Detecting Singleton Review Spammers Using Semantic Similarity

Vlad Sandulescu

Adform, Denmark

joint work with

Martin Ester

School of Computing Science

Simon Fraser University

Burnaby, B.C., Canada

2015.05.19 - WWW 2015, Florence, Italy

Online reviews

- 31% of consumers read online reviews before actually making a purchase (rising)
- by the end of 2014, 15% of all social media reviews will consist of company paid fake reviews

Ken K.
Burke, VA
0 friends
4 reviews

4/12/2011

Immediately upon entering, we became aware of the fact that this is a unique and charming hotel. The main lobby is decorated by live vines overlapping the open-feeling roof and by chandeliers, quite a contrast. The hotel staff were courteous, welcoming and efficient. The room was tastefully decorated with plush, comfortable bedding and the street noises of New York were never noticeable. The location is convenient to everything in the area of Columbus Circle and Carnegie Hall and there is a subway nearby. Overall a lovely experience.

Ken K.
Burke, VA
0 friends
4 reviews

4/12/2011

Immediately upon entering, we became aware of the fact that this is a unique and charming hotel. The main lobby is decorated by live vines overlapping the open-feeling roof and by chandeliers, quite a contrast. The hotel staff were courteous, welcoming and efficient. The room was tastefully decorated with plush, comfortable bedding and the street noises of New York were never noticeable. The location is convenient to everything in the area of Columbus Circle and Carnegie Hall and there is a subway nearby. Overall a lovely experience.

Behavioural features
text analysis

- Behavioural approach gives good results for ”elite” users
- Textual analysis = mostly cosine similarity, but also linguistic cues of deceptive writing - using more verbs, adverbs and pronouns
- ”husband” or ”vacation” = highly suspicious based on their incidence in fake reviews
- ~ 90% of reviewers write a single review under one user name

