
Obiettivo
Preprocessing
Text representation and Modeling
Conclusioni
Text Mining and Search
Amazon reviews classification
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
amazon
Ivan Mera 783086

OBIETTIVO

Generale: Creare una rete neurale per classificare testi

Specifico: Confrontare due tecniche di text representation


Dataset
From
JSON
To
CSV
Quattro macro categorie:
• Industrial
• Luxury
• Musical Instruments
• Appliances
GloVe
EDA & Feature Engineering
Text Preprocessing

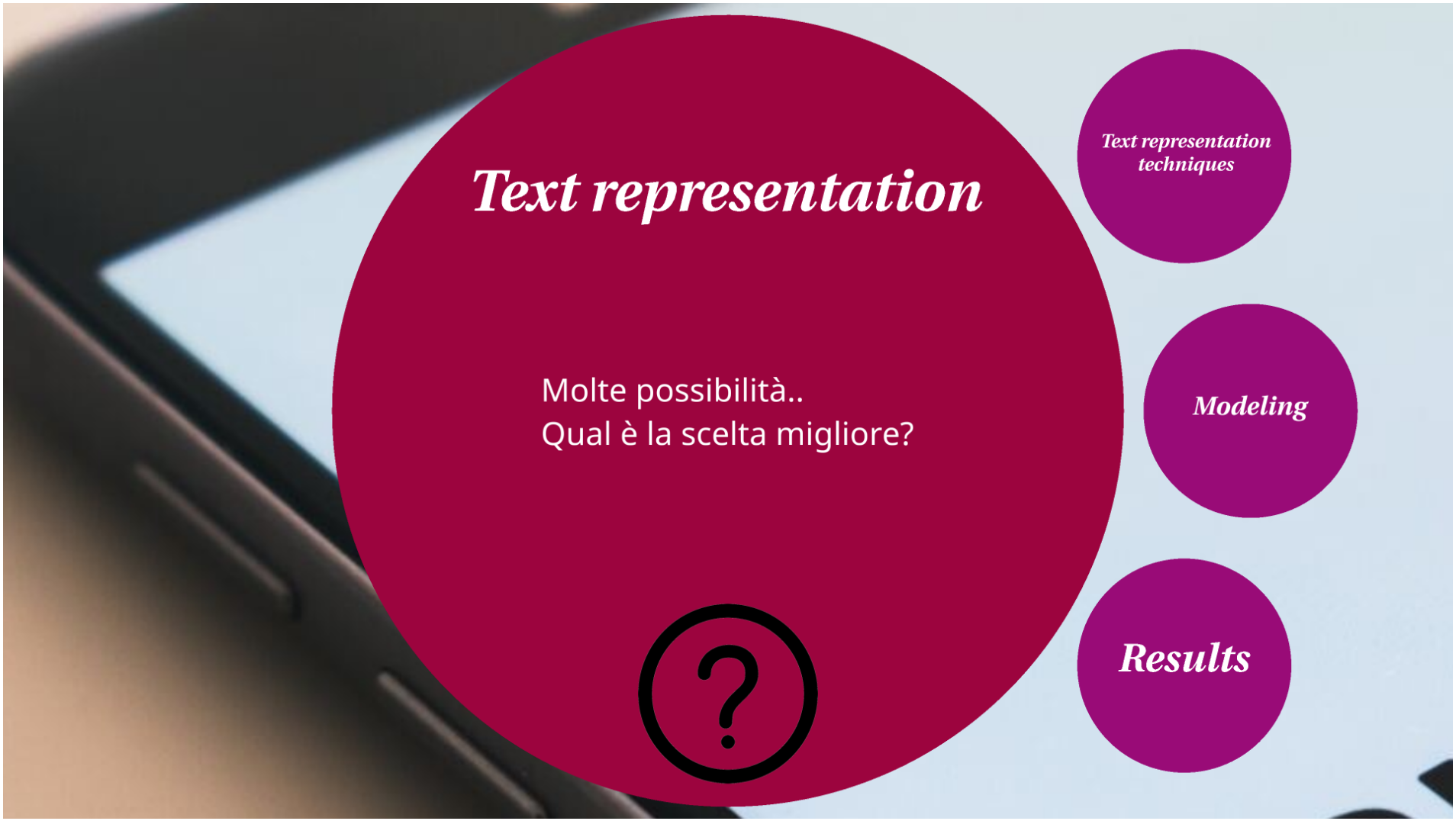
Operazioni

- Controllo missing values
- Numero di recensioni per prodotto
- Variabile target


