Spoiler Detector

Francesca Pezzuti, Lorenzo Massagli

Highlights

Main features of the application:

- In **SpoilerDetector,** users can search movies, read reviews and write reviews

- **SpoilerDetector** uses a classifier to identify **spoiler reviews** to avoid that user mistakenly read spoilers

Dataset Description

- Source:

  https://www.kaggle.com/rmisra/imdb-spoiler-dataset?select=IMDB_reviews.json

- Description

  573.913 reviews from 1998 to 2018

  1572 movies

- Volume

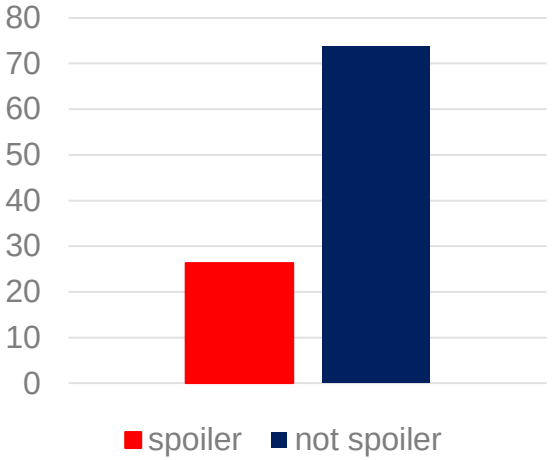
  976MB

WORKFLOW


Data analysis
Data preprocessing
Feature extraction
Learning & Model Evaluation
Classification

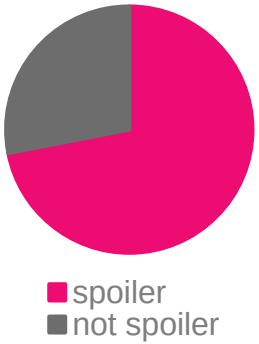
Data analysis

Class distribution

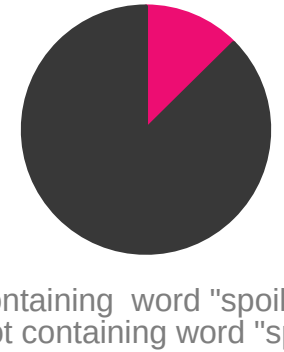


Spoiler distribution & selected words

Reviews containing the word "spoiler"



Spoiler reviews



DATA PREPROCESSING

CLEANING

- removed **links**
- removed **accents**
- removed **punctuation**
- removed **repeated characters**
- removed **special characters**
- expanded **contractions** of words
- removed **meaningless words** using a dictionary

STEMMING

- performed tests with/without stemming

EXAMPLE

"http://google.com/ this is treee okk test fòr cleaning, isn't it dsiadohaspi anomalies ?"

CLEANED TEXT

"tree test text cleaning anomaly"

STEMMED TEXT

"tree test text clean anomali"

LEARNING

Features representation

- Count vectorizer
- Term Frequency – Inverse Document Frequency (TF-IDF)

Models

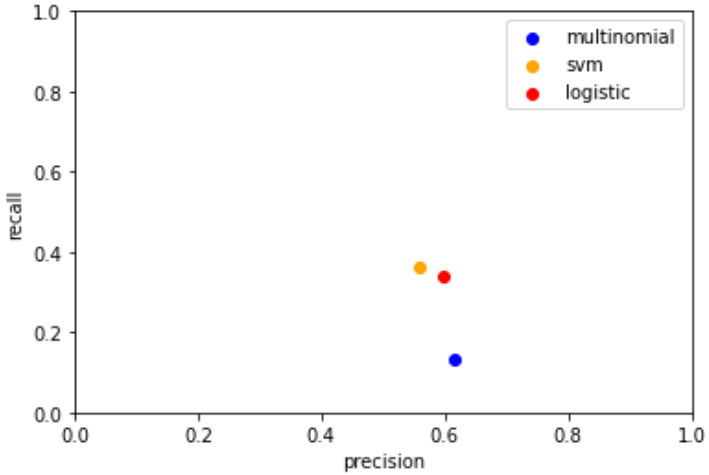
- Naïve Bayes Multinomial
- Support Vector Machine
- Logistic Regression

Training & Test

Test:	2017	**(45K reviews)**
Training:	2008-2016	**(288K reviews)**

Models Evaluation

Comparison between models



After performing a **10-folds cross-validation**, we plotted in the precision-recall graph the mean results of each model.

