- slide 3: Current research on detecting fake review has taken two main directions: behavioral features and text analysis. Behavioral features are the things in the red rectangles, like review date, review stars, number of friends of reviewer and so on. Textual analysis refers to finding linguistic cues of deceptive writing. Most studies used cosine similarity, but some claimed that using more personal pronouns is more dominant in fