Segmentation of text with emojis

🥳🥳🥳🥳🥳🥳🥳🥳🥳🥳
🥳🥳

Presented by: Zakharov Artem, Shatalov Andrey and Keseli Timur, БПАД212

Project supervisor: Titova Natalia Nikolaevna

RACI Matrix

TASK	Zakharov Artem😎	Shatalov Andrew🤩	Keseli Timur🤡
Data Preprocessing	R	R	R
Data Analysis	R	R	R
KNN	I	I	R
BERTweet	I	R	I
MLP	I	R	I
Siamese	R	I	I
Presentation	R	R	R

Project aim

Combine text and emoji analysis in text segmentation. Our dataset requires to classify tweets based on text and emoji into 4 classes:

0: sad 😿

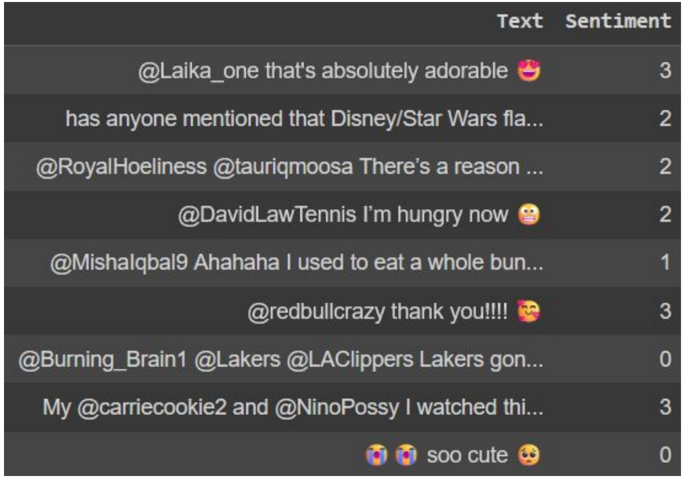
1: happy 🤣

2: angry 🤬

3: love 🥰

Dataset

This dataset contains 3085 Twitter tweets labelled with 0–3 values, where 0 = sad, 1 = happy, 2 = angry and 3 = love



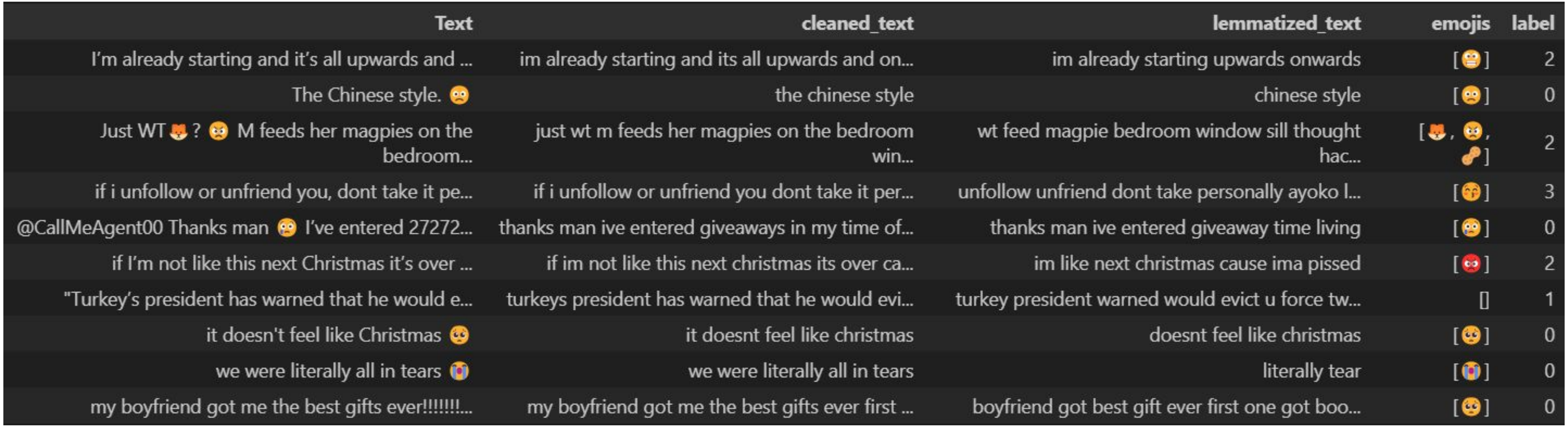
Text	Sentiment
@Laika_one that's absolutely adorable 😍	3
has anyone mentioned that Disney/Star Wars fla...	2
@RoyalHoeliness @tauriqmoosa There’s a reason ...	2
@DavidLawTennis I’m hungry now 😬	2
@MishalqbaI9 Ahahaha I used to eat a whole bun...	1
@redbullcrazy thank you!!!! 🥰	3
@Burning_Brain1 @Lakers @LAClippers Lakers gon...	0
My @carriecookie2 and @NinoPossy I watched thi...	3
😭 😭 soo cute 🥺	0

Figure 1. Data Samples.

