Roberto Turrin - Moviri, ContentWise R&D
Daniele Loiacono – Politecnico di Milano, DEIB
Andreas Lommatzsch - TU Berlin, DAI-Labor

An analysis of the 2014 RecSys Challenge

Challenge description

- Engagement

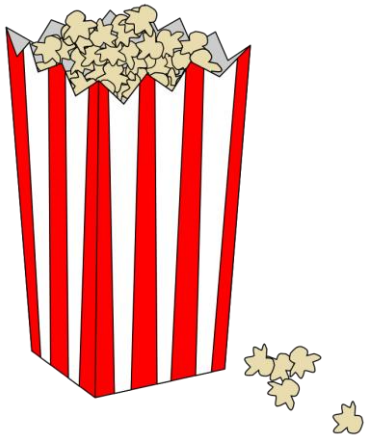
Challenge description

- Engagement
- **Context**:
  - movie ratings tweeted by users (using smartphone) with IMDb account connected to their Twitter accounts.
  - data about the tweets

"I rated The Matrix 9/10 http://www.imdb.com/title/tt0133093/ #IMDb"

Challenge description

- **Task**: predicting which movies generate the highest user engagement
  - –participant's algorithms should generate a **ranked list of tweets** which are ranked based on the amount of interaction.
  - –The interaction is defined as the sum of **retweet** and **favorite count**.



Challenge description

- The evaluation is based on nDCG@10.
  - computed for each user in the test set
  - then averaged over all users.

Rival

Blind retweets

- Twitter Users Don't Always Click the Links They Retweet
  - a weak correlation between retweets and clicks

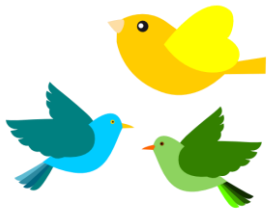
http://blog.hubspot.com/blog/tabid/6307/bid/33815/New-Data-Indicates-Twitter-Users-Don-t-Always-Click-the-Links-They-Retweet-INFOGRAPHIC.aspx

Agenda

- base analysis
- enrichment
- exploration
- reference predictor

Observations

- Each user has the same impact on the overall performance, regardless of how many tweets he/she posted.
- User role in the social network does not influence his/her nDCG@10



Dataset

Dataset	Users	Items	Tweets	Dates
Training	22,079	170,285	170,285	28/02/2013 - 08/01/2014
Test	5,717	4,226	21,285	08/01/2014 - 11/02/2014
Evaluation	5,514	4,559	21,287	11/02/2014 – 24/03/2014
All	24,924	15,142	212,857	28/02/2013 – 24/03/2014

