Appello Febbraio 2022

TEXT MINING PROJECT

Sentiment analysis and topic modelling of Amazon e-books reviews

Alfredo Galli & Riccardo Rubini


Liked it! Lauren Dane is a great writer. I wish the story was longer but was a good read for the price
A very basic love story that is so boring and not believable. Can't believe Jane wrote this. Disappointing and not her best
It gave me some great recipes and ideas for meals to make for my family and its always fun to find new ideas on cooking


Liked it! Lauren Dane is a great writer. I wish the story was longer but was a good read for the price
A very basic love story that is so boring and not believable. Can't believe Jane wrote this. Disappointing and not her best
It gave me some great recipes and ideas for meals to make for my family and its always fun to find new ideas on cooking

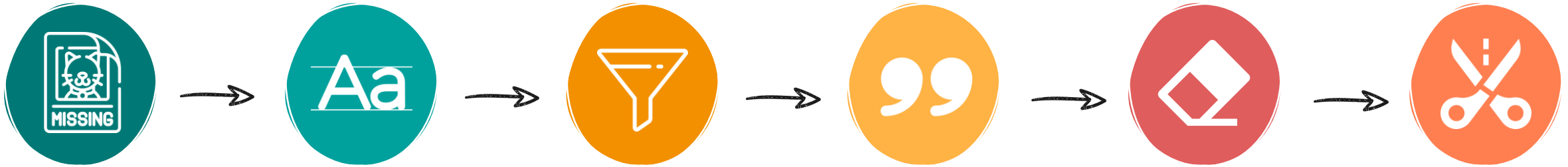
Descrizione Dataset

Il dataset utilizzato è stato ottenuto unendo i due file JSON, uno per le recensioni uno per i metadati, provenienti dalla repository *Amazon Review Data (2018)* di *Jianmo Ni*

- **486 mila** il numero totale di recensioni
- **331 caratteri** la lunghezza mediana
- **4,34 stelle** il voto medio
- **40 categorie** relative ai libri


well written
first book
read book
next book
recommend book
short story
really enjoyed
reading book
book serie
loved book
book read
relationship
feeling
husband
enjoyed reading
father
show
good story
wait see
love character
something
forward next
man
set
either
found
great job
problem
put book
woman
mind
might
easy read
still
son
add
plan
talk
child
often
instead
alway
sure
will love
though
story well
learn
hand
nothing
tell
copy book
fun read
year old
well developed
great read
great story
sex scene
use
book great
children
fact
exchange honest
wait next
although
seem
especially
name
perfect
girl
rather
book really
case
said
sister
everything
one thing
kid
course
enjoyed book
reason
story will
book one
book author
book well
life
realize
call
best friend
main character
looking forward
parent
needed
fell love
human
good
characters well
almost
work
novella
daughter
mean
moment
forward reading
someone
love serie
world
now
honest review
even
guy
experience
sometime
book love
idea
book good
read next
really
brother
good book
used
made
high school
women
bring
help
deal
meet
friend
fast paced
read author
yet
novel
need
enjoyed story
Highly recommended
secret
series book
men
highly recommend
read first
start
head
books read
loved story
books serie
even though
good read
making
must read
story line
thinking
other
situation
goe
book book
believe
actually
short storie
mother
love story
sort
book will
fall love
run
great book
least
become
may
family
thought
live
point
read one
next one
will definitely
one favorite
character development
really liked
kind
love book
quick read
beautiful
hero
knew
end book
past
begin
couple
maybe
wait read
look forward
first time
heroine
find
well done
writing style
second book
twists turn

Pre-processing workflow



Rimozione NA → lowecasing → Filtraggio stopwords → Rimozione punteggiatura → Rimozione white space → Stemming

