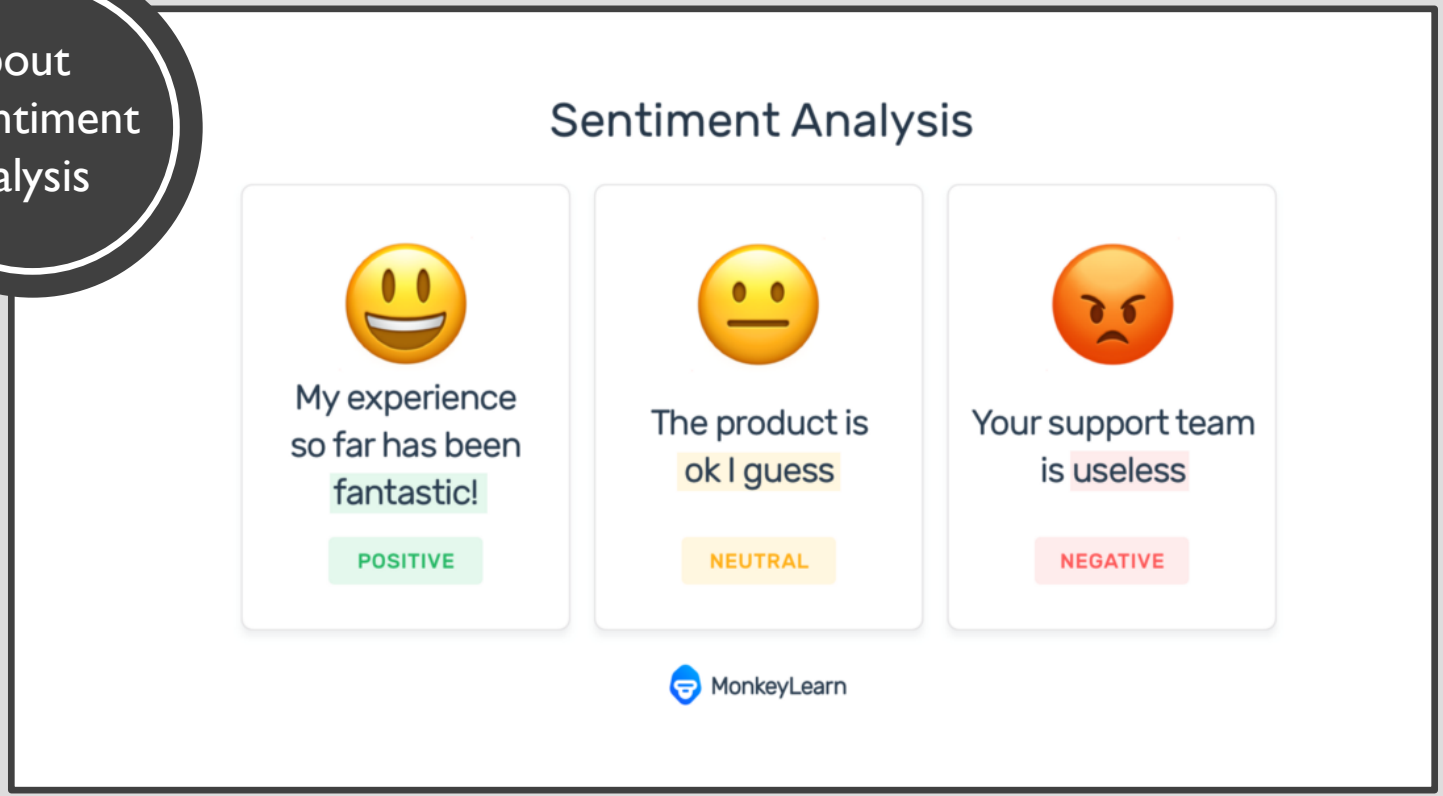
SENTIMENT ANALYSIS

Web and text analysis

Ophélie de la Brassinne Bonardeaux
Lucie Navez

About sentiment analysis



DATASET

Twitter dataset

- 1.6 millions tweets
- Extensive use of words that don't exist (ex:"*looooovvve*")
- Example:
  - *"Damn, back to school tomorrow"*
  - *"this week is not going as i had hoped"*

IMBD

- 50.000 movie reviews
- Example:
  - *"In one of her first movies, Romy Schneider shines as young queen Victoria of Britain, as she is suddenly put into the throne at the age of 18, learns to govern despite the machinations of the politicians, and eventually romances and marries Prince Albert of Saxony. Kitschy and campy (though surprisingly faithful to the real events), this romantic piece is irresistible. Seeing this movie about British royals spoken in German adds to its quaint charm. On that front, one wonders why an Austrian movie was made about an English queen"*