- **What about the singleton reviewers?**

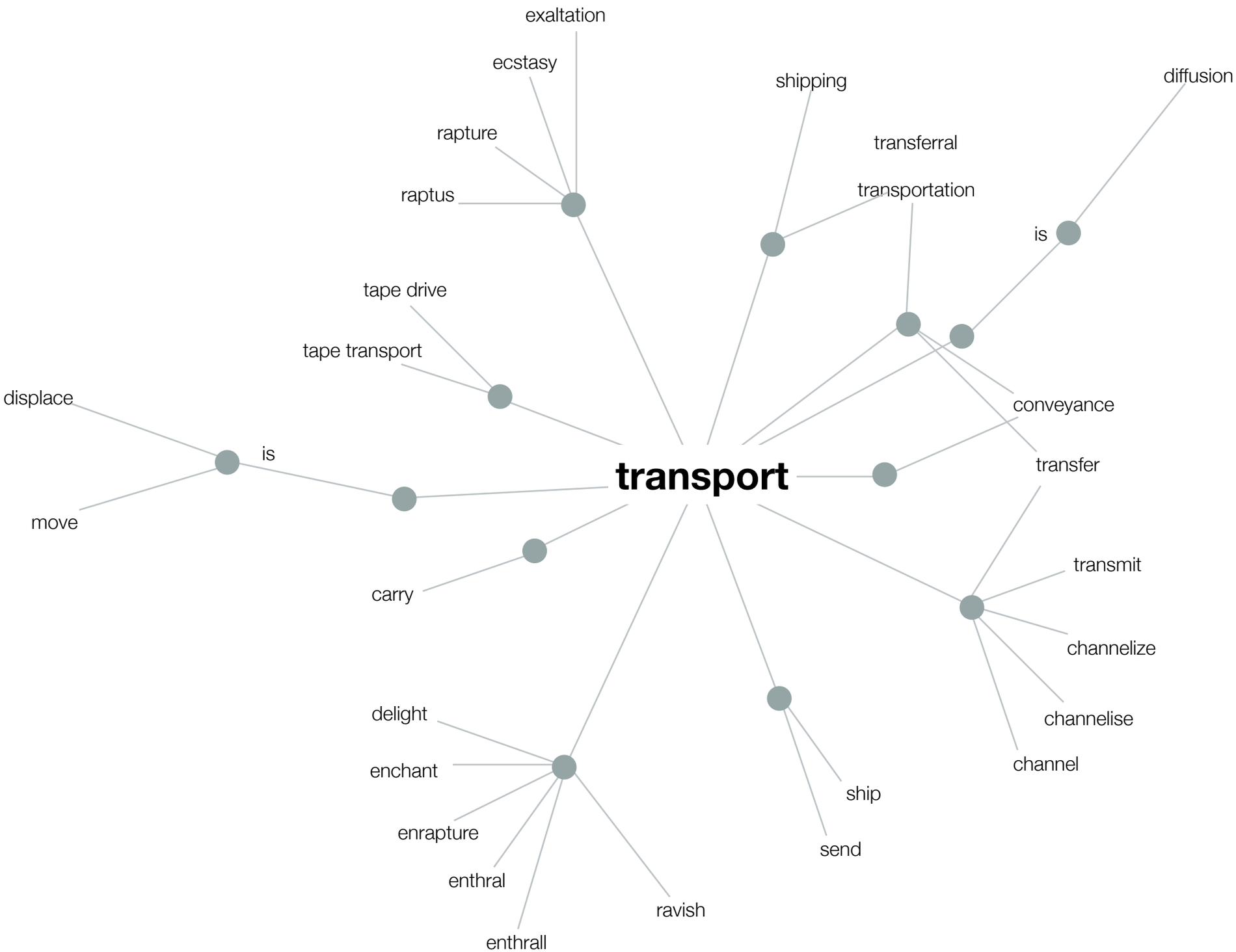
Hypothesis

- Semantic similarity measures should outperform vectorial based models in detecting more subtle similarities between fake reviews written by the same author
- A spammer's imagination is limited, so he will partially reuse some of the aspects between reviews, through paraphrase and synonyms