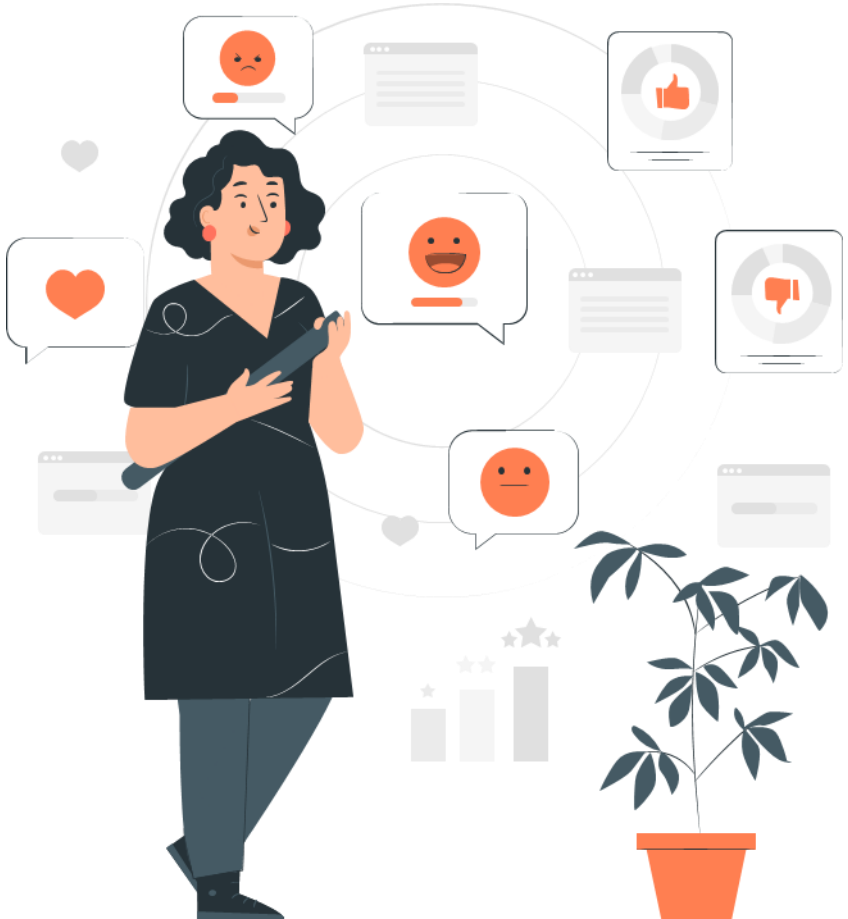
Testo non processato

“I really enjoyed the book. Had the normal back against the wall moments. It even had laugh out loud moments”

Testo processato

“realli enjoy book normal back wall moment even laugh loud moment”

SENTIMENT ANALYSIS



Creazione
Variabile Target

La variabile target è stata creata partendo dal rating espresso in stelle nel seguente modo:

- Si sono **eliminate le recensioni con rating pari a 3** essendo ambigue
- Le recensioni con **valutazione superiore a 4** sono state etichettate come **positive**
- Le recensioni con **valutazione inferiore a 2** sono state etichettate come **negative**