LIBRAIRIES

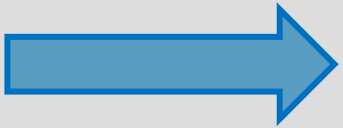
NLKT

Pytorch

Gensim

DATA PROCESSING

1. Put everything to lowercase
2. Remove unwanted characters/words like HTML beacons, emojis, handles, …
3. Remove stop-words
4. Lemmatization
5. Stemming
6. Remove non-English words
7. Expand English contractions (he'll, can't, ...)
8. Remove extensive repetitions of words (e.g.: "*fuck, fuck, fuck, fuck, ...*" or "*shit, shit, shit, shit, ...*")



Followed with padding and encoding

VOCABULARY

- All words in the pre-processed train set
- <PAD>
- <SOS>
- <EOS>
- <UKN>

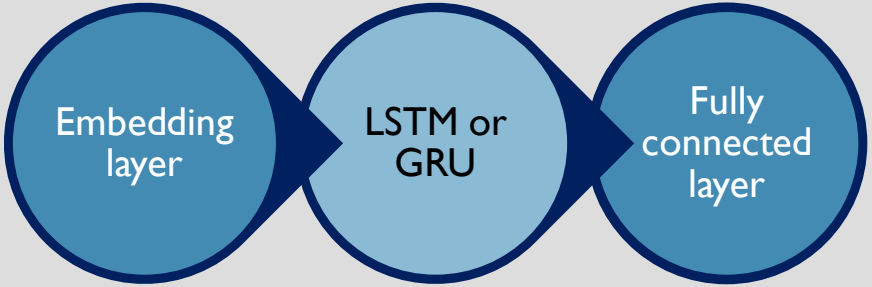
	Twitter	IMBD
Vocabulary size	30.982	30.619
Max sentence length	20	692
Sequence length	12	100

WORD AND DOCUMENT EMBEDDING

- Pre-trained GloVe
- Pre-trained Word2Vec
- Pre-trained FastText
- Word2Vec trained on our data

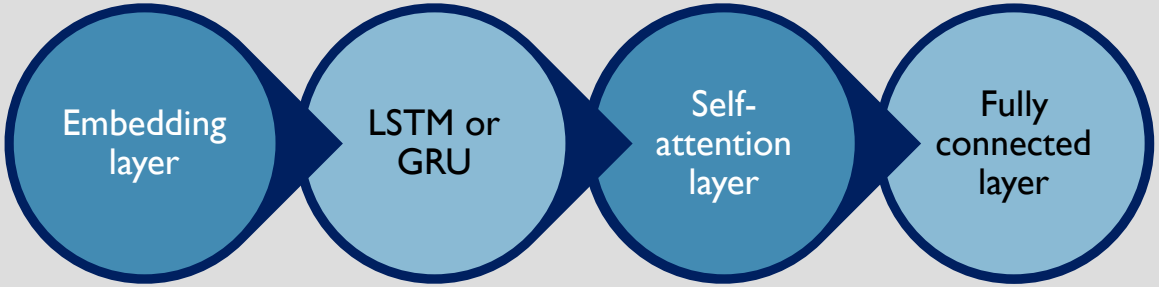
- Doc2Vec
- Word2Vec averaged with TF-idf weights