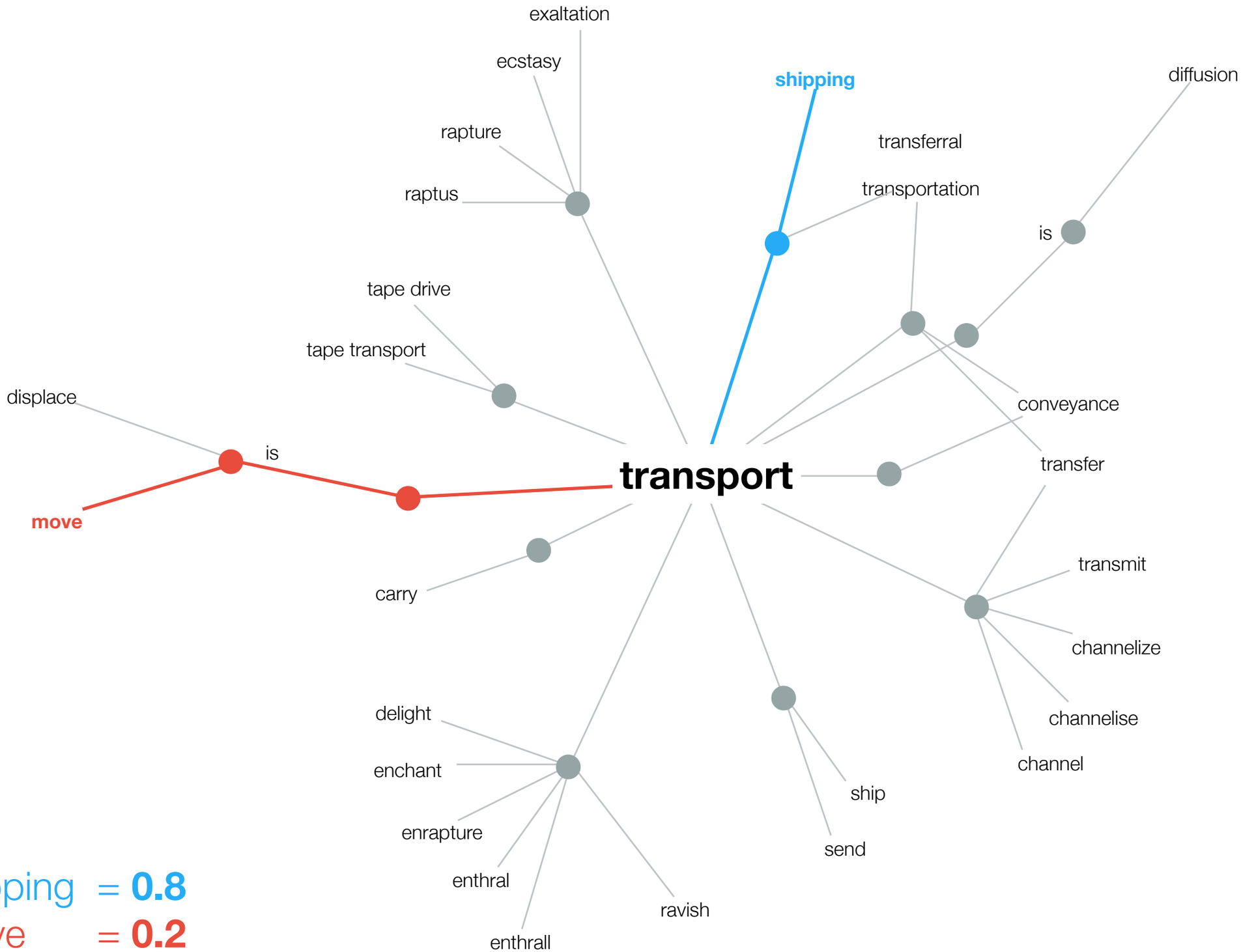
Goals

- Detect opinion spam using semantic similarity (WordNet) and topic modeling (LDA)
- Compare to vectorial similarity models (cosine)

Wordnet synsets



Wordnet synsets



transport - shipping = **0.8**

transport - move = **0.2**

Vectorial-based measures

For T1 and T2, their cosine similarity can be formulated as

$$\cos(T_1, T_2) = \frac{T_1 T_2}{\|T_1\| \|T_2\|} = \frac{\sum_{i=1}^{n} T_{1i} T_{2i}}{\sqrt{\sum_{i=1}^{n} (T_{1i})^2} \sqrt{\sum_{i=1}^{n} (T_{2i})^2}}$$

Knowledge-based measures

For T1 and T2, their semantic similarity (Mihalcea et al.) can be formulated as:

$$sim(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w \in \{T_1\}} (maxSim(w, T_2) * idf(w))}{\sum_{w \in \{T_1\}} idf(w)} + \frac{\sum_{w \in \{T_2\}} (maxSim(w, T_1) * idf(w))}{\sum_{w \in \{T_2\}} idf(w)}\right)$$

transport - ”The shop now offers night delivery”