Data preprocessing

1. Cleaning sentences
2. Deleting stop words
3. Lemmatization
4. Extracting emojis
5. Vectorisation of words / emojis
6. Concatenation of vectors (dimension = 600)

Data preprocessing



Text	cleaned_text	lemmatized_text	emojis	label
I'm already starting and it's all upwards and ...	im already starting and its all upwards and on...	im already starting upwards onwards	[😬]	2
The Chinese style. 🙁	the chinese style	chinese style	[🙁]	0
Just WT🦊? 😠 M feeds her magpies on the bedroom...	just wt m feeds her magpies on the bedroom win...	wt feed magpie bedroom window sill thought hac...	[🦊, 😠, 🥜]	2
if i unfollow or unfriend you, dont take it pe...	if i unfollow or unfriend you dont take it per...	unfollow unfriend dont take personally ayoko l...	[😚]	3
@CallMeAgent00 Thanks man 😢 I've entered 27272...	thanks man ive entered giveaways in my time of...	thanks man ive entered giveaway time living	[😢]	0
if I'm not like this next Christmas it's over ...	if im not like this next christmas its over ca...	im like next christmas cause ima pissed	[😡]	2
"Turkey's president has warned that he would e...	turkeys president has warned that he would evi...	turkey president warned would evict u force tw...	[]	1
it doesn't feel like Christmas 🥺	it doesnt feel like christmas	doesnt feel like christmas	[🥺]	0
we were literally all in tears 😭	we were literally all in tears	literally tear	[😭]	0
my boyfriend got me the best gifts ever!!!!!!!...	my boyfriend got me the best gifts ever first ...	boyfriend got best gift ever first one got boo...	[🥺]	0

Figure 2. Data before and after Preprocessing.

Data preprocessing, Word2Vec

Word2Vec Pre-trained model:

- Google News: pre-trained vectors trained on part of Google News dataset (about 100 billion words).
- The model contains 300-dimensional vectors for 3 million words and phrases.

Emoji2Vec Pre-trained model:

- Proved to have better accuracy with Google News Word2Vec on analysing emoji texts and encodes 300 dimensional vectors.
- Contains description of 1661 emoji symbols.

