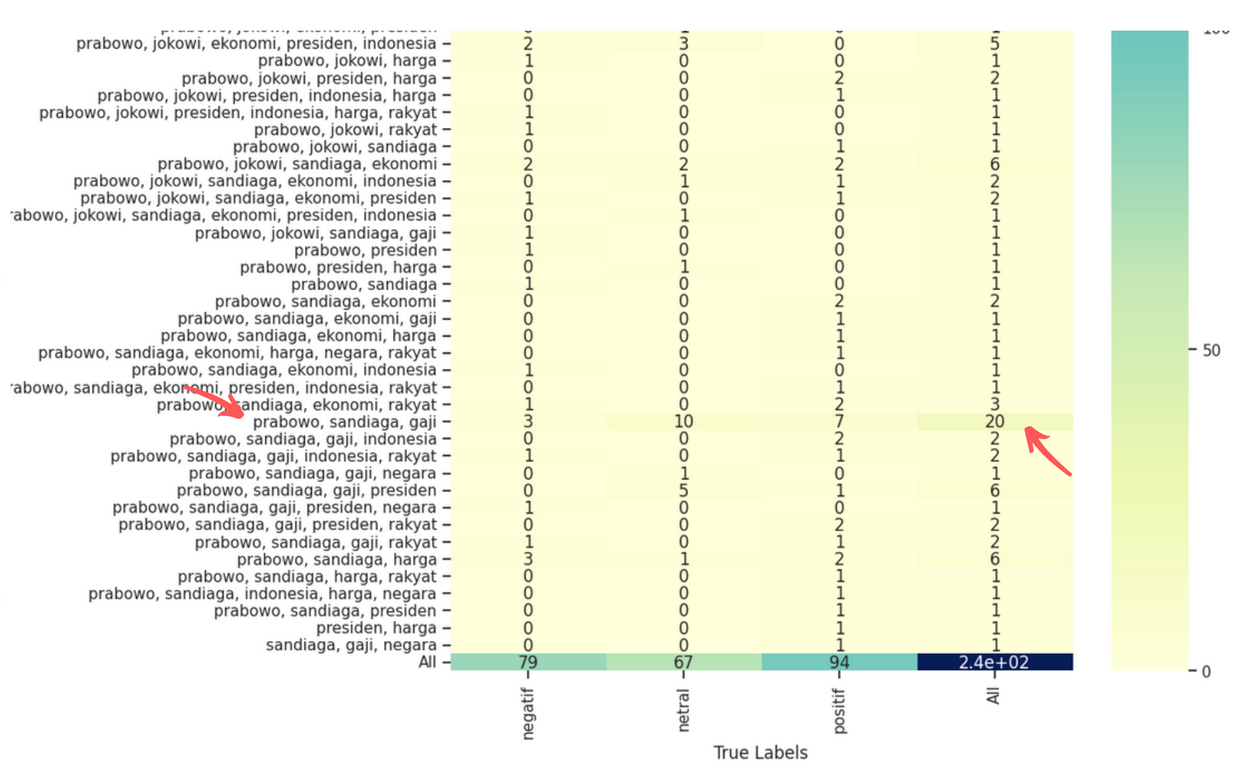
Error Analysis – 1



Error analysis dilakukan terhadap misklasifikasi pada 10 kata terbanyak yang digunakan dalam dataset.

Error Analysis - 2





Model LSTM belum dapat memprediksi sentimen dari tweet yang berkaitan dengan ekonomi, **jokowi ekonomi**, jokowi harga dan **prabowo sandiaga gaji**.

Error Analysis - 3

Kesimpulan

ekonomi --> Model belum dapat memprediksi kalimat yang tidak mengandung kata negatif, dan jika terdapat kata negatif model cenderung melabeli kalimat tersebut negatif.