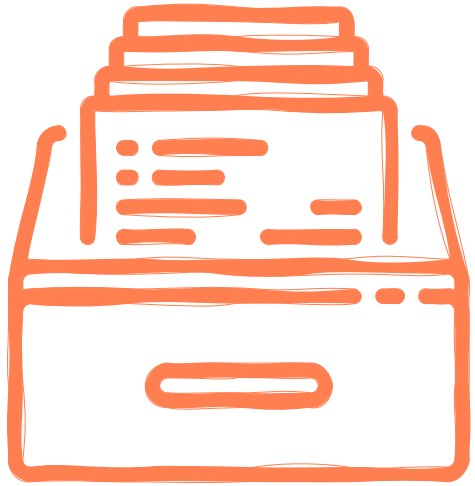
92.1%

La percentuale di recensioni classificate come **positive**

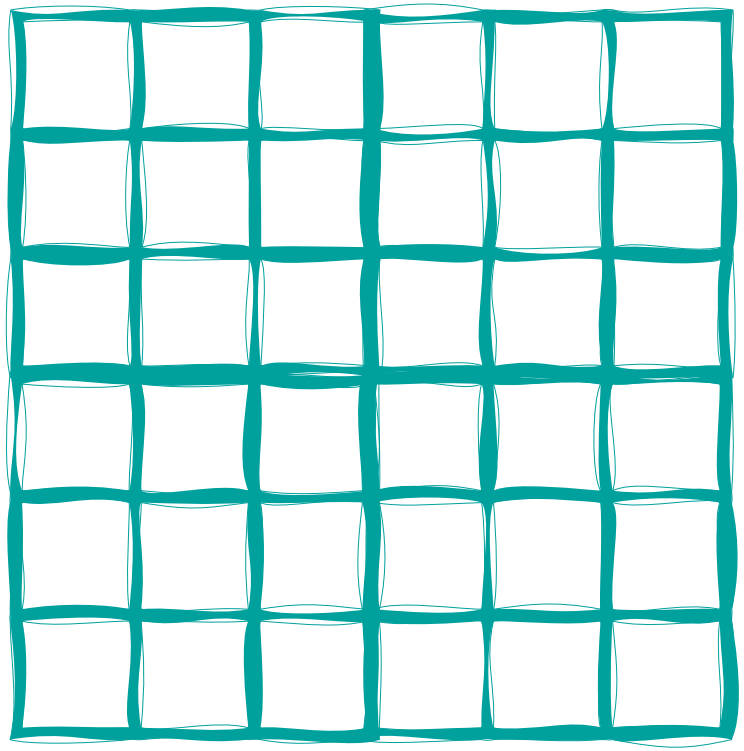
7.9%

La percentuale di recensioni classificate come **negative**

Recensioni Raw



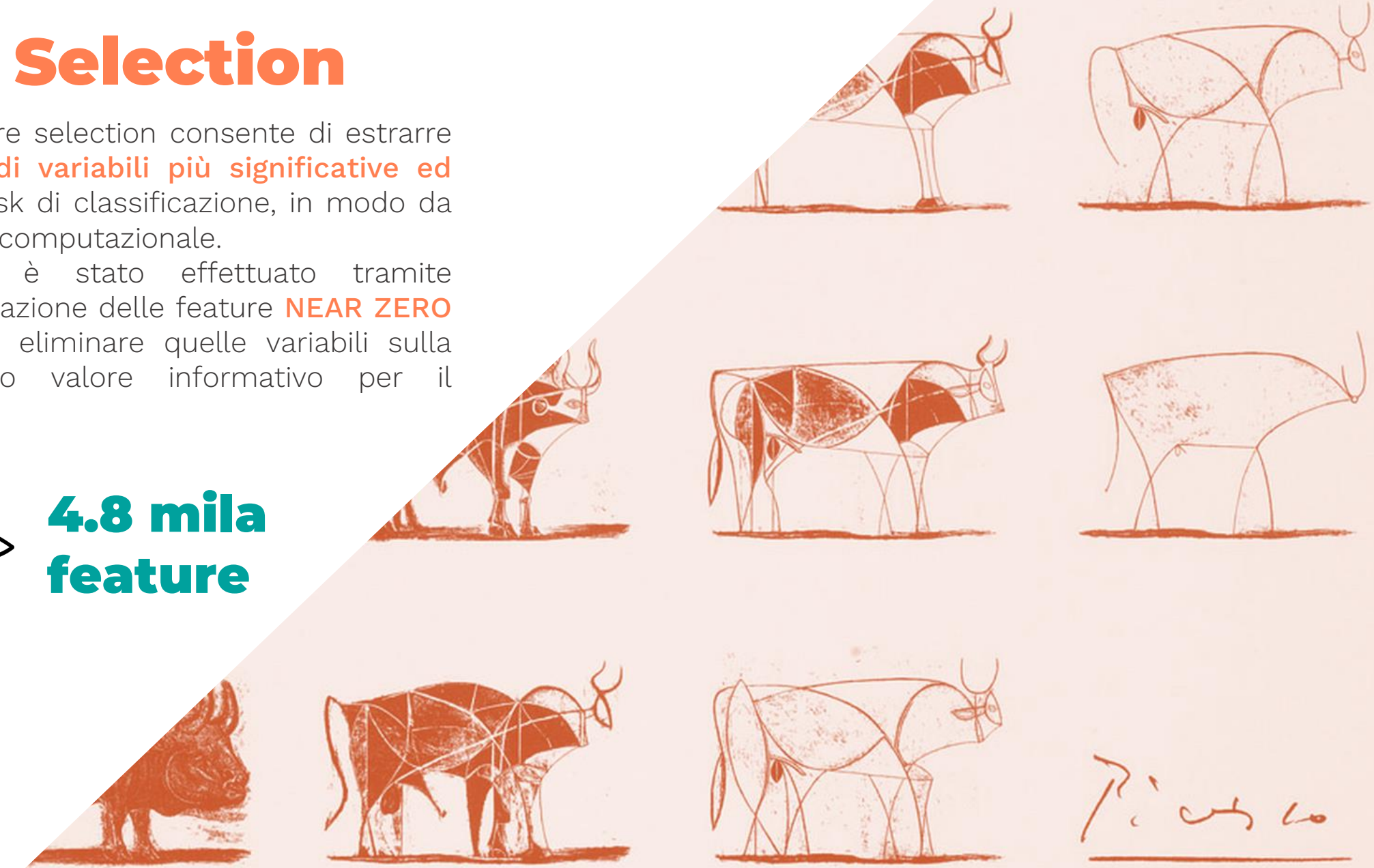
Rappresentazione TF-IDF con bi-grammi



Feature Selection

Il processo di feature selection consente di estrarre dal dataset il **set di variabili più significative ed essenziali** per la task di classificazione, in modo da alleggerire lo sforzo computazionale.
Questo passaggio è stato effettuato tramite l'approccio di eliminazione delle feature **NEAR ZERO VARIANCE**, così da eliminare quelle variabili sulla carta dallo scarso valore informativo per il classificatore