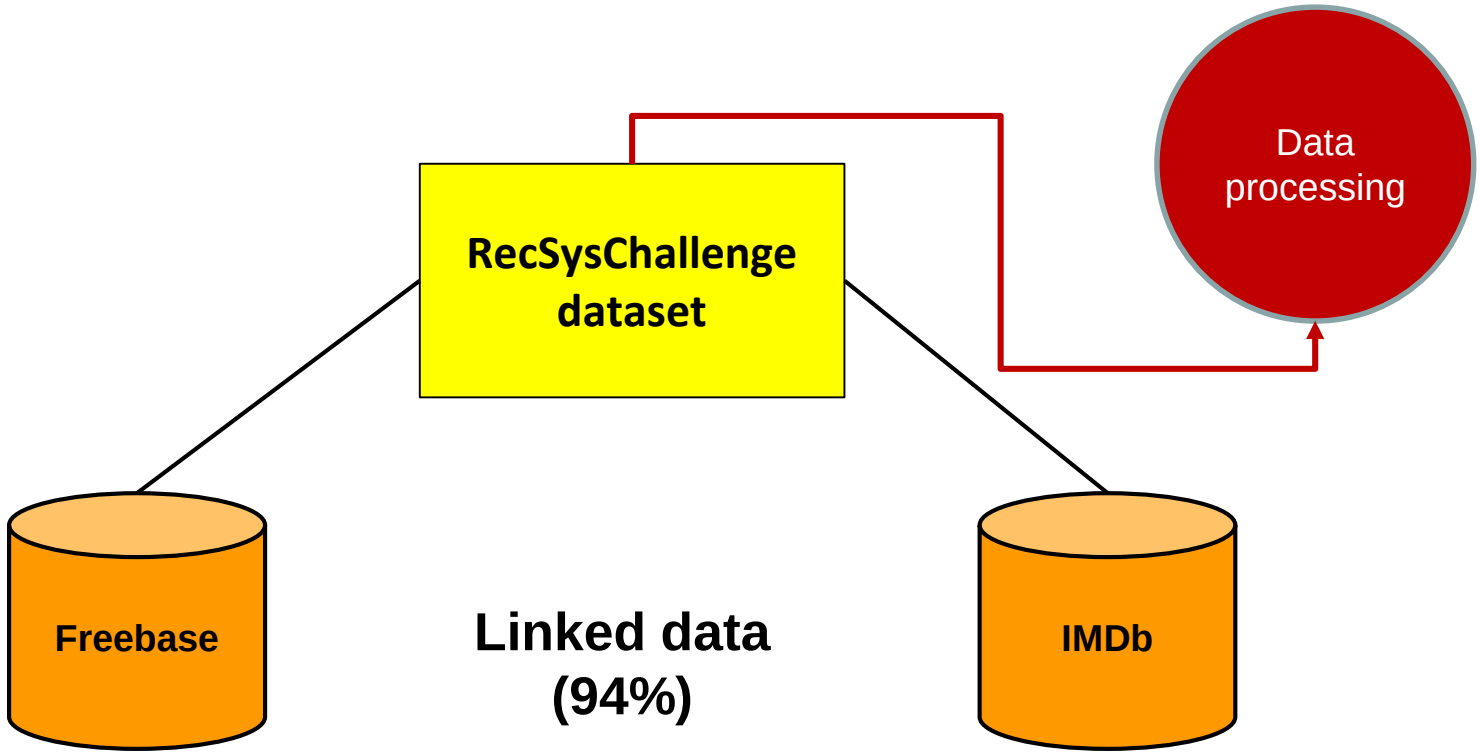
Dataset

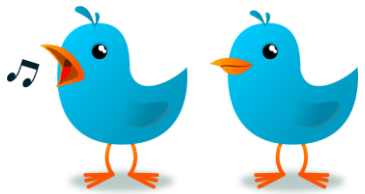
- user identifier
- tweet identifier
- timestamp of the message
- rating (in a 1-to-10 rating scale)
- additional information such as
  - IMDb url of the movie
  - the tweet has mentions
  - it is a retweet
  - tweet language
  - some user properties (e.g., the number of followers and the number of friends).

#tweets	#Users	
	All	Engaging
1	9477	3394
2-5	6406	1020
6-10	2319	106
11-20	1837	35
21-50	1482	16
51+	562	52
	22079	4577

21% users have engaging tweets
79% users have no engaging tweets

Enrichment


Data processing
RecSysChallenge dataset
Freebase
Linked data (94%)
IMDb



Enrichment: Freebase

- type, category, and genre, runtime and censure rating of the movie
- movie release date
- main movie language and main country
- number of won awards and estimated budget
- if the movie was adapted from a book
- number of festivals the movies attended
- if the movie is part of a series or a prequel/sequel

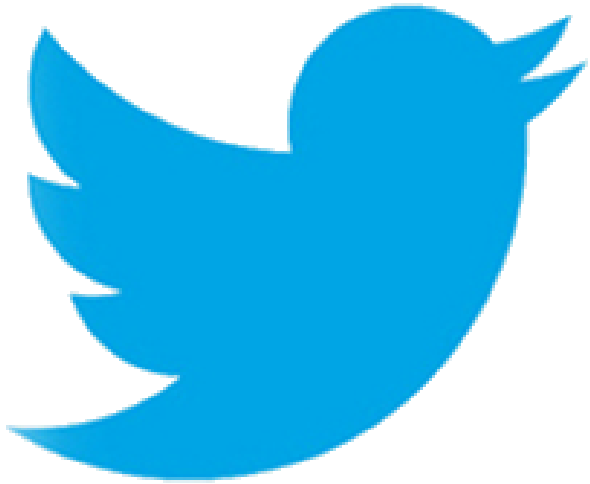
Enrichment: IMDb

- average IMDb rating
- number of IMDb raters

Enrichment: data processing

- related number of days between movie release and tweet

Problem re-formulation



"engaging"