Contoh 1
Tweet: Eh bung, tolong bedakan startup sebagai monetizing dan hobi yg di monetize, startup jelas ada nilai keekonomian yg besar, tapi hobi? Hanya segelintir orang yg melakukannya, dan bukan kebutuhan. Ya Allah masa harus saya jelaskan seh bedanya panjang lebar bung...
Label: Negatif
Predict: Positif

Contoh 2
Tweet: Alhamdulillah, dalam membangun ekonomi dan pemerataan yang adil pasti akan mengalami batu aral dan pasti akan ada pihak yang merasa jadi korban.... tapi apa yang bapak kerjakan dan lakukan saat ini saya percaya dan yakin suatu saat anak dan cucu saya yang merasakannya
Label: Positif
Predict: Negatif

jokowi ekonomi --> Model belum dapat memprediksi kalimat singkat yang jelas-jelas positif dan terdapat hasil preprocess yang tidak sesuai ekspetasi serta typo yang tidak teratasi.

Contoh 1
Tweet: Isu SDG dibahas saat Pk Jokowi menjelaskan kalau pertumbuhan ekonomi tdk berarti kalau ketimpangan tdk diperhatikan
Final Tweet: isu bahas jahat kelamin jokowi tumbuh ekonomi tidak timpang tidak perhati
Label: Positif
Predict: Netral

Contoh 2
Tweet: Nyatnaya Pemerinta di Era Jokowi Sangat Mendukung Perkembangan Ekonomi Digital. #17aprilcoblosbajuputih #coblos01JokowiKHMaruf #DaritimurUntukPresidenku #CoblosPutihJokowi #Jokowi2Periode #jokowiAmin
Final Tweet: nyatnaya pemerinta era jokowi dukung kembang ekonomi digital
Label: Positif
Predict: Negatif

Error Analysis - 4

Kesimpulan

jokowi harga --> Model kesulitan untuk memprediksi label netral yang memiliki kata-kata negatif atau positif dan label yang tidak sesuai

Contoh 1
Tweet: Tidak Cuma Cek Harga, Jokowi Juga Akan Perbaiki Pasar #2019JokowiKyaiMaruf #JokowiKHMarufUnggulanKita #DerinDiPRO2FM #InterBarça
Final Tweet: tidak cek harga jokowi baik pasar
Label: Netral
Predict: Negatif

Contoh 2
Tweet: Sebelumnya harga BBM di Papua senilai Rp 100.000/liter, sungguh drastik, di era Jokowi mampu menekan harga premium seharga Rp6.450/liter setara dengan harga BBM di seantero nusantara. #PiliOrangbaik #PilihJelasIslamnya #PilihBajuPutih
Final Tweet: harga bahan bakar minyak papua nilai rupiah liter sungguh drastik era jokowi tekan harga premium harga rp6 liter tara harga bahan bakar minyak antero nusantara
Label: Netral
Predict: Positif

prabowo sandiaga gaji --> Banyak sekali label yang tidak sesuai sehingga membuat model bingung.

Contoh 1
Tweet: Dengan mencoblos PROBOWO SANDI anda ikut beramal. Karena selama 5 tahun, gaji Prabowo sandi akan di sumbangkan ke fakir miskin,kaum duafa,dll
Label: Positif
Predict: Netral

Contoh 2
Tweet: JikaTerpilih, Prabowo – Sandi Tak akan Ambil😂Gaji
Label: Netral
Predict: Positif

