Why is emoji prediction difficult?

TAR project

Matija Bertović, Antun Magdić, Ante Žužul

University of Zagreb
Faculty of Electrical Engineering and Computing

June 1, 2020

This song is lit!

This song is lit! 🔥

This song is lit! 🔥

And to all, a Merry Christmas!

This song is lit! 🔥

And to all, a Merry Christmas! 🎄

This song is lit! 🔥

And to all, a Merry Christmas! 🎄

I need new friends...

This song is lit! 🔥

And to all, a Merry Christmas! 🎄

I need new friends... 😔

This song is lit! 🔥

And to all, a Merry Christmas! 🎄

I need new friends... 😔 😂

This song is lit! 🔥

And to all, a Merry Christmas! 🎄

I need new friends... 😔 😂

I'm so happy to have you

This song is lit! 🔥

And to all, a Merry Christmas! 🎄

I need new friends... 😔 😂

I'm so happy to have you ❤️ 💕 💙 💜 😍

Dataset

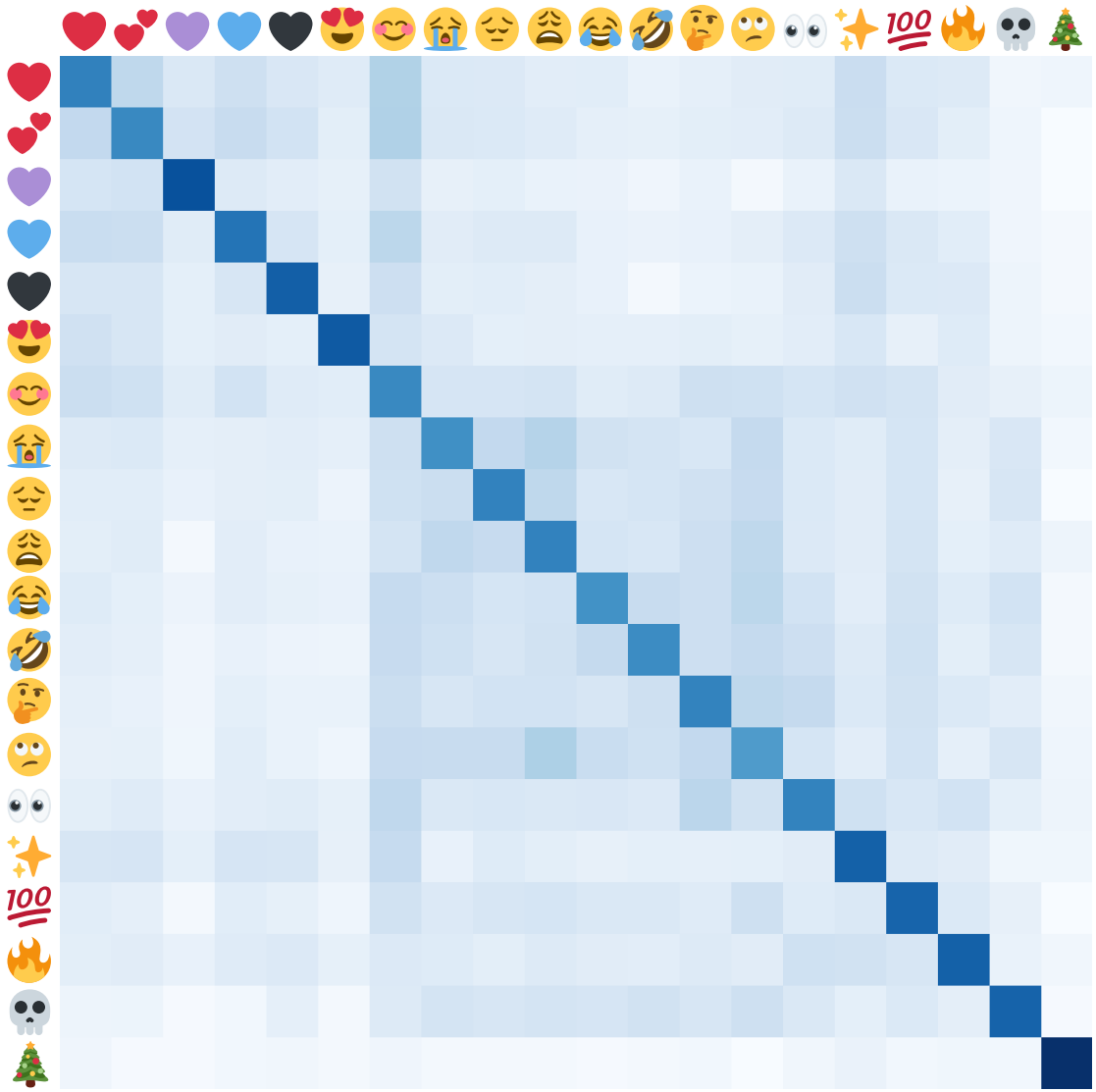
- 10 million tweets collected
- Only tweets with single emoji kept
- Final data: 200 000 tweets ($20 \times 10\,000$)
- Train: 120 000, validation: 40 000, test: 40 000 (all balanced)
- Tweet text $\mapsto$ emoji