1.5 mln feature → **4.8 mila feature**



Logistic Lasso

Come classificatore la scelta è ricaduta su un **modello logistico penalizzato**, nel particolare con metodo Lasso. Si è fatta questa scelta così da:

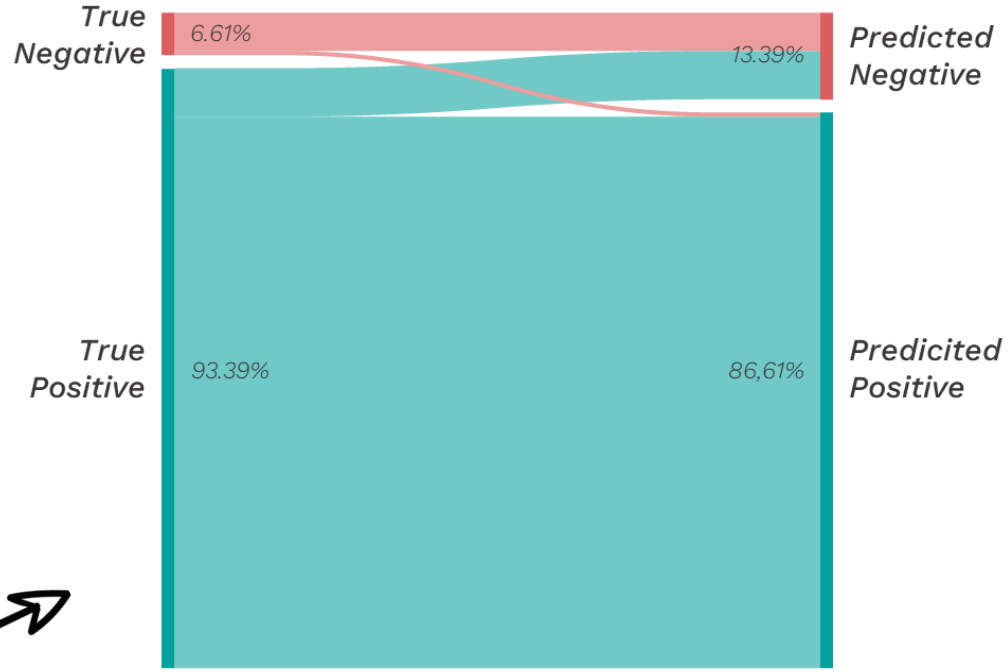
- Ottenere **risultati interpretabili**
- Eseguire una **model selection** già direttamente nella fase di training del modello. In totale le feature con coefficiente diverso da zero (i.e. selezionate) sono state 3268

Inoltre, dato la situazione unbalanced nella variabile target, si sono utilizzati pesi in modo da dare più rilevanza alle classificazioni negative corrette, in modo da bilanciare il classificatore.

