IE 684 Web Mining Project

Music Lyrics Sentiment Analysis

Martin Böckling, Chih-Yen Ou, Yi-Hsuan Peng, David Probst, Fabio Westphal

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

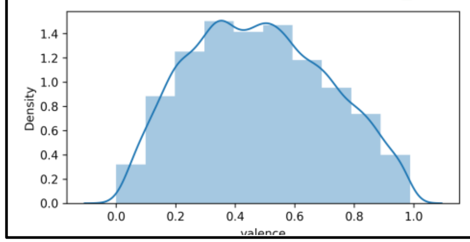
Use Case

- Music has a big impact on us
- Every song conveys a certain sentiment
  - through melody & sound
  - **through lyrics**
- Task: predict sentiment of a song just by its lyrics
- Benefit: provide better recommendations based on mood

Dataset Preparation

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Spotify dataset from Kaggle:
Song_titles
Artist
Year 1921-2021
Valence (0-1.0)



Genius API by Genius Media Group.
Python package: Lyricsgenius
song = Genius_api.search_song(title,artist)

Matched by: Song_title, artist

All_songs_dataset: Lyrics with space between lines

All_songs_comma: Lyrics with comma between lines

Data Cleaning

Remove songs with no lyrics matching

Clear out wrong lyrics

Remove duplicate songs

Filter out non-English songs

Normalize 175 Slang words

Preprocessing

Remove bracket [Intro],[Chorus],[Verse]

Convert all text into lowercase

Replace bracket [Intro],[Chorus],[Verse] with dot '.'

Keep punctuation

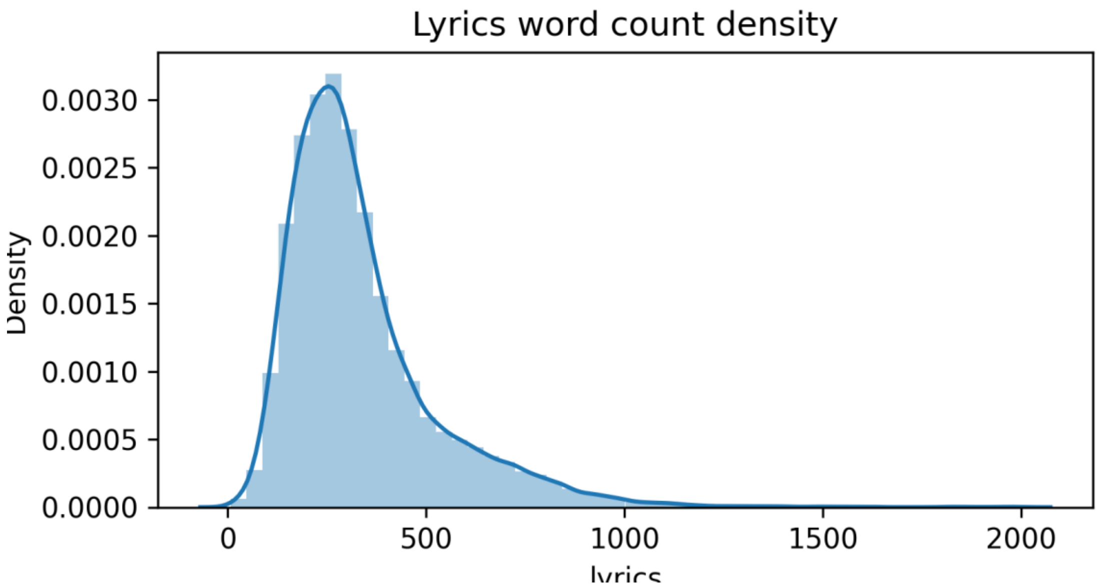
see it,She'll make the first move, smooth, int...
(down, down).Baby, whatever you want, girl, yo...
That lipstick on your collar, well, it ain't m...