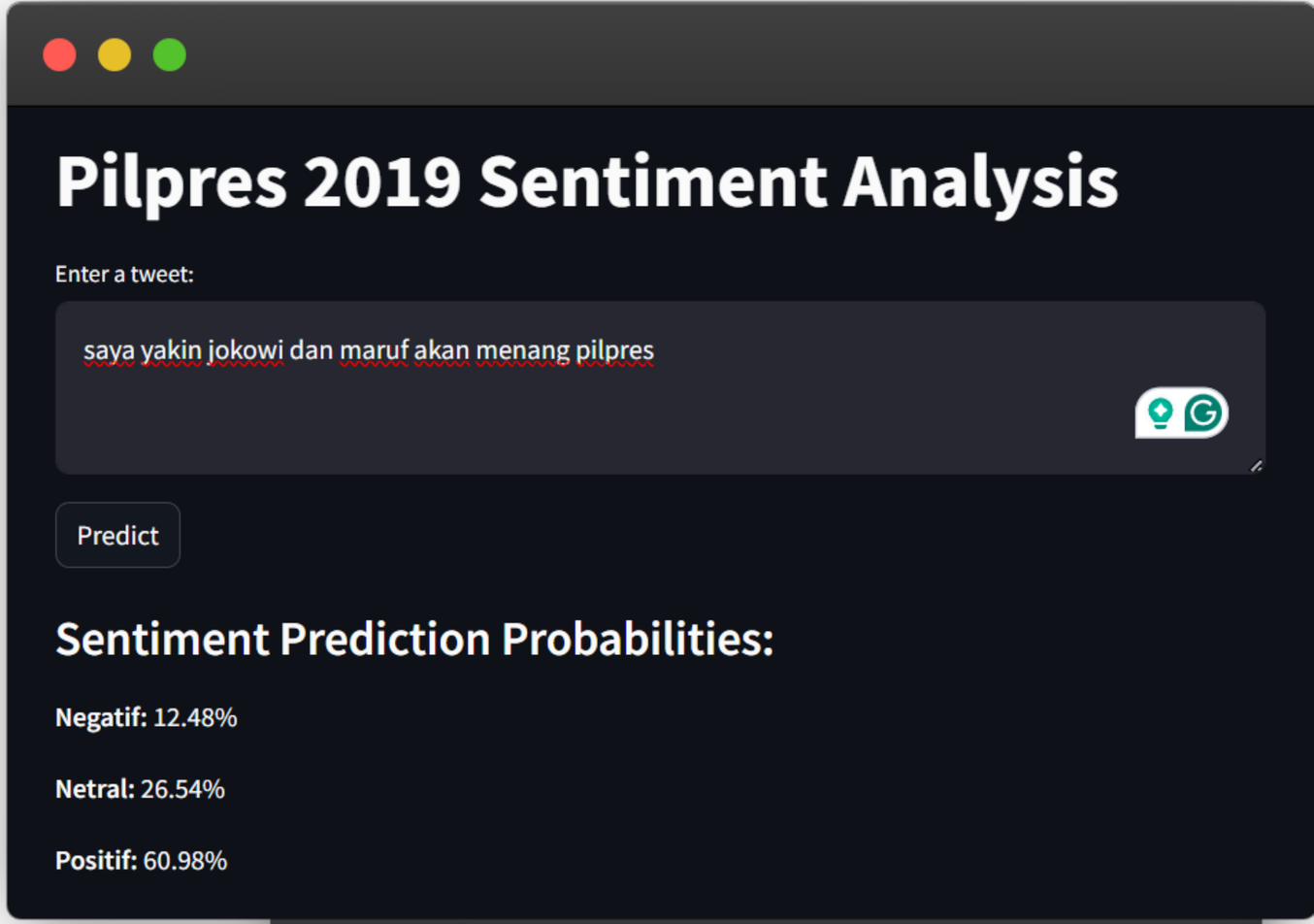
App Deployment

Proses deployment masih berada pada tahap **local deployment**. Berikut ini adalah screenshot dari prototype aplikasi

Back-end: Flask

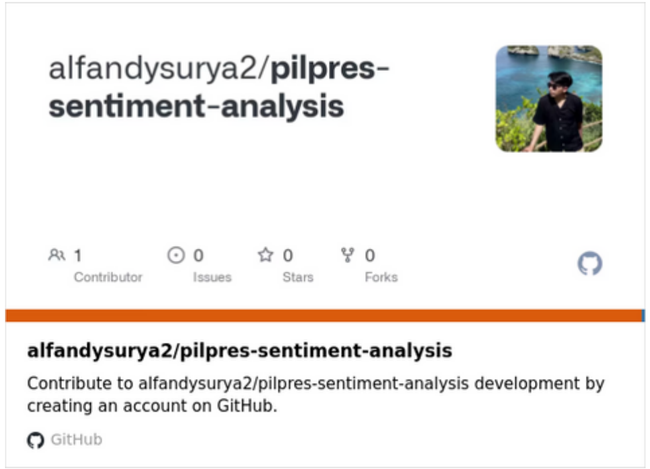
Front-end: Streamlit

Chart: plotly



Our Github Repository

Link:



Directory Structure:

```
├── dataset
│   └── tweet.csv
├── models
│   ├── model_rf_sg_tuned_sw.joblib
│   ├── word2vec_model_sg_min_8_window_6_sw.bin
│   └── [dan file model lainnya]
├── plugins
│   └── text_prep_function.py
├── resources
│   ├── emoji_dictionary.json
│   └── slang_word_dictionary.json
├── .gitignore
├── app_flask.py
├── app_streamlit.py
├── notebook_eda.ipynb
├── notebook_modelling.ipynb
└── README.md
```

Summary, Evaluation & Future Improvements

Summary:

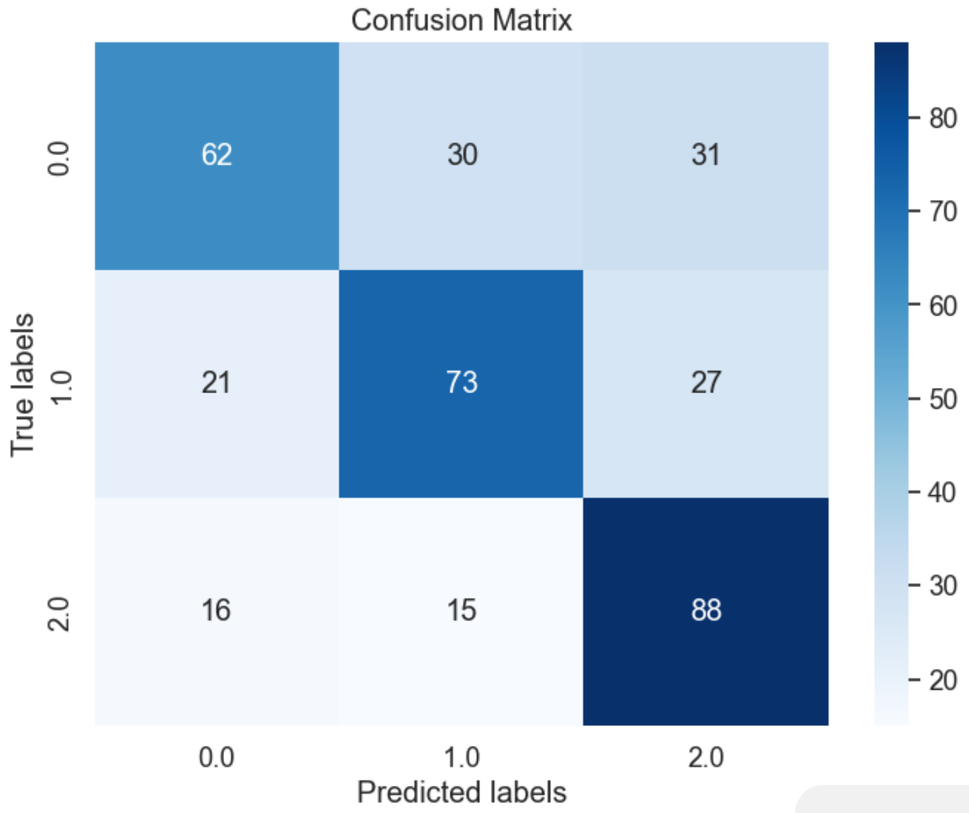
- Model Random Forest terbaik adalah **model RF + stopwords + Word2Vec SkipGram** dengan performa akurasi train 71%, validation 61% dan test 55%.
- Model LSTM terbaik adalah **model LSTM tanpa text processing** dengan performa akurasi train 98%, validation 58,6%, dan test 57,5%.

