
# ANALISI RECENSIONI AMAZON

*Luca Gabellini (777786)*
*Matteo Provasi (782922)*
*Pierluigi Tagliabue (835211)*


# STEP

**01**

**Dati**

Caricamento dei dataset con recensioni prodotti

**02**

**Text preprocessing & representation**

Token, lemming, stopwords, …

**03**

**Modelli**

Implementazione text classification

**04**

**Analisi & conclusioni**

Capacità classificative dei modelli

# Step 1

## Dati

- File
- Formato
- Dataset

01

# DATI

## File

- Dataset Amazon con recensioni di prodotti
- Selezionati 9 file, uno per categoria

## Formato

- Dati originali in formato JSON
- Dopo il caricamento le recensioni sono state memorizzate in un file *pandas*

## Dataset

- Campionamento casuale senza reinserimento
- Variabili:
  - TextReview
  - Category

# Step 2

## Text preprocessing

- Tokenization
- Normalization
- Lemmatization
- Stopwords removal

# 02 TEXT PREPROCESSING - 1

## Tokenizzazione

- Applicato non singolarmente ma nel preprocessing ad ogni step

- Per ogni documento
  - Split
  - Funzione di preprocessing
  - Join

## Normalizzazione

Due funzioni:

- Lower case: parole in minuscolo

- Rimozione della punteggiatura e dei caratteri speciali

# 02 TEXT PREPROCESSING - 2

## Lemmatizzazione

Parole riportate nella loro forma base eliminando le inflessioni del contesto.

Applicato su:

- Verbi
- Sostantivi
- Aggettivi

## Stopwords

Rimosse le stopwords utilizzando una lista predefinita

Lista estesa con altri termini dopo un primo step.

# 02 TEXT PREPROCESSING - 3

*Esempio di text preprocessing su una recensione del dataset:*

1. **Documento originale:**
 GREAT light so far....after 8 months it is going strong ...i hope it lasts....i can believe it has a motion sensor....!!

2. **Testo convertito a lower case:**
great light so far....after 8 months it is going strong ...i hope it lasts....i can believe it has a motion sensor....!!

3. **Rimozione punteggiatura:**
great light so far  after 8 months it is going strong  i hope it lasts  i can believe it has a motion sensor

4. **Testo lemmatizzato:**
great light so far after 8 month it be go strong i hope it last i can believe it have a motion sensor

5. **Rimozione stopwords:**
great light far month strong hope last believe motion sensor

# 02 TEXT REPRESENTATION

## Train & test

L’interno dataset è stato diviso in due parti:

- Train: contenente il 75% delle osservazioni, sarà utilizzato per allenare i modelli.
- Test: contiene il 25% delle osservazioni, dati di classe «ignota» su cui si valuteranno le performance dei modelli.
- Stratificato per categorie

## Bag of words

- Rappresentazione più consona per il task. Ogni recensione viene convertita in un vettore con componenti date dai pesi TF-IDF.
- Algoritmo *TfidfVectorizer*
- addestrato sul train, restituisce una rappresentazione dei dati in forma matriciale (document-term matrix).
- Scelta delle 5000 features più rilevanti

# Step 3

## Modelli

- Descrizione
- Naïve Bayes
- Random Forest
- Logistic

03

# MODELLI

## Naïve Bayes

- Probabilità di appartenenza calcolata sfruttando il teorema di Bayes
- Correzione di Laplace per ovviare alla presenza di parole sconosciute

## Random Forest

- Media i risultati di *n* alberi classificativi
- Ogni albero contiene un subset di massimo $\sqrt{k}$ features
- Split dei nodi in base all'Indice di Gini

## Logistico

- Metodo *Multinomial*
- Probabilità di appartenenza calcolata con la funzione *softmax*
- Recensione assegnata alla classe con probabilità più elevata

# Step 4

## Analisi & risultati

- Precision/Recall
- Confusion matrix
- Commenti

# 04 ANALISI & RISULTATI - 1

Test $\chi^2$ per rilevanza condizionata:

$$X^2(f,t) = \frac{N(AD - CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

**# Baby:**
1. baby
2. diaper
3. bottle

**# Cell Phones and Accessories:**
1. phone
2. case
3. charge

**# Digital Music:**
1. album
2. song
3. quot

**# Musical Instruments:**
1. guitar
2. string
3. pedal

**# Office Products:**
1. printer
2. ink
3. paper

**# Pet Supplies:**
1. dog
2. cat
3. treat

**# Tools and Home Improvement:**
1. tool
2. light
3. Bulb

**# Toys and Games:**
1. doll
2. toy
3. kid

**# Video Games:**
1. game
2. graphic
3. play

# 04 ANALISI & RISULTATI - 2

## Performance modelli

Tutti i modelli registrano buone performance classificative in termini di accuracy:

- Random forest (100 alberi), meno performante
- Accuracy più elevata con il logistico

## Precision & recall

La classe *Musical Instruments* presenta un problema:

- Alta precision
- Bassa recall

Il numero di osservazioni in questa classe è significativamente minore rispetto a tutte le altre

# 04 ANALISI & RISULTATI - 3

## Ricampionamento

Per avere una migliore recall si sono prese tutte le osservazioni disponibili per *Musical Instruments*

La numerosità diventa uguale a quella della seconda classe meno numerosa

Nuovo dataset con solo questa categoria

## Preprocessing

- Si sono svolti tutti gli step di preprocessing elencati in precedenza
- Dal dataset originale sono rimosse le osservazioni relative a *Musical Instruments* e sostituite con il nuovo dataset
- Nuovo split train/test

# 04 ANALISI & RISULTATI - 4

## Accuracy

| | Accuracy globale | |
|---|---|---|
| Modello | Dataset originale | Dataset ricampionato |
| Naive Bayes | 0.883 | 0.883 |
| Random Forest | 0.864 | 0.864 |
| Logistic | 0.908 | 0.906 |

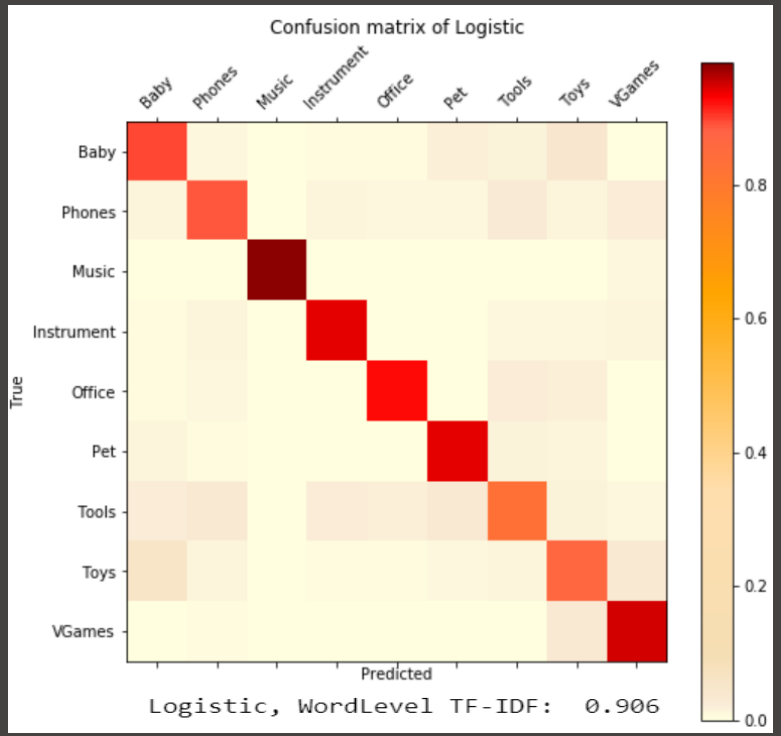
I valori medi rimangono invariati fra nuovo e vecchio dataset

## Precision & recall

Recall relativa alla classe ricampionata, *Musical Instruments*:

| | Dataset originale | | Dataset ricampionato | |
|---|---|---|---|---|
| Modello | Precision M.I. | Recall M.I. | Precision M.I. | Recall M.I. |
| Naive Bayes | 0.94 | 0.31 | 0.97 | 0.78 |
| Random Forest | 0.93 | 0.34 | 0.92 | 0.82 |
| Logistic | 0.91 | 0.62 | 0.94 | 0.85 |

# 04 ANALISI & RISULTATI - 4




Confusion matrix e riassunto della classificazione del miglior modello (logistic new dataset)

| | precision | recall |
|---|---|---|
| Baby | 0.90 | 0.88 |
| Cell_Phones_and_Accessories | 0.89 | 0.94 |
| Digital_Music | 0.98 | 0.97 |
| Musical_Instruments | 0.94 | 0.85 |
| Office_Products | 0.93 | 0.85 |
| Pet_Supplies | 0.94 | 0.92 |
| Tools_and_Home_Improvement | 0.83 | 0.88 |
| Toys_and_Games | 0.87 | 0.86 |
| Video_Games | 0.95 | 0.94 |
| micro avg | 0.91 | 0.91 |
| macro avg | 0.91 | 0.90 |
| weighted avg | 0.91 | 0.91 |


# 04 CONCLUSIONI

- Si è rivelato interessante l’esito dei termini più rilevanti per ciascuna categoria, che ha permesso di dare una visione sintetica delle features più rilevanti.
- Notevoli migliormenti con il ricampionamento.
- Random Forest modello «peggiore».
- Al variare del numero di features considerate, le performance del modello logistico rimangono stabili.


# Grazie dell’attenzione