"non-engaging"

binary problem

Binary upper bound

Tweet

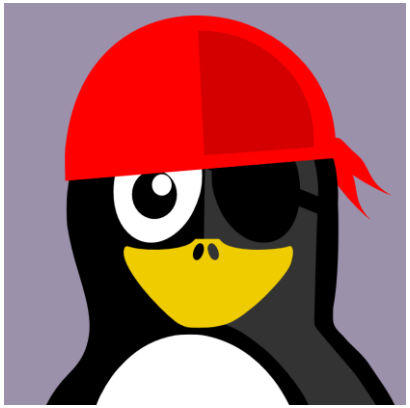
engaging

non-engaging

nDCG@10 = 0.9877

Non-engaging baseline

Tweet



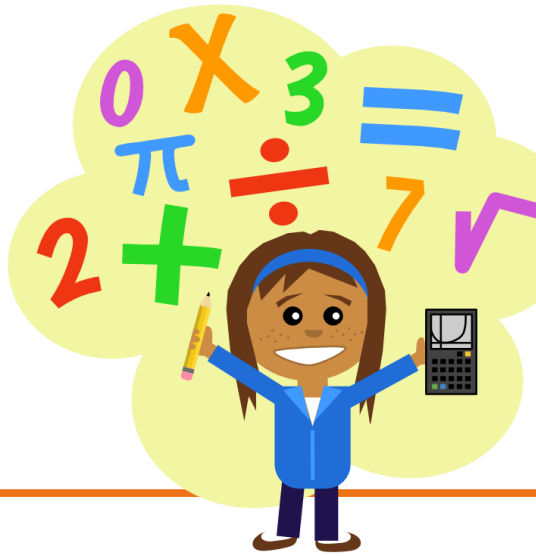
non-engaging

nDCG@10 = 0.7509

Nominal attributes

attribute	source	value	overall #tweets	proportion of tweets non-engaging	proportion of tweets engaging	average engagement	information gain
Movie lang	Freebase	Arabic	506	90.91%	9.09%	0.29	0.0268
		English	148056	95.46%	4.54%	0.23	
		French	2423	94.35%	5.65%	0.07	
		German	1608	89.05%	10.95%	0.16	
		Hindi	1127	94.50%	5.50%	0.07	
		Italian	705	95.46%	4.54%	0.06	
		Japanese	1521	95.46%	4.54%	0.07	
		Korean	601	94.68%	5.32%	0.07	
		Latin	1339	94.32%	5.68%	0.07	
		Russian	651	95.39%	4.61%	0.05	
		Spanish	4098	92.90%	7.10%	0.10	

- 10% of Arabic, German tweets are engaging **vs.** 4.5% of English tweets
- 6.4% of January tweets are engaging **vs.** 4% of September tweets
- 11% of German movies tweets are engaging **vs.** 4.5% of English movie

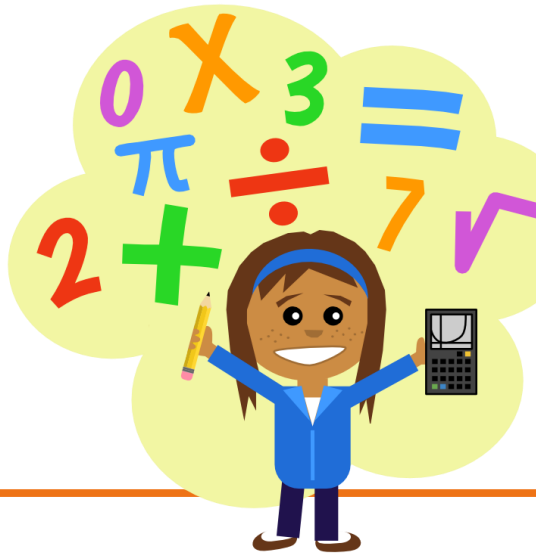


Numeric/Boolean attributes

attribute	source	Engaging tweet non-engaging	Engaging tweet engaging	overall	information gain
User rating	tweet	7.27	7.99	7.31	0.0220
Has mentions	tweet	0%	22.11%	2.03%	0.2017
Is a retweet	tweet	0.89%	24.72%	1.06%	0.2390
Has been retweeted	tweet	0%	17.36%	0.83%	0.2168

- 22% of engaging tweets have mentions **vs**. 0% of non-engaging tweets
- 24% of engaging tweets are a retweet itself v**s.** 0.9% of non-engaging tweets
- 17% of engaging tweets have been retweeted **vs.** 0% of non-engaging tweets