Ken K.
Burke, VA
0 friends
4 reviews

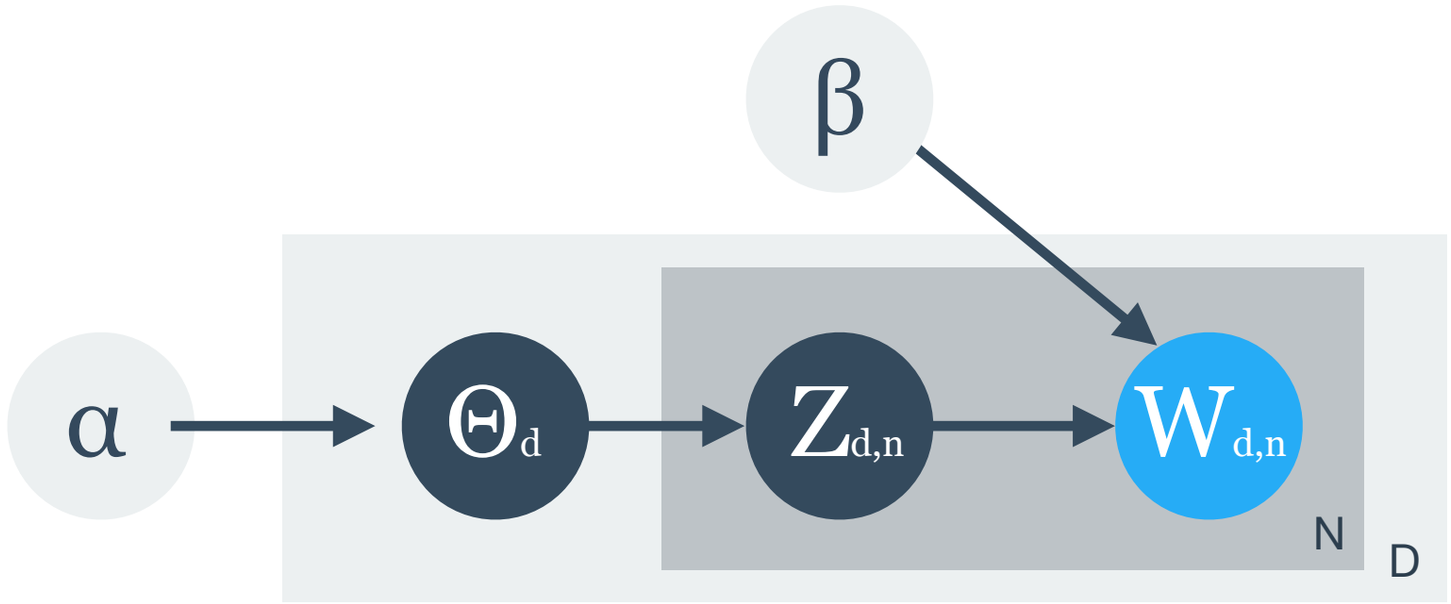
4/12/2011

Immediately upon entering, we became aware of the fact that this is a unique and charming hotel. The main lobby is decorated by live vines overlapping the open-feeling roof and by chandeliers, quite a contrast. The hotel staff were courteous, welcoming and efficient. The room was tastefully decorated with plush, comfortable bedding and the street noises of New York were never noticeable. The location is convenient to everything in the area of Columbus Circle and Carnegie Hall and there is a subway nearby. Overall a lovely experience.

Aspect-based opinion mining

- opinion phrases : <*aspect, sentiment*>
- opinion phrases: <*hotel, unique*>, <*hotel, charming*>, <*staff, courteous*>
- different words = same aspect (laptop, notebook, notebook computer)
- reviews = short documents = latent topics mixture = review aspects mixture
- reviews similarity = topics similarity => topic modeling problem
- advantage: language agnostic, not like WordNet

Topic Modeling for opinion spam detection



Θ_d represents the topic proportions for the **d**th document

$Z_{d,n}$ represents the topic assignment for the **n**th word in the **d**th document

$W_{d,n}$ represents the observed word for the **n**th word in the **d**th document

β represents a distribution over the words in the known vocabulary

---

$$KL(P\|Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i).$$

$$JS(P \parallel Q) = \frac{1}{2}KL(P \parallel M) + \frac{1}{2}KL(Q \parallel M), \quad where \quad M = \frac{1}{2}(P + Q)$$

$$IR(P, Q) = 10^{-\beta JS(P\|Q)}$$

yelp

TRUSTPILOT

Ott dataset

57K crawled reviews

from 660 New York restaurants

Recommended reviews = truthful
Not recommended = fake

9K labeled reviews

from 130 US and UK businesses

800 labeled reviews

from TripAdvisor and AMT

One submission per turker, rejected short, illegible or plagiarized reviews

Preprocessing

- Stop words removal, POS tagging (extracted NN, JJ, VB)

”I am working hard on my presentation at WWW”

I/PRP **am**/VBP **working**/VBG **hard**/RB **on**/IN **my**/PRP **presentation**/NN **at**/IN **WWW**/NNP

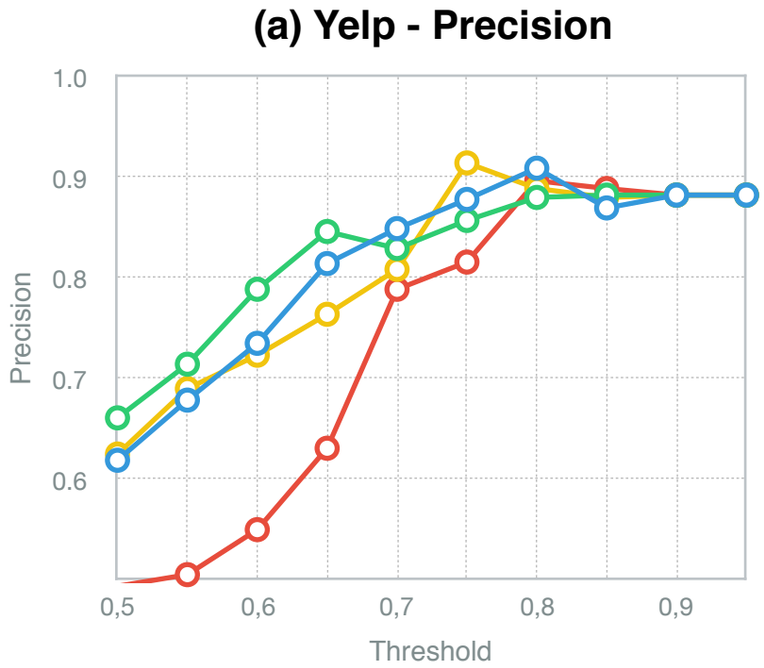
- am $\xrightarrow{\text{lemma}}$ be, working $\xrightarrow{\text{lemma}}$ work
- Cosine (all POS), Cosine (NN, JJ, VB), Cosine with lemmatization, Semantic