Performance
Test Set

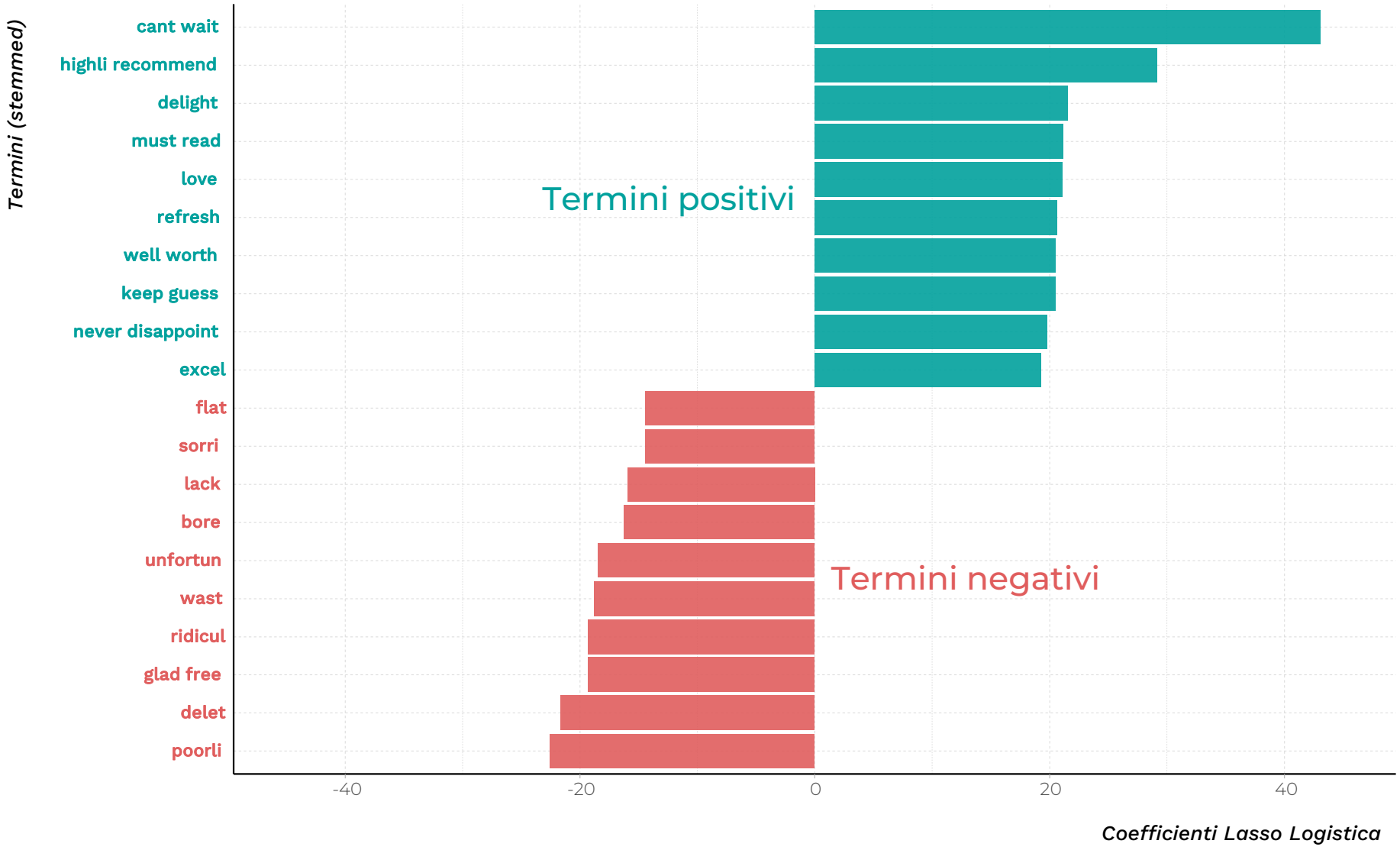
True class \ Predicted class	Positive Reviews	Negative Reviews
Positive Reviews	85.95%	7.44%
Negative Reviews	0.66%	5.95%

Calcolato su 110k recensioni di test



- **ACCURACY:** 92%
- **RECALL POSITIVE:** 92%
- **RECALL NEGATIVE:** 90%

Termini più discriminanti



TOPIC MODELING

Topic Modeling

Con l'applicazione del topic modeling il nostro obiettivo era quello di **estrarre** dalla collezione di recensione gli **argomenti principali latenti** discussi al loro interno. Per far ciò sono stati utilizzati due approcci:

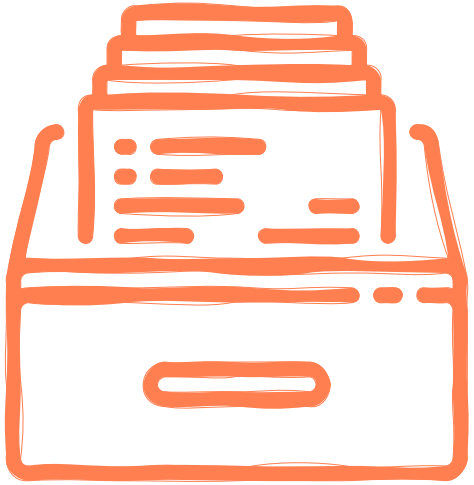
1. **NMF:** approccio puramente **algebrico** di riduzione della dimensionalità della matrice documenti-termini.
2. **LDA:** approccio di tipo **probabilistico**