Experiment 1

- We compare various models:
  - Naïve Bayes (NB)
  - Logistic regression (LR)
  - Feed forward neural network (NN)
  - Bidirectional LSTM (BLSTM)
- GloVe vs TF-IDF
- Word order
- Naïve assumption

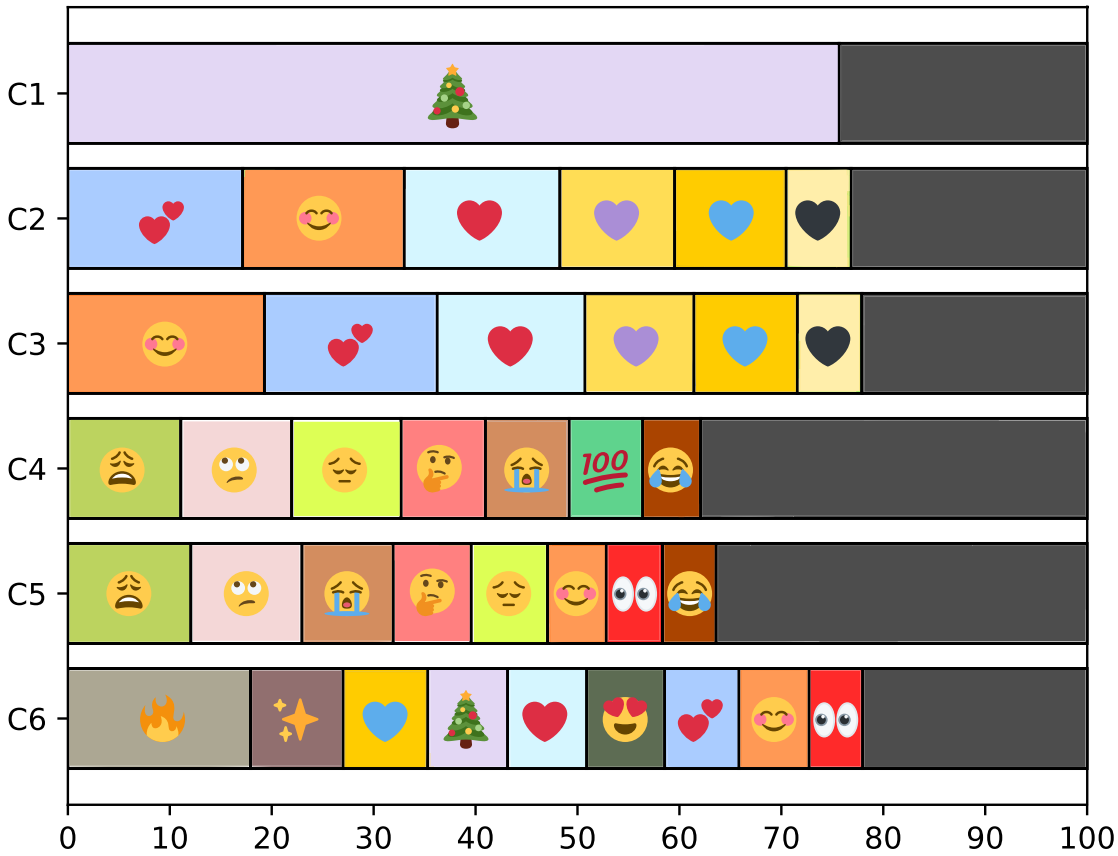
Model	Accuracy (%)
NB	51.15
LR GloVe	33.78
LR TF-IDF	53.35
NN GloVe	45.67
NN TF-IDF	51.05
BLSTM	51.40

Experiment 1



Experiment 2

- K-Means with GloVe (50 clusters)



Conclusion

- Main difficulties:
  - Synonymy among emojis
  - Subjective meanings
  - Sarcasm
- More information is needed for better performance