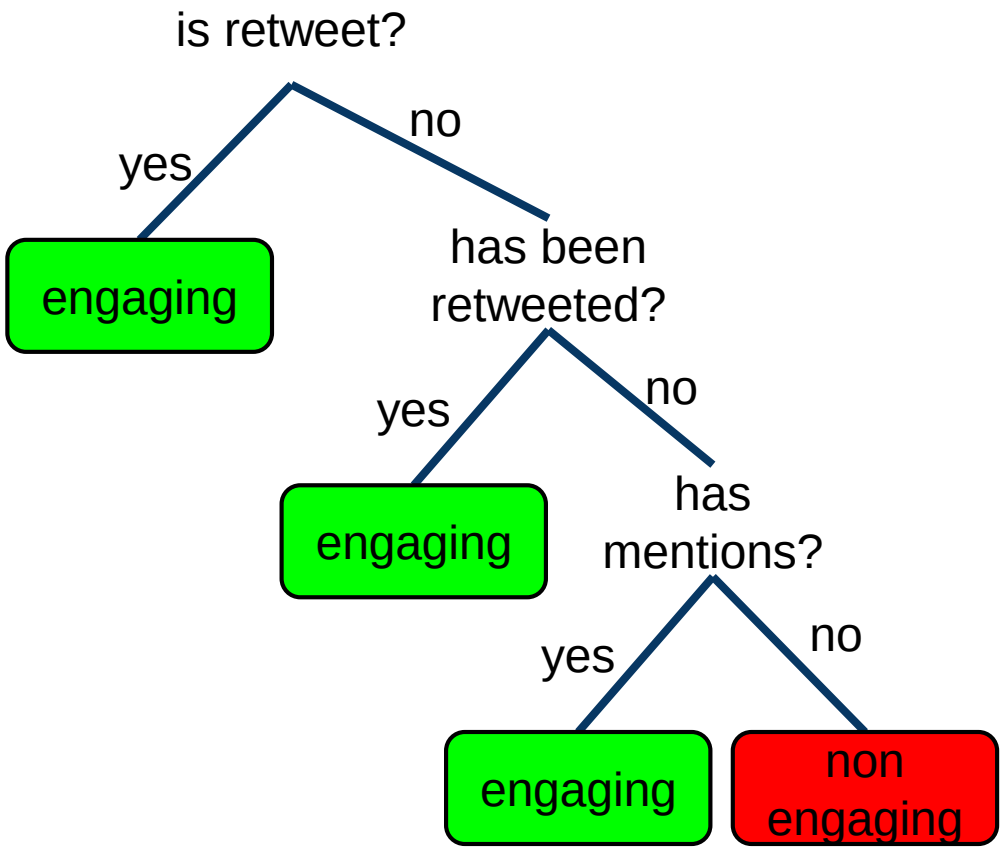
- Avg.#rating of engaging is 184 **vs.** 161 of non-engaging
- Difference between user rating and IMDb rating 0.74 for engaging **vs.** 0.27 for non-engaging
- Engaging tweets have won 2.74 awards **vs.** 2.28 and attended 2 festival **vs.** 1.5 (on average)



Machine learning

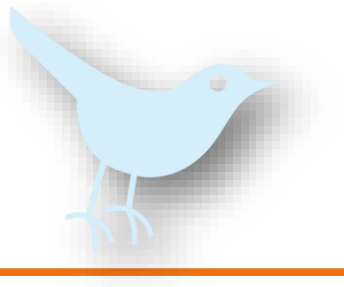
- Naive Bayes
- Bayesian Networks
- Decision Trees
- Pair learning

Machine learning: decision tree



Linear model

#	Added attrib	w_i	nDCG@10	increment
1	User rating	1000	0.8131	



Linear model

#	Added attrib	w_i	nDCG@10	increment
1	User rating	1000	0.8131	
2	#user followers	10	0.8146	0.0015

Linear model

#	Added attrib	w_i	nDCG@10	increment
1	User rating	1000	0.8131	
2	#user followers	10	0.8146	0.0015
3	#user favorites	1	0.8168	0.0022
4	#user friends	-3	0.8200	0.0032
5	Tweet language	n.a.	0.8212	0.0012

..and finally

eng(t) =

- rating(t) +
- has_mentions(t) +
- has_retweets(t) +
- is_a_retweet(t)

nDCG@10 = 0.8352

Conclusion

- Unbalance towards tweets with no engagement
- Most relevant attributes related to Tweet content, e.g.,: rating, mentions, retweet status