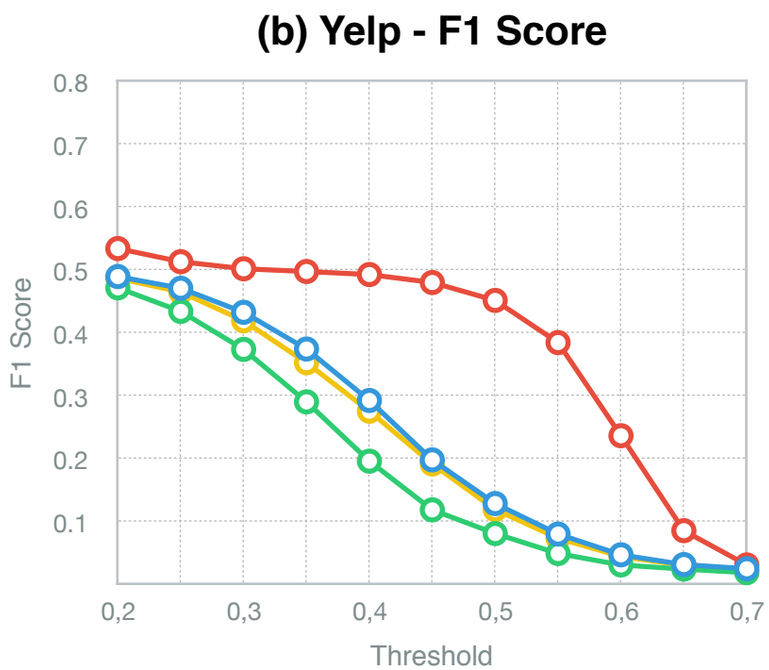
Pairwise similarity

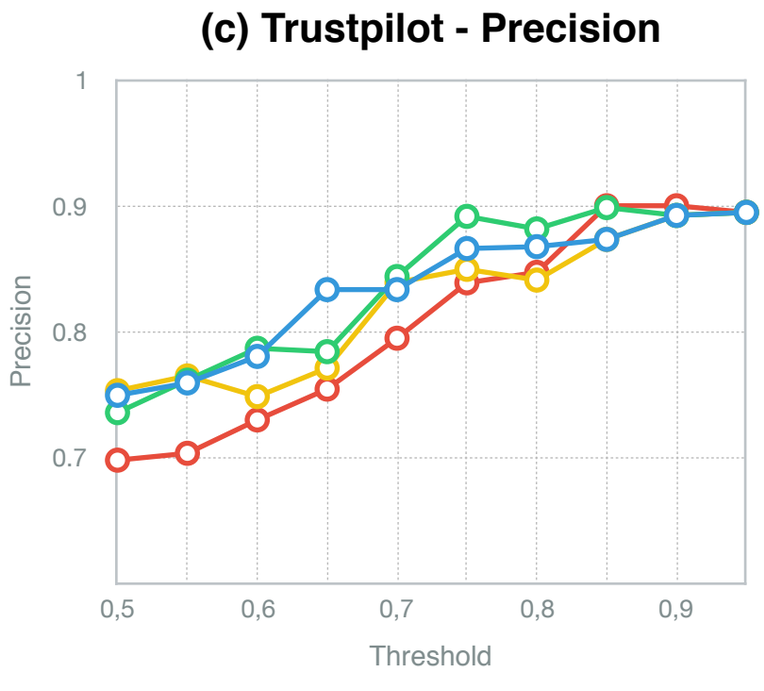
- ∀ pairs (Ri, Rj) ∈ business B
- if sim(Ri, Rj) > T, T ∈ {.5, 1} ⇒ Ri and Rj are fake, else truthful