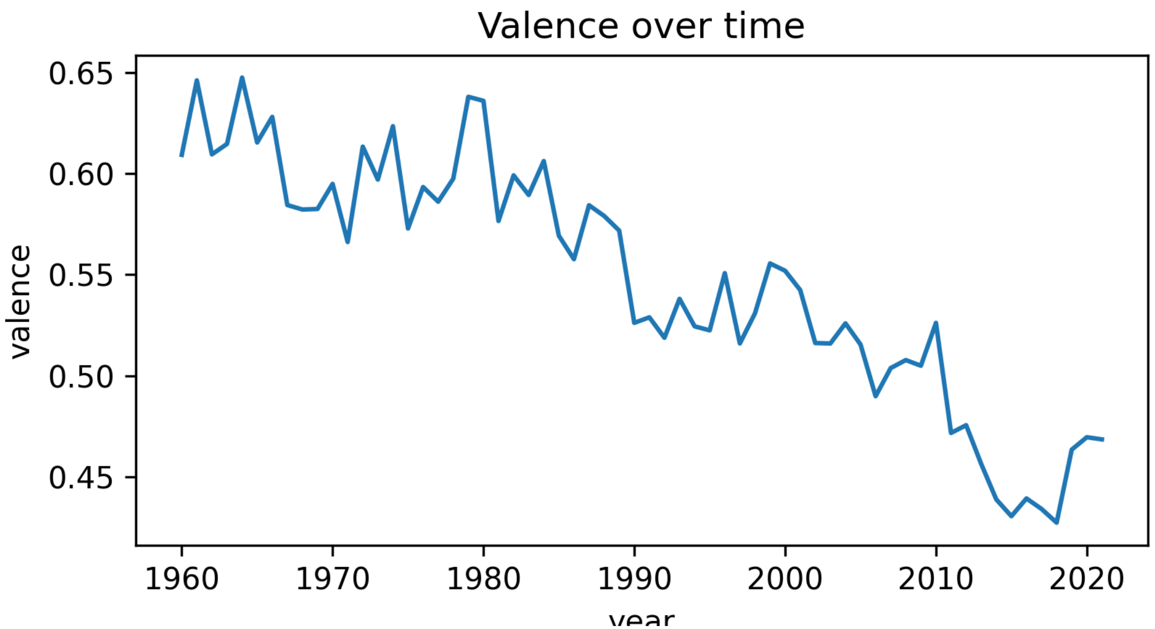
For Classic Classifiers

All_songs_corpus:
Use Spacy to remove stopword and lemmatize the text

2 classes	-	6207	Baseline: 54%
	+	5248	
3 classes	-	4558	Baseline: 40%
	o	3325	
	+	3586	

Data Understanding (1)





Data Understanding (2) - Keyword Extraction

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- First baseline approach using TF-IDF

"I miss the old Kanye, straight from the Go Kanye
Chop up the soul Kanye, set on his goals Kanye
I hate the new Kanye, the bad mood Kanye
The always rude Kanye, spaz in the news Kanye
I miss the sweet Kanye, chop up the beats Kanye..."

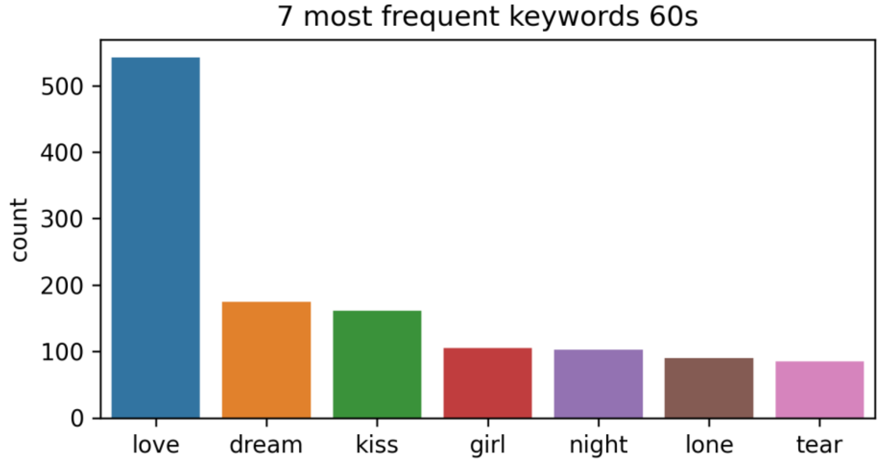
I love Kanye – Kanye West

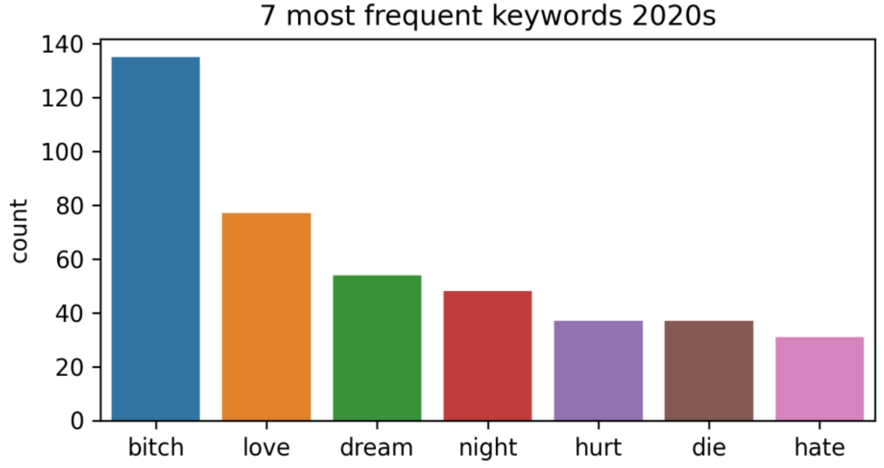
```
seven 1.0
blah 0.963
rumour 0.997
doo 0.964
gon 0.994
warn 0.994
bye 0.959
rape 0.991
fix 0.989
ooh 0.973
halo 0.989
meet 0.988
ruby 0.988
amazing 0.988
fire 0.987
infinity 0.984
scotty 0.985
thuggish 0.985
para 0.985
aight 0.984
donna 0.984
fah 0.983
molly 0.983
bah 0.982
anchor 0.982
kanye 0.982
radar 0.981
bam 0.981
sing 0.98
womanizer 0.98
cabron 0.98
hosanna 0.978
funkytown 0.977
fertilizer 0.976
superhero 0.976
```