Per evitare di includere parole troppo rare o troppo popolari, le quali avrebbero potuto inficiare sull'efficacia e l'efficienza degli algoritmi, si è scelto di escludere:

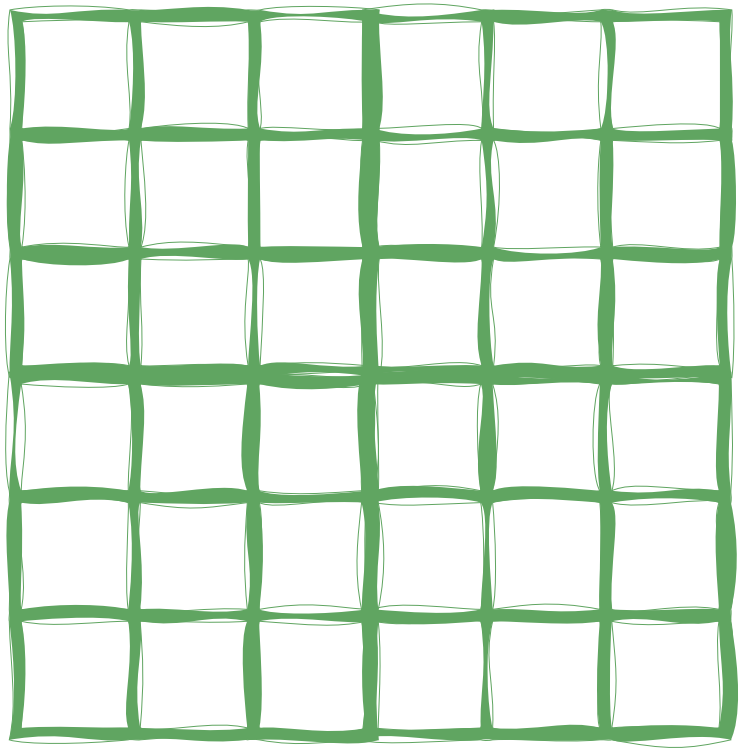
- Termini presenti in **meno di 20 recensioni**
- Termini presenti in **più del 30% delle recensioni**