Data analysis

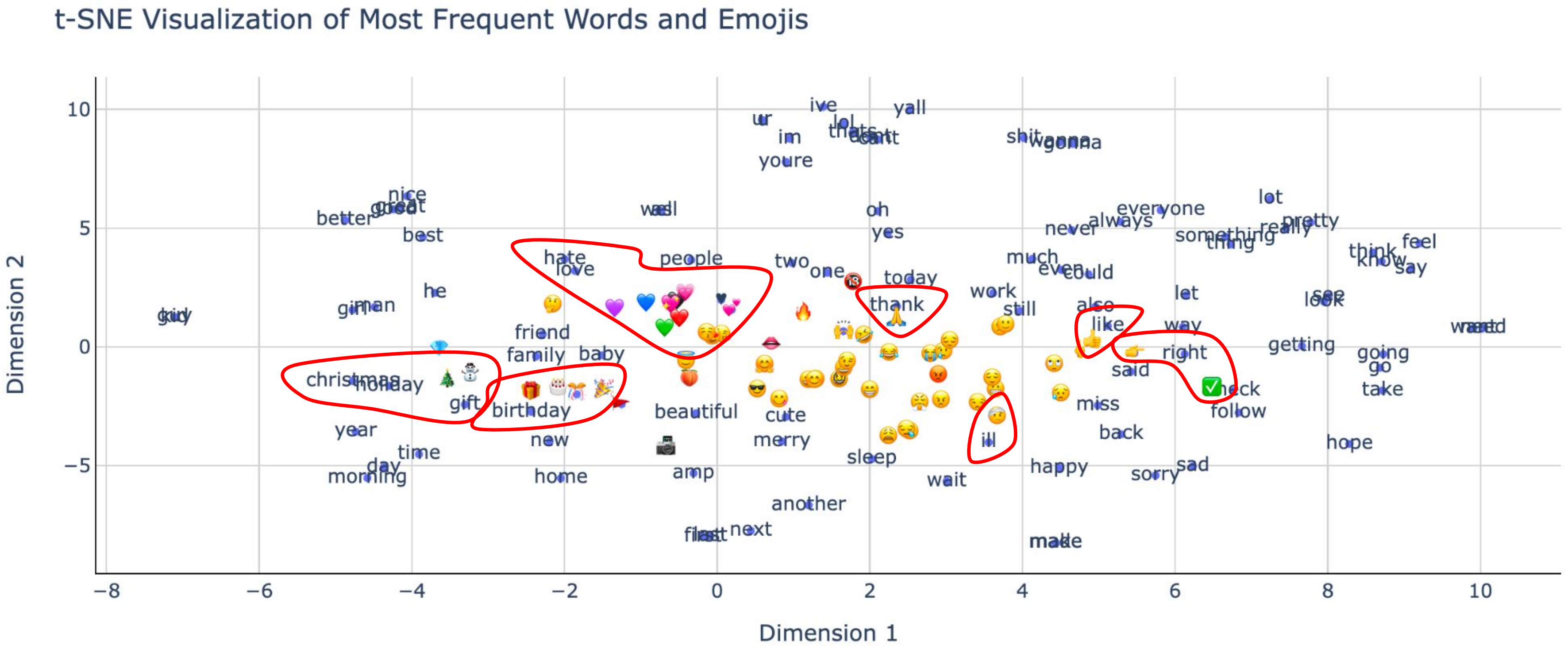


Figure 3. Stochastic neighbour embedding.

Data analysis

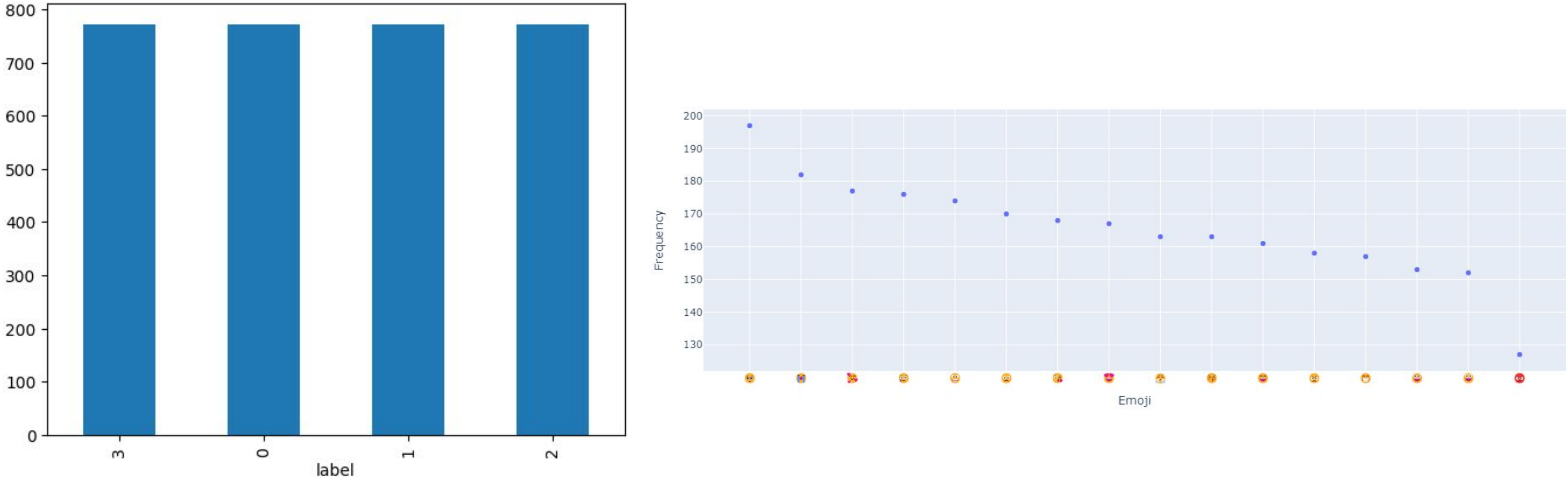


Figure 4 & 5. Class and Emoji Distribution.

Models and metrics

1. KNN
2. Random Forests on BERT Embeddings
3. SIAMESE
4. SimpleNN

1. Accuracy
2. Precision
3. Recall
4. F1 Score

KNN

Implements K-Nearest Neighbors for classification.