RNN



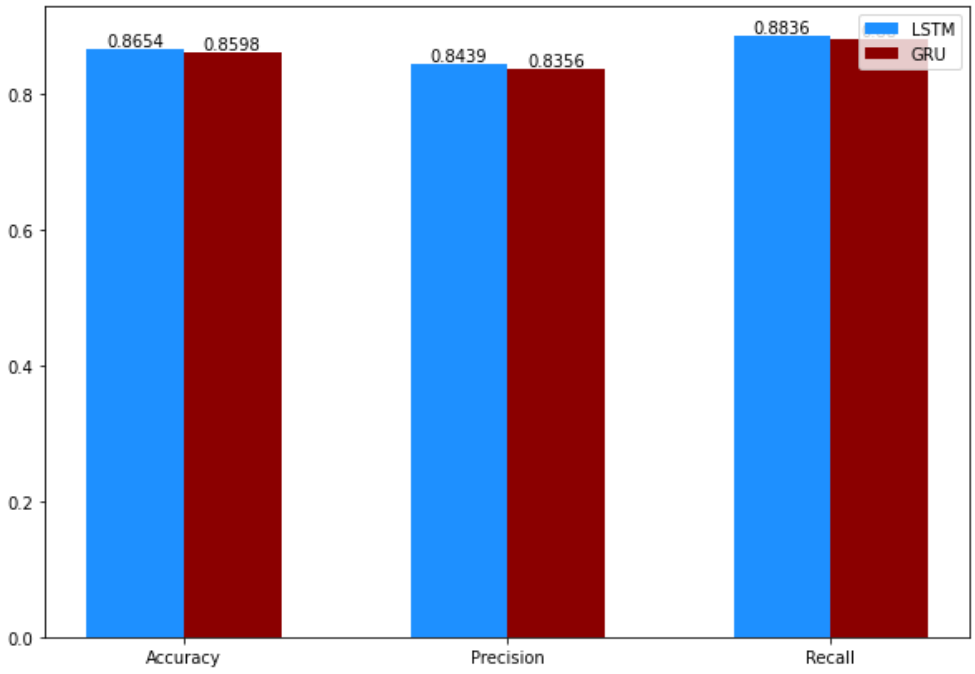
- Embedding dimensions: 300
- Hidden layer dimension: 32
- Dropout: 20%
- Bidirectional
- Fully connected layer
- Dataset: IMBD