NMF representation

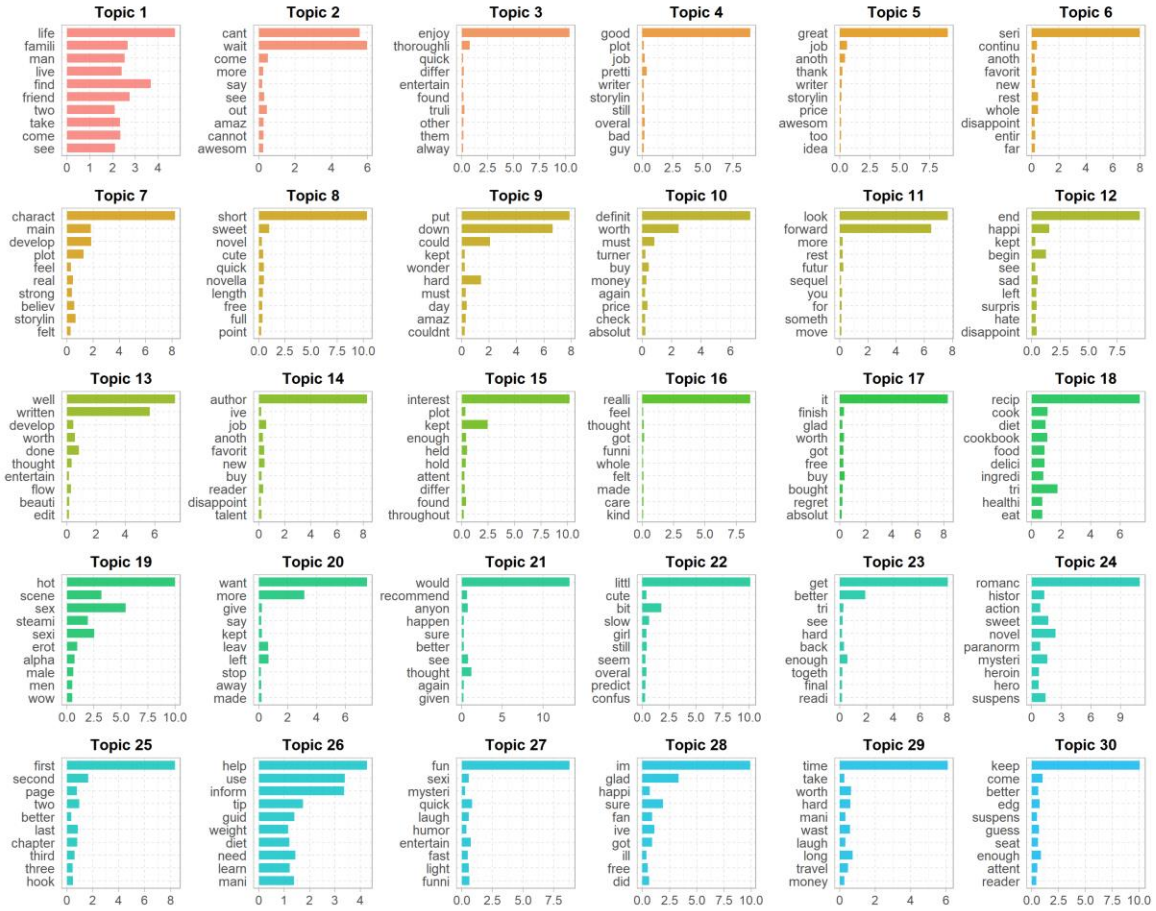
Recensioni Raw

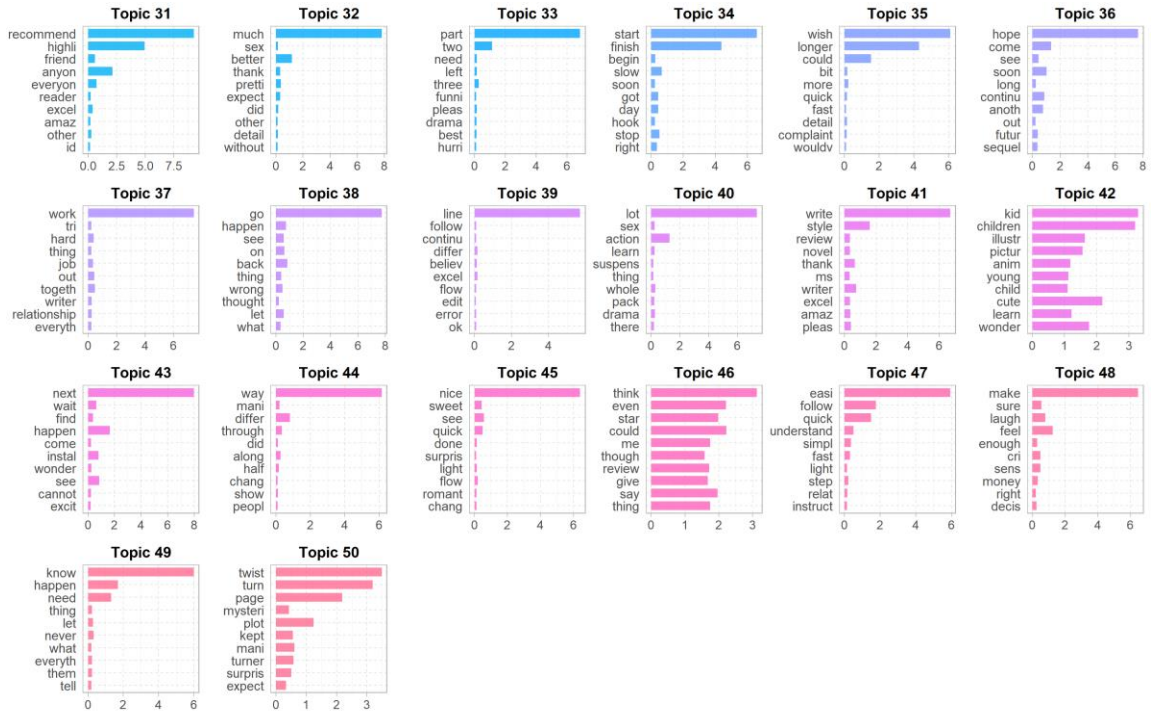


Rappresentazione TF-IDF con uni-grammi



NMF top 50 Topic

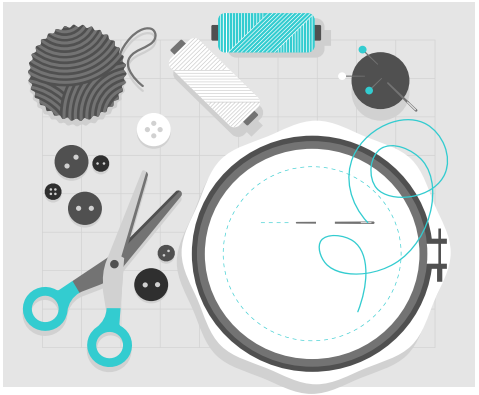




Confronto NMF Topic con Book Category

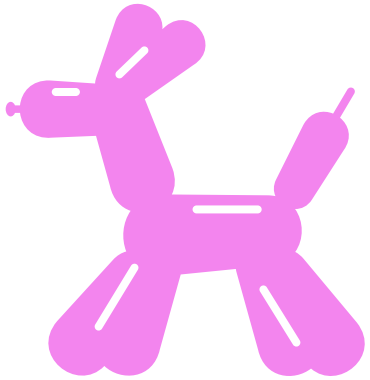
Book category

Crafts, Hobbies & Home

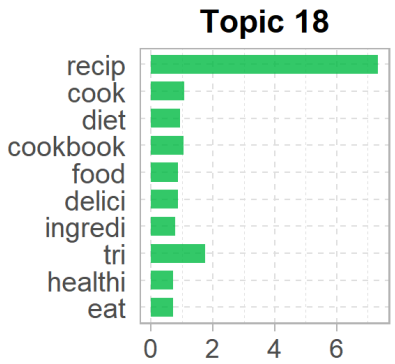


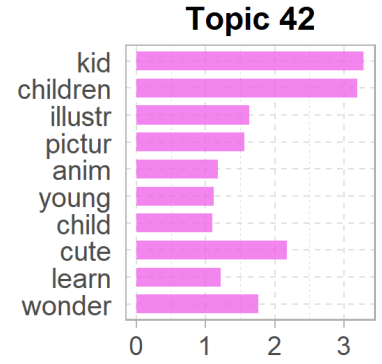
Cookbooks, Food & Wine

Children's eBooks



Main Topic





LDA Topic


Topic1
charact
bit
felt
seem
realli
would
littl
good
much
main
0
20000
40000
60000
80000
Topic2
brother
sister
her
father
famili
year
find
man
mother
marri
0
10000
20000
Topic3
im
it
go
want
end
say
know
get
got
realli
0
20000
40000
Topic4
author
review
use
write
reader
publish
kindl
mitch
amazon
word
0
10000
20000
30000
Topic5
life
peopl
help
chang
thing
us
person
live
way
make
0
10000
20000
30000
40000
50000
Topic6
enjoy
recommend
well
charact
great
look
written
forward
romanc
author
0
20000
40000
60000
Topic7
guy
get
girl
he
night
friend
olivia
date
meet
boyfriend
0
10000
20000
Topic8
heart
emot
beauti
feel
logan
kate
pain
tear
past
amaz
0
10000
20000
30000
Topic9
world
human
war