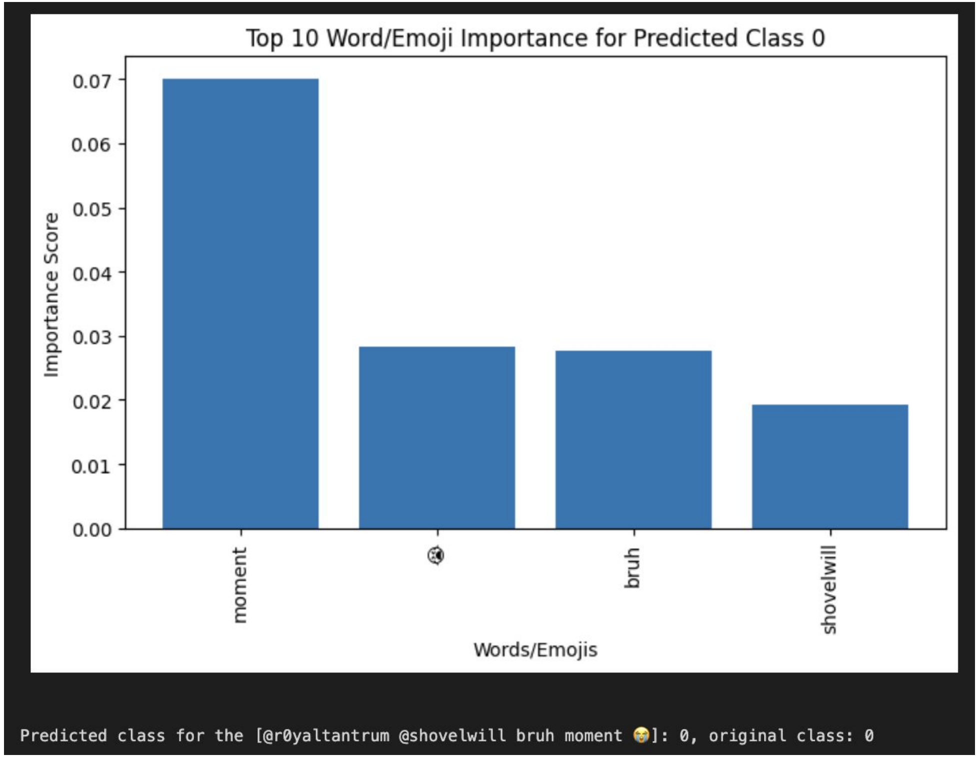
- Uses 4 neighbors with Euclidean distance as the metric.
- Trained on normalized input data to classify sentiment.

Siamese

Siamese Neural Network with triplet loss for classification.

- Normalizes input data to improve model performance and convergence.
- Input data is reshaped to add a time dimension for convolutional processing.
- Consists of two shared convolutional layers for feature extraction.
- Employs a fully connected dense layer for generating embeddings.
- Utilizes cosine similarity on sNN embeddings for classification.

Word Importance for SiameseNN



```
Predicted class for the [@r0yaltantrum @shovelwill bruh moment 😭]: 0, original class: 0
```

- Compute mean embeddings
- Calculate the gradient of the similarity score with respect to the input features (words and emojis).
- The absolute values of these gradients indicate how much each feature (word/emoji) contributes to the similarity score.

BERTweet + Random forest

- Performs data augmentation on text data using synonym replacement.
- Extracts emojis from text using regular expressions.
- Combines text and emoji embeddings using BERTweet.
- Trains a Random Forest model to classify sentiment.
    -

MLP

Utilizes a Multi-Layer Perceptron (MLP) with multiple dense layers.

- Four hidden layers with ReLU activation functions.
- Dropout layers for regularization to prevent overfitting.
- Output layer uses softmax activation for multi-class classification.
- Trained using the Adam optimizer.

Results

	Accuracy	Precision	Recall	F1 Score
KNN	0.7277	0.7286	0.7277	0.7280
BERTweet + RF	0.7615	0.7632	0.7615	0.7611
MLP	0.7763	0.7818	0.7763	0.7771
Siamese	0.7812	0.8086	0.7812	0.7855

References

1. Dataset, Kaggle, https://www.kaggle.com/datasets/juyana054/sentiment-data-16-emoji
2. Code, Google Colaboratory,
3. Emoji2Vec, GitHub, https://github.com/uclnlp/emoji2vec
4. Google News vector model, GitHub, https://github.com/mmihaltz/word2vec-GoogleNews