SVM and **Logistic regression** are the two models that show a better trade-off between precision and recall, so we decided to go more in detail with this two models

	Precision	Recall
Multinomial	0.62	0.13
SVM	0.56	0.36
Logistic	0.60	0.34

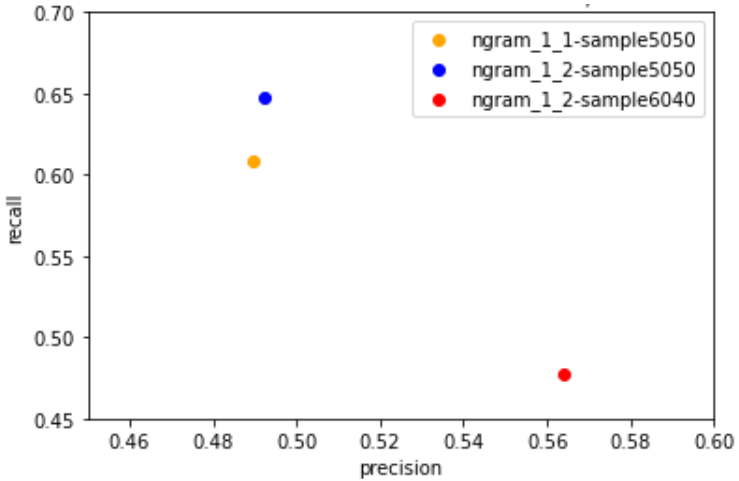
Models evaluation

Comparison between SVM & Logistic

SVM		
Parameters	Avg Precision	Avg Recall
ngram 1-1 sample 50-50	0.47	0.60
ngram 1-1 sample 60-40	0.52	0.45
ngram 1-2 sample 50-50	0.47	0.63
ngram 1-2 sample 60-40	0.52	0.50

Logistic regression		
Parameters	Avg Precision	Avg Recall
ngram 1-1 sample 50-50	0.49	0.61
ngram 1-2 sample 50-50	0.49	0.65
ngram 1-2 sample 50-50	0.56	0.47

Model Selection – Logistic



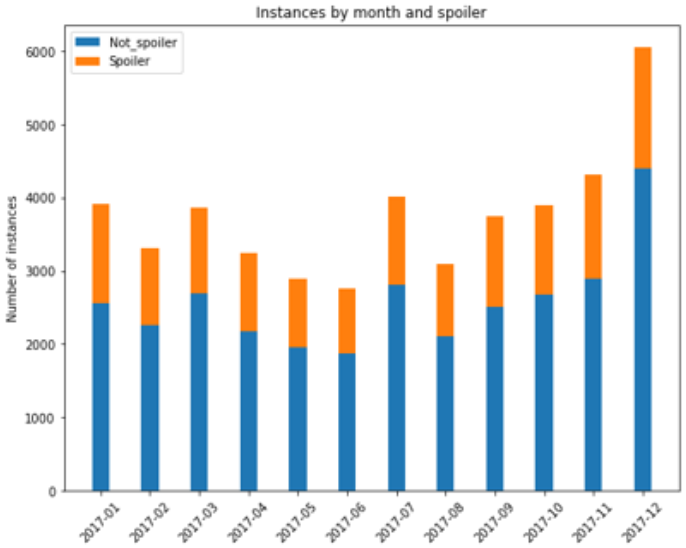
Logistic Regression		
Parameters	Avg Precision	Avg Recall
ngram 1-1 sample 50-50	0.49	0.61
ngram 1-2 sample 50-50	0.49	0.65
ngram 1-2 sample 50-50	0.56	0.47