RNN + ATTENTION

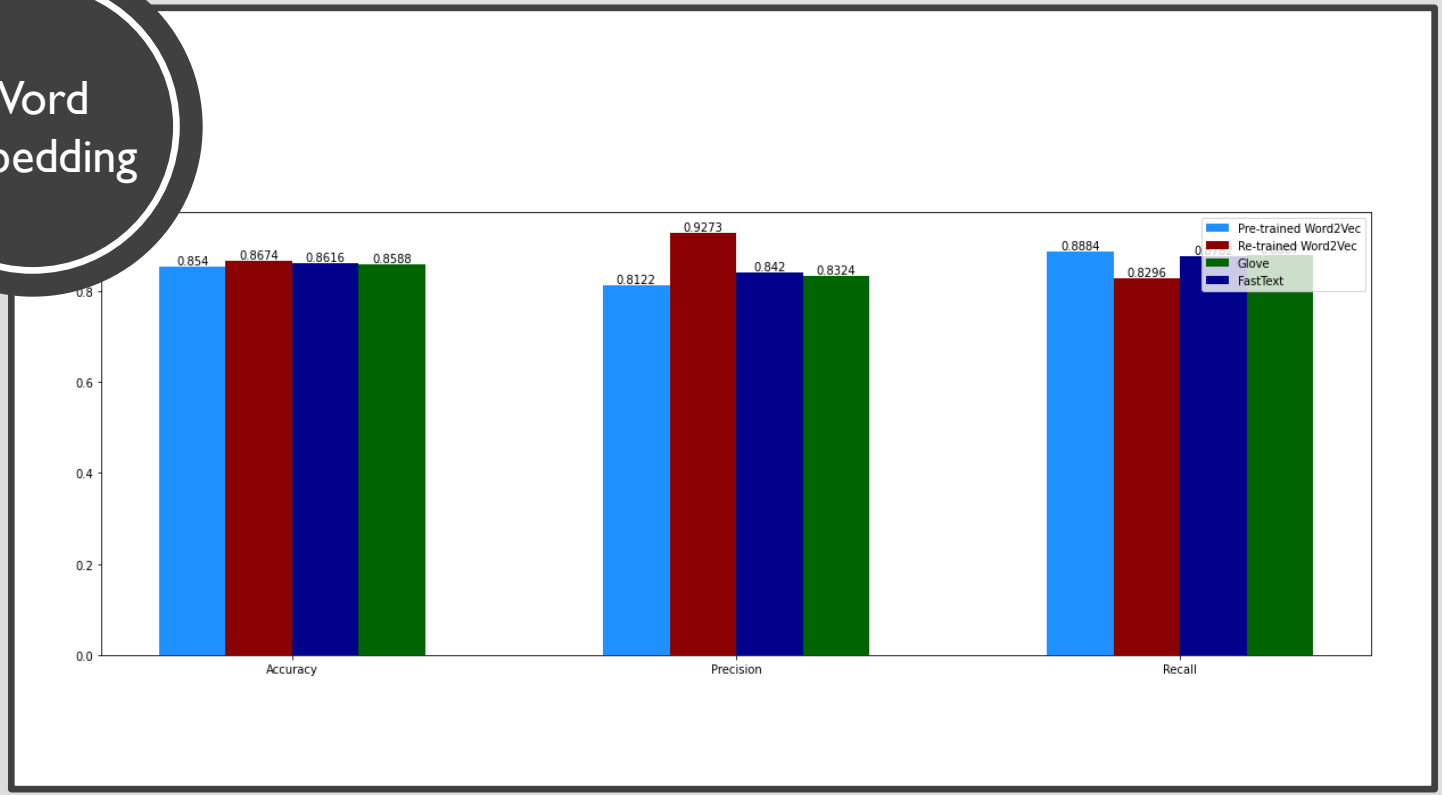


- Embedding dimensions: 300
- Hidden layer dimension: 32
- Dropout: 20%
- Bidirectional
- Fully connected layer
- Dataset: IMBD

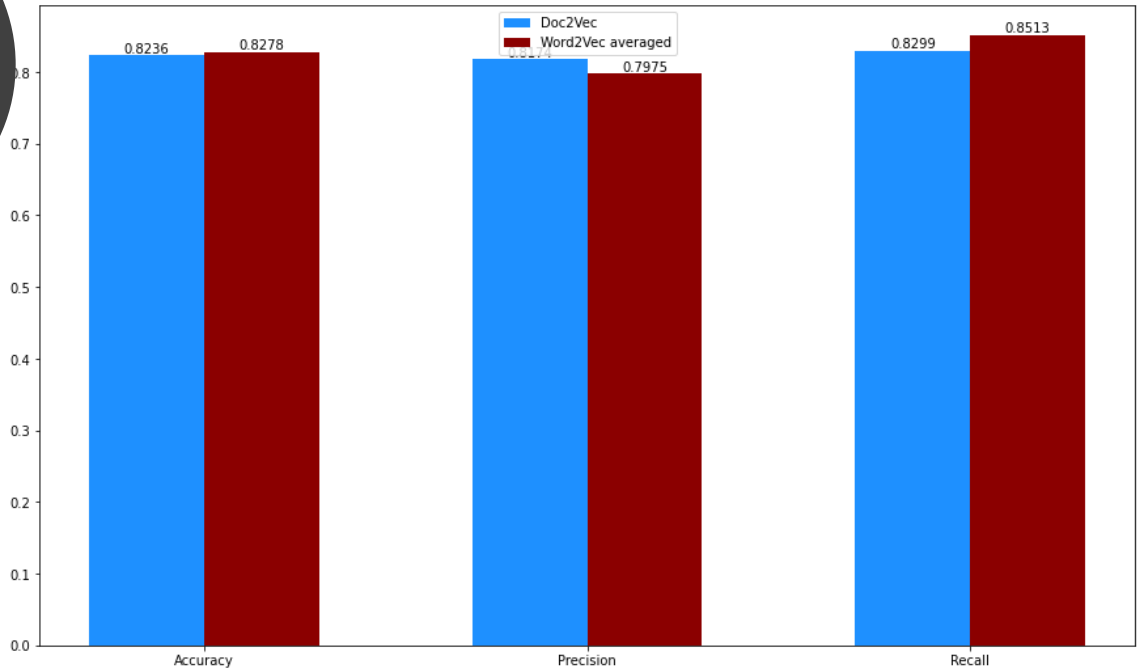
GRU VS LSTM


0.8654
0.8598
0.8439
0.8356
0.8836
LSTM
GRU
0.0
0.2
0.4
0.6
0.8
Accuracy
Precision
Recall