- Performa kedua model cukup mirip namun dibandingkan dengan RF, LSTM mengalami overfitting yang cukup parah karena selisih akurasi dengan validation dan test yang besar.
- Dari hasil error analysis, performa model belum cukup bagus dikarenakan:
  - Terdapat beberapa pola tweet dengan label yang tidak sesuai.
  - Belum dapat memahami sarkasme atau tweet yang tidak mengandung kata negatif
  - Adanya noise dari label netral yang memiliki kata negatif atau positif
- Terdapat juga kesalahan pemrosesan data khusunya di bagian singkatan atau slang words
- Penggunaan stopwords di kasus ini berpengaruh positif terhadap performa model

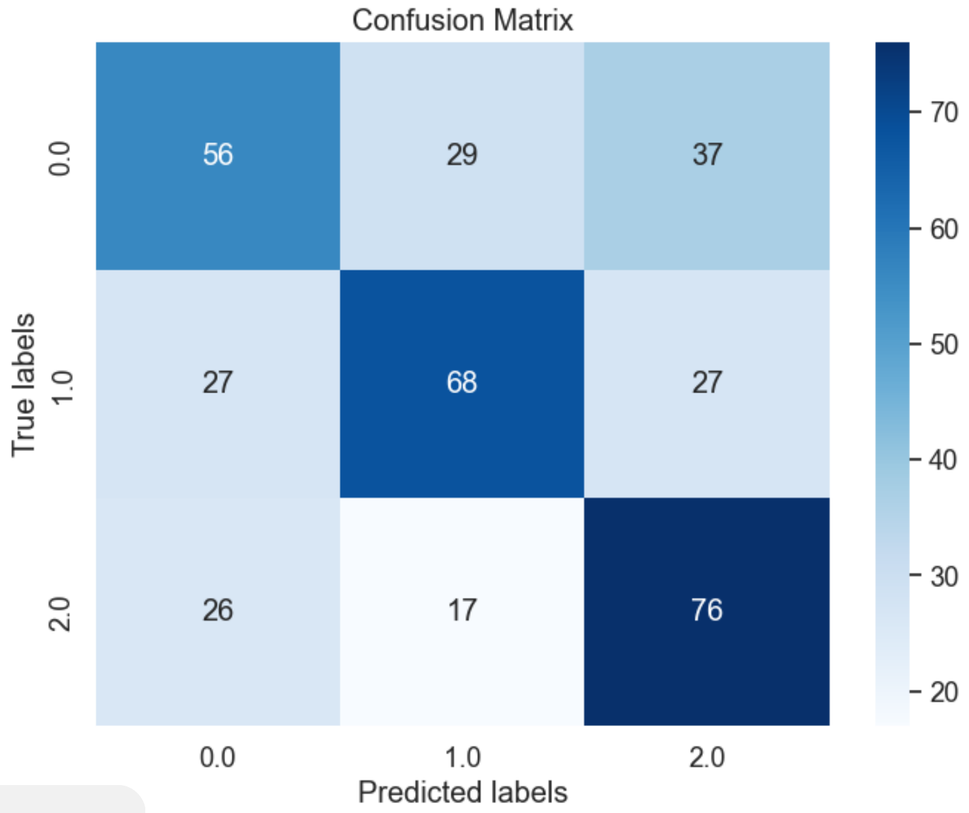
Summary, Evaluation & Future Improvements

Confusion Matrix Evaluation:

Validation Set



Test Set



Label

0 : Positif, 1 : Netral, 2: Negatif

Summary, Evaluation & Future Improvements

Future Improvements:

- Karena data yang dimiliki sedikit sehingga menyebabkan overfitting, selanjutnya bisa dilakukan penambahan data.
- Memperbaiki label yang salah sehingga model tidak bingung dalam melakukan klasifikasi.
- Memperbaiki preprocessing yang salah seperti pada *typo* dan *slang words*.
- Menggunakan embedding Word2Vec dan menambahkan Regularization untuk model LSTM
- Melakukan eksperimen dengan model lain selain dari Random Forest
- Memperbanyak data dengan label negatif dan positif agar dapat mengimprove performa model untuk sentimen tersebut
- Tidak menghapus Hashtag karena berisikan informasi penting terkait sentimen

Thank you, any question?