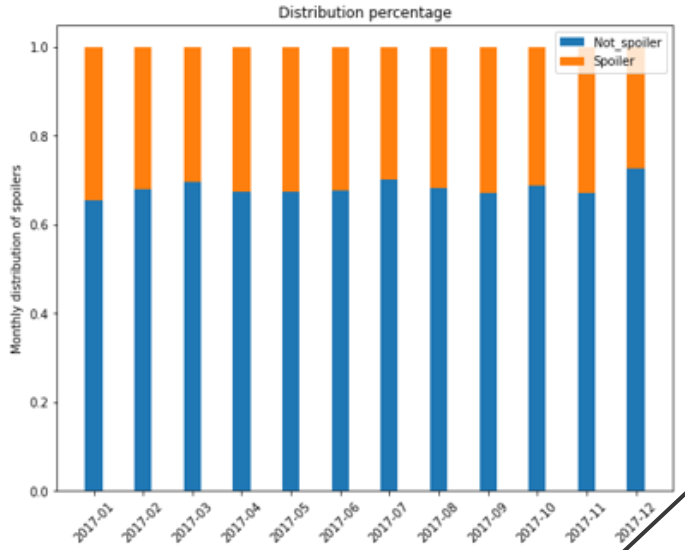
Model deploy - training

Training parameters

- Ngram range: 1,2
- Distribution: 50/50
- Stemming: false

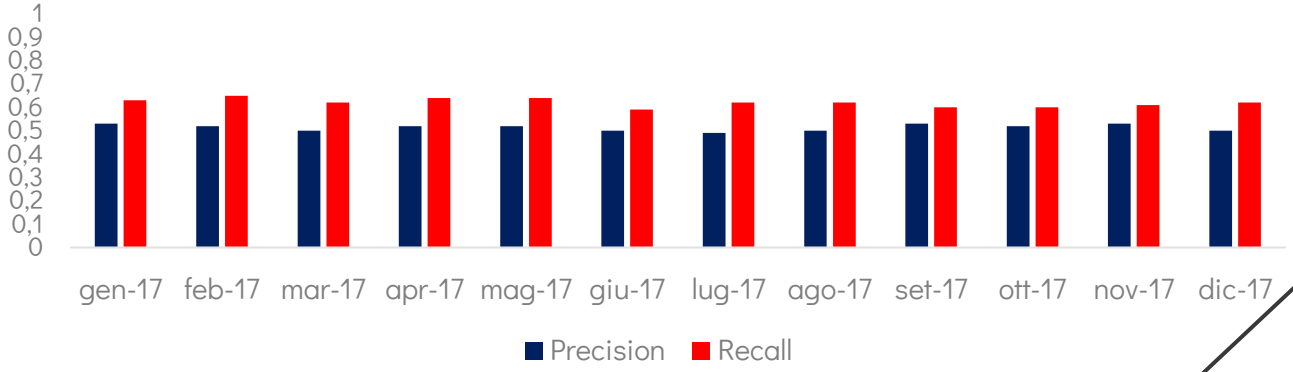
Model deploy - Evaluation





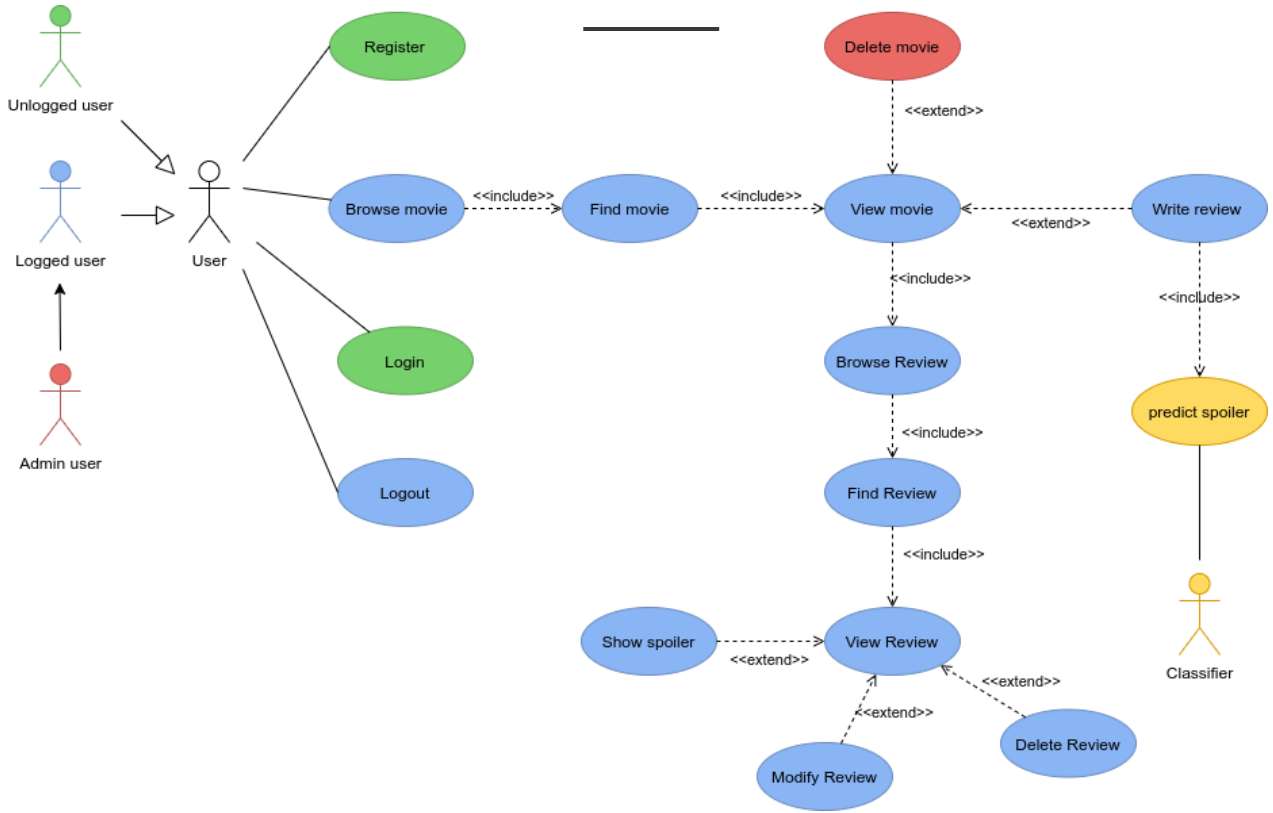
Model deploy - Evaluation

0,62	0,51
Average **recall**	Average **precision**

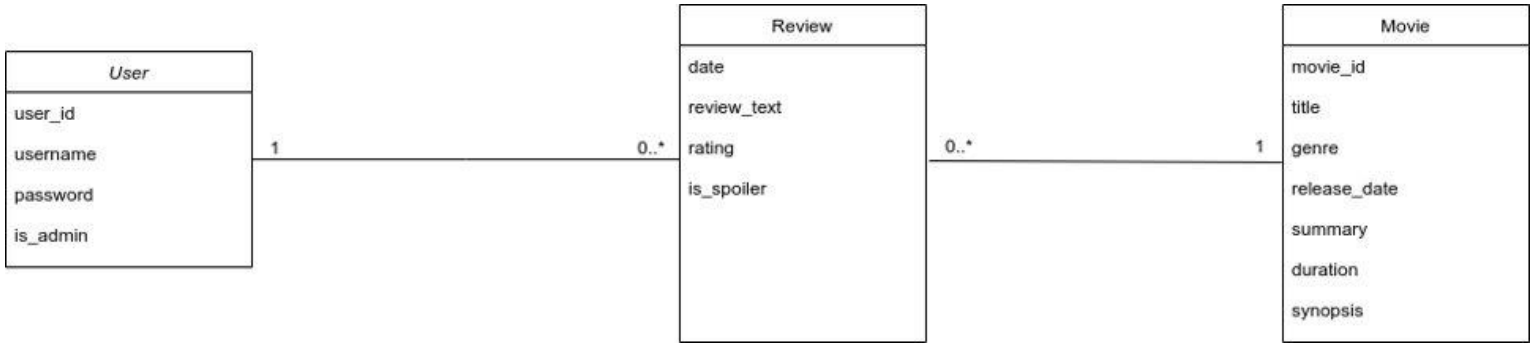


Application

USE CASE DIAGRAM



CLASS DIAGRAM



Data Model

User document

```
{
"_id": {"$oid": "62080e64a44c99e53aff21cb"},
 "username": "humorousPorpoise7",
 "password": "j0FH&aaP",
 "reviews": [
    {"_id": {"$oid":"62080e64a44c99e53aff1b68"},
     "title":"8 1/2",
     "review_date":"26 November 1998",
     "review_text":"8 1/2 has been known to bring a
        tear to my eye....",
     "rating":10,
     "is_spoiler":false
    },
    {"title": "I, Tonya",
     "review_date": "17 February 2022",
     "review_text": "Very cool movie!",
     "rating":7,
     "is_spoiler":false},
  "is_admin":0
}
```

Movie document

```
{"_id": {"$oid":"62080e64a44c99e53aff1908"},
 "title": "Old School",
 "genre": ["Comedy"],
 "release_date": "2003-02-21",
 "summary":"Mitch, Frank and Beanie are ...",
 "synopsis":"Attorney Mitch Martin comes back ...",
 "duration": "1h 28min",
 "cover_url": "https://i.imgur.com/R7KOk45.png",
 "reviews":[
        {"_id": {"$oid":"62080e64a44c99e53aff1fa2"},
        "username": "sadPie6",
        "review_date": "22 February 2003",
        "review_text": "Very cool movie!",
        "rating":8,
        "is_spoiler":false
        }]
}
```

Application – Spoiler Detector

Spoiler Review

Spoiler Revealed