Data Understanding (3) - Keyword Extraction

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Second approach using KeyBERT
  - Easy to use
  - Pretrained
  - Calculating cosine similarity between word and document embeddings created by BERT





Algorithms

Model	Model Setting
SVM	TF-IDF, Default setting with 5-fold cross validation (CV)
*Random Forest	TF-IDF, Grid search CV
BERT	Pretrained model: bert_based_uncased Max_len = 256 Batch size = 32
XLNet	Pretrained model: xlnet_based_cased Max_len = 256 Batch size = 16

*SVM+Random Forest+Stochastic Gradient Descent+Multinomial Naïve Bayes+Xgboost

BERT vs. XLNet

Similarity:

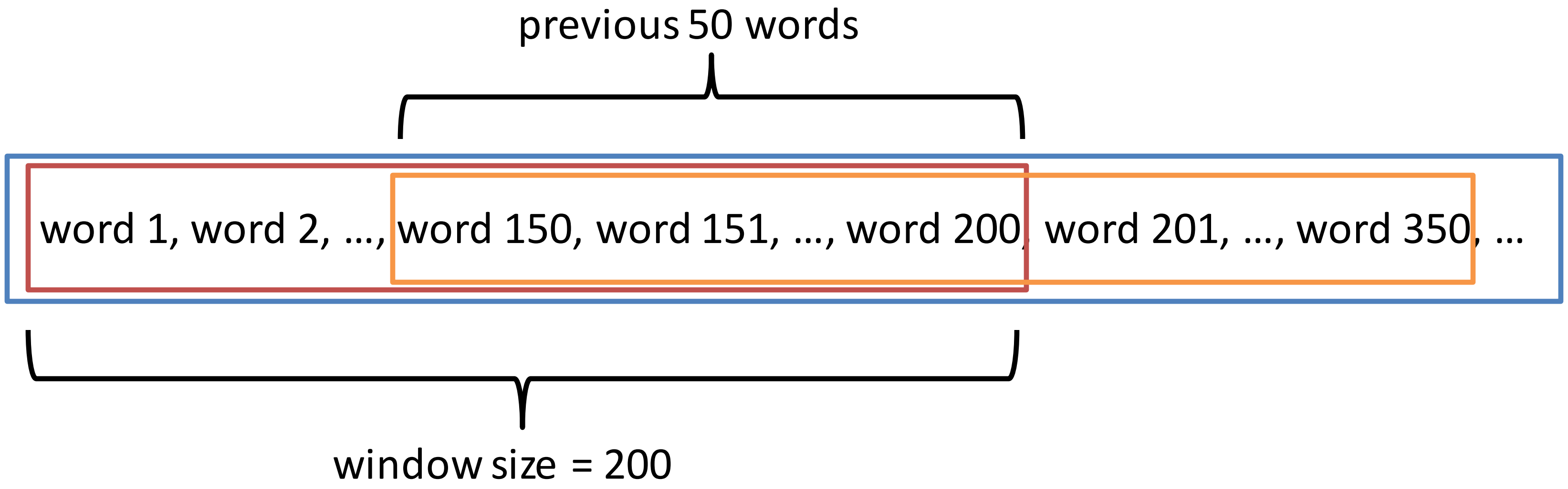
1. Pretraining and finetuning on bidirectional language understanding
2. Contains Transformer encoder

Differences on XLNet:

1. No pretrain-finetune discrepancy problem
    - Permutation Language Modeling (PLM)
    - Remove [MASK] token
2. Long text understanding
    - TransformerXL
    - Accept max_len > 512
3. Better performance
    - Language understanding, reading comprehension, text classification tasks

Sliding Window

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics



Model results for two class sentiment

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Model	Dataset	Sliding Window	Accuracy
Baseline	All songs dataset	False	0.5419
SVM	All songs corpus	False	0.61
Random Forest		False	**0.63**
BERT	All songs dataset	False	0.6308
		True	**0.7419**
	All songs comma dataset	False	0.6393
		True	**0.7307**
XLNet	All songs dataset	False	0.625
		True	**0.7283**
	All songs comma dataset	False	0.6027
		True	**0.6997**

Model results for three class sentiment

Model	Dataset	Sliding Window	Accuracy
Baseline	All songs dataset	False	0.3974
SVM	All songs corpus	False	0.44
Random Forest		False	**0.5**
BERT	All songs dataset	False	0.4717
		True	**0.5908**
	All songs comma dataset	False	0.4665
		True	**0.5567**
XLNet	All songs dataset	False	0.4339
		True	**0.593**
	All songs comma dataset	False	0.4435
		True	**0.5539**

Conclusion & Outlook

- BERT worked best for lyrics sentiment classification
- Feature extraction with sliding window helpful
- Biggest challenge: long text with meaningless repetitive words
- Next step: extend classes to more sentiments
  - angry, happy, relaxed, sad...
- Consider acoustic data for holistic view
- Look into decline in positivity

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Thank you!

References can be found in the report.