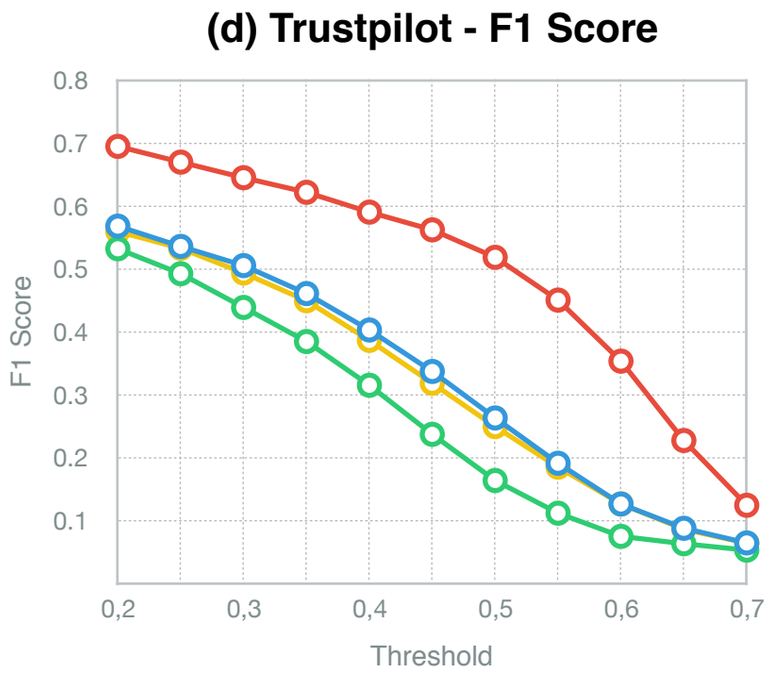
Semantic similarity results

Yelp/Trustpilot - classifier performance with vectorial and semantic similarity measures









cos cpnl cpl mih

CPL-↑P,T>0.75
↑T⇒↑P

P=90%, T>0.8
Semantic ↑ F1-score

P=90%, T>0.85
Trustpilot's spammers are lazy
Yelp's spam is higher quality

Distribution of truthful and deceptive reviews - Ott

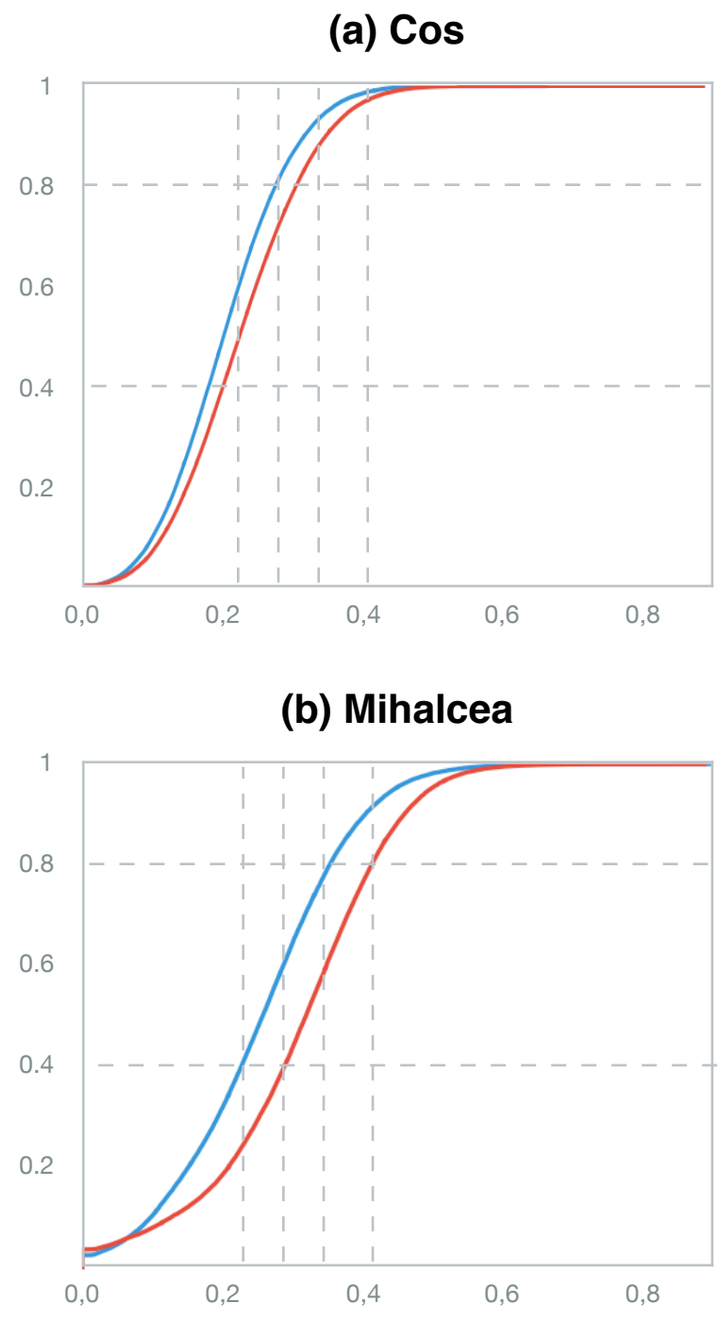
Cumulative percentage of reviews vs. similarity values