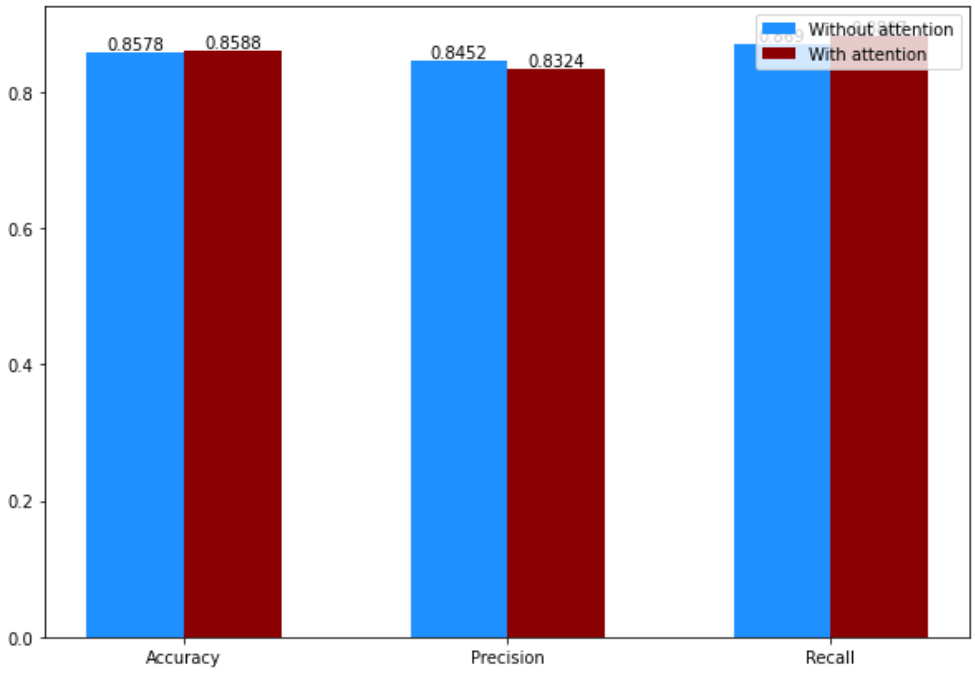
Word embedding



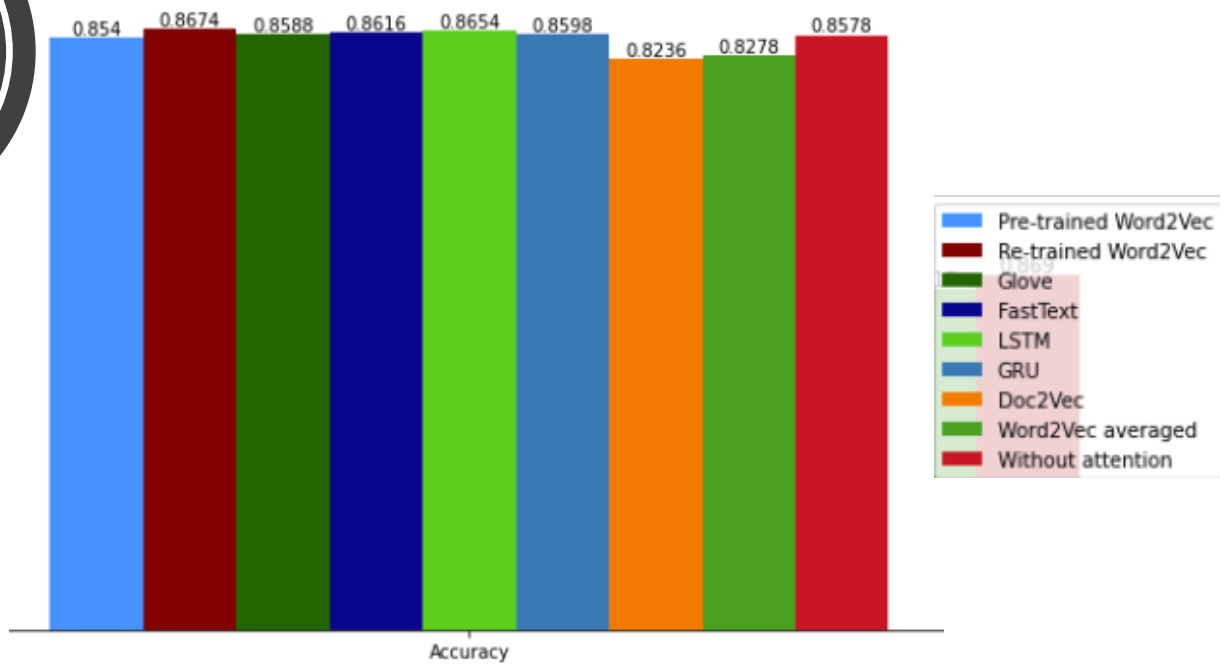
Document embedding



Attention


0.8578
0.8588
0.8452
0.8324
Without attention
With attention
0.0
0.2
0.4
0.6
0.8
Accuracy
Precision
Recall

Accuracy


0.854
0.8674
0.8588
0.8616
0.8654
0.8598
0.8236
0.8278
0.8578
Accuracy
Pre-trained Word2Vec
Re-trained Word2Vec
Glove
FastText
LSTM
GRU
Doc2Vec
Word2Vec averaged
Without attention