Dataset
From
JSON
To
CSV
Quattro macro categorie:
• Industrial
• Luxury
• Musical Instruments
• Appliances
GloVe
EDA & Feature Engineering
Text Preprocessing

Operazioni

- Upper case
- Link web
- Contrazioni
- Emoji
- Caratteri speciali
- Spazi bianchi
- Stop words
- Lemmatization


Text representation
Molte possibilità..
Qual è la scelta migliore?
Text representation techniques
Modeling
Results

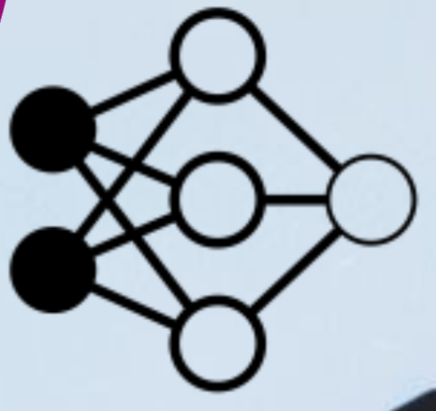
Approches

- TF-IDF
- WORD EMBEDDINGS

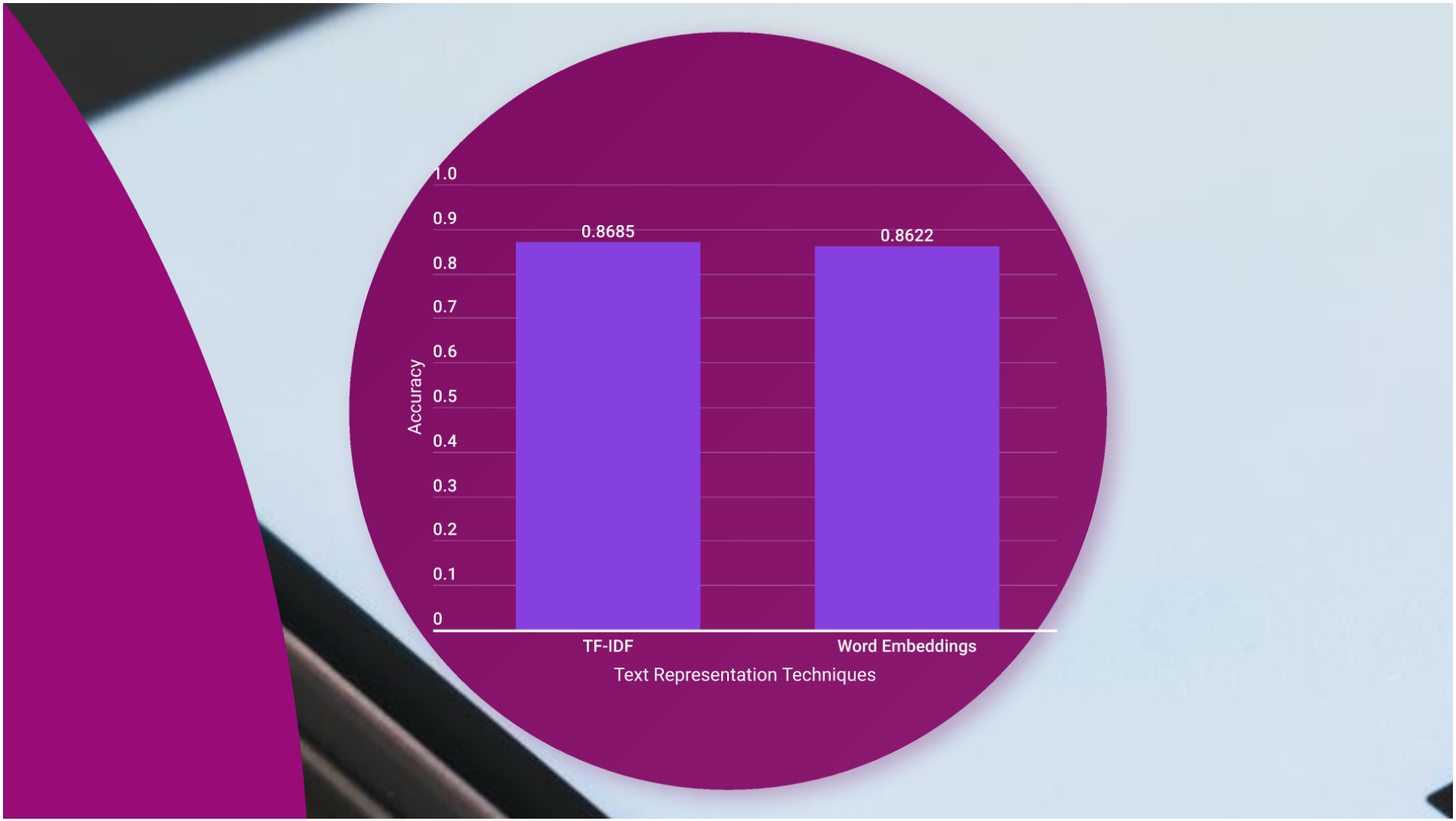
Neural networks architectures

- Dense layers
- CNN + GRU

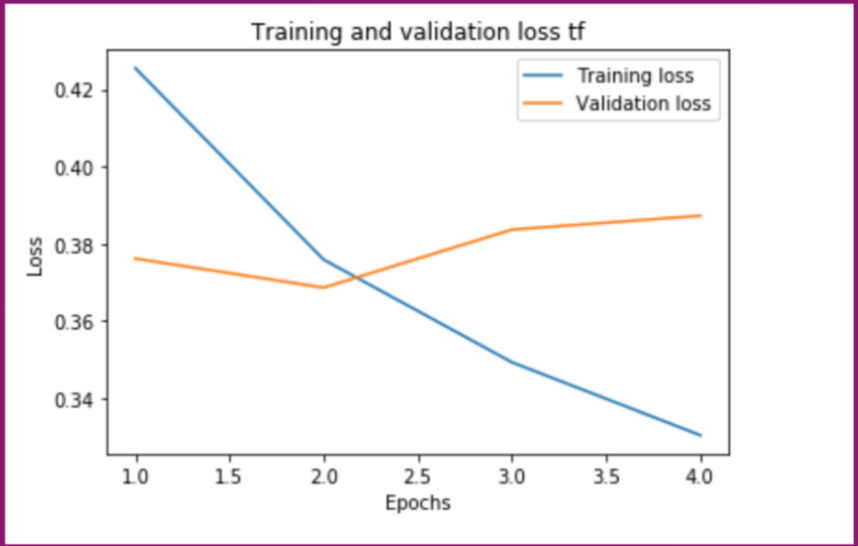
SUDDIVISIONE DATI:
Train/test: 80% - 20%
Train/Validation: 90% - 10%

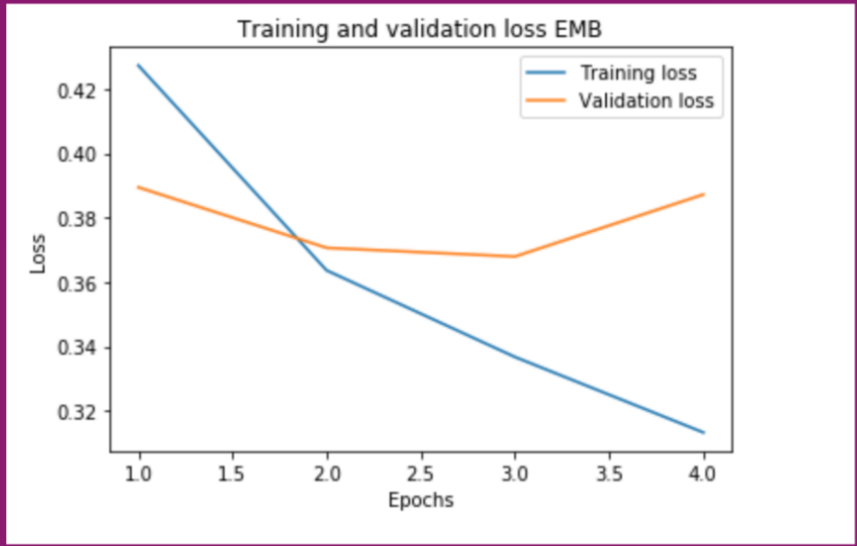



WHO WINS?
Accuracy
Robustness


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Accuracy
0.8685
0.8622
TF-IDF
Word Embeddings
Text Representation Techniques

Overfitting?






IMPROVEMENTS
• Più capacità computazionale
• Più training data
• Ottimizzazione