Vectorial ~ 2% diff

- 80% reviews ↑ 0.32
- 80% reviews ↑ 0.34

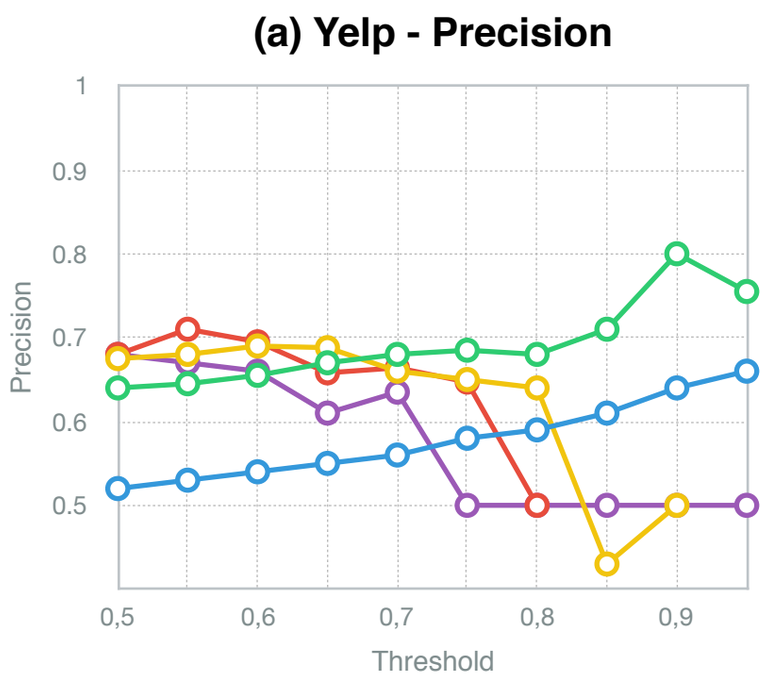
Semantic ~ 6-10% diff

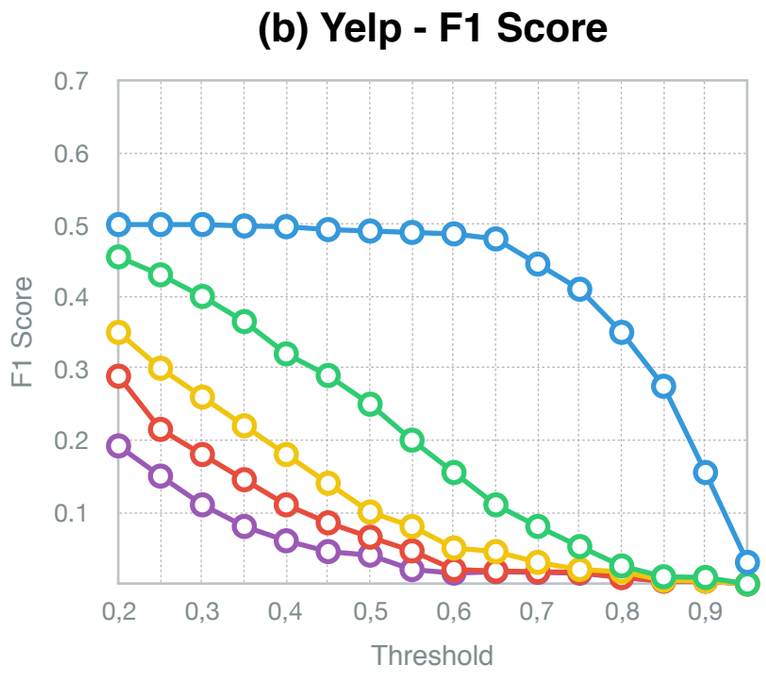
- 40% reviews ↑ 0.22
- 40% reviews ↑ 0.32
- 80% reviews ↑ 0.38
- 80% reviews ↑ 0.44