power
surviv
earth
battl
evil
finn
action
0
10000
20000
Topic10
recip
diet
inform
weight
guid
food
tip
eat
help
use
0
5000
10000
15000
20000
Topic11
wait
next
cant
seri
great
instal
come
see
sloan
hope
0
25000
50000
75000
100000
Topic12
hot
sex
sexi
scene
steami
sexual
alpha
male
erot
men
0
10000
20000
30000
40000
50000
Topic13
honest
seri
xchang
receiv
copi
novella
mate
shifter
ben
review
0
5000
10000
15000
20000
Topic14
kid
children
child
cute
young
parent
dog
age
old
aaron
0
5000
10000
15000
Topic15
tale
magic
adventur
king
steve
collect
jax
mason
fantasi
island
0
2500
5000
7500
10000
12500
Topic16
lili
cole
laugh
funni
h
humor
mysteri
murder
loud
hilari
0
2500
5000
7500
10000

RISULTATI PROGETTO

PERFORMANCE
Il modello classifica in maniera corretta 9 recensioni su 10, indipendentemente dalla loro polarità

SIGNIFICATO
Le parole più discriminanti in un verso e nell'altro emerse dal modello sono logiche e interpretabili

INTERPRETABILITÀ
I topic emersi sia con NMF che con LDA sono interpretabili e alcuni vanno oltre la pura sfera dell'opinione

COERENZA
Il topic NMF sono risultati allineati a quelle che sono le categorie d'appartenenza dei libri

GRAZIE PER L'ATTENZIONE