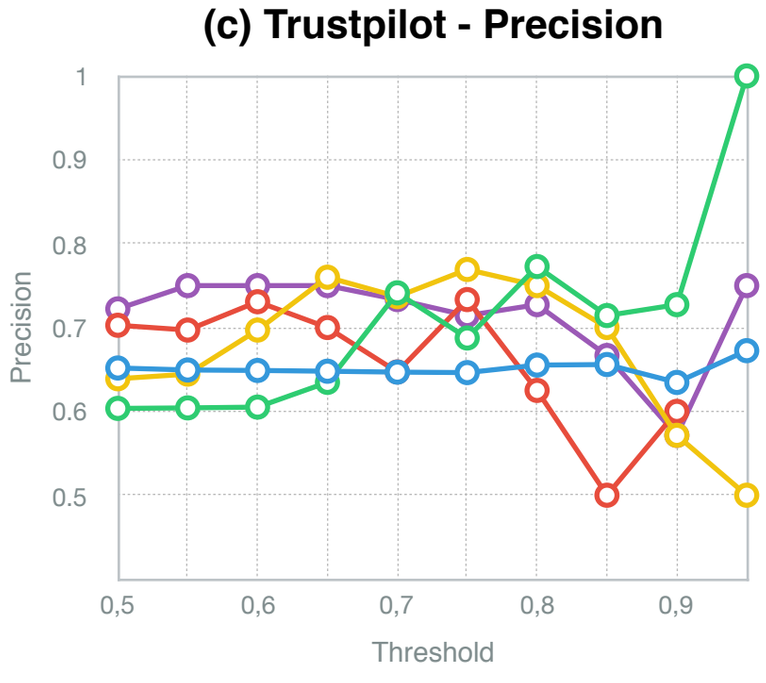


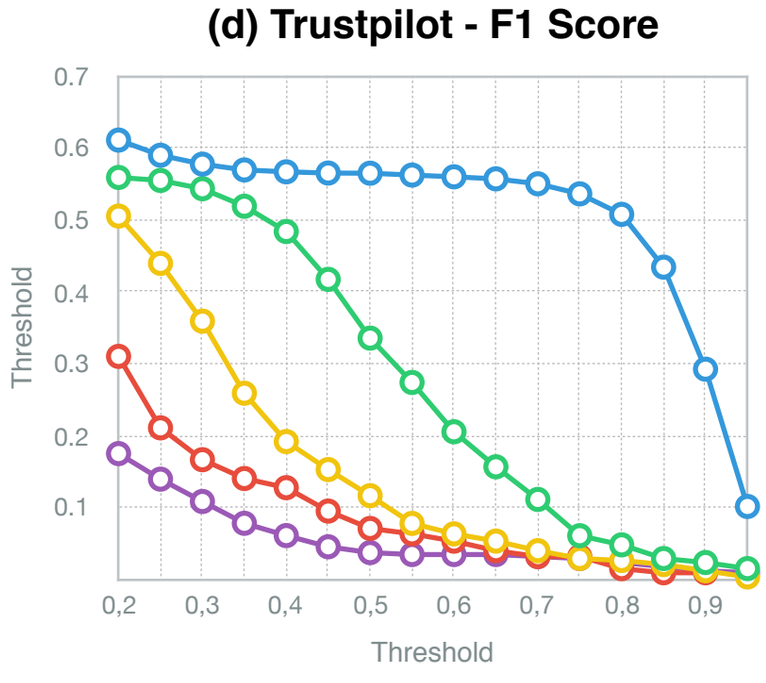
Bag-of-words LDA model results

Yelp/Trustpilot - classifier performance for IR similarity with bag-of-words LDA









IR10 IR30 IR50 IR70 IR100

- topics ∈ {10 - 100}
- #30-P>70%
- topics↑⇒P↓
- topics↑⇒F1↓
- Trustpilot reviews are much shorter
- Everybody kind of talks about the same aspects

Bag-of-opinion-phrases LDA model results

Yelp - classifier performance for IR similarity with bag-of-opinion-phrases LDA

- Yelp - smoother precision increase as both #topics and threshold ↑
- Trustpilot - poor results due to reviews length and topic sparseness and smaller dataset
- (aspect,sentiment) predict same author better

Key points

- Singleton review spammers detection using two new methods
- Yelp(57K), Trustpilot(9K), Ott(800) datasets
- Semantic similarity with WordNet => can outperform the vectorial-based measures
- Topic modeling with LDA using new bag-of-opinion-phrases approach
- Shape of reviews in Ott dataset => semantic similarity shows a more distinctive gap
- Comparison with cosine similarity and variations

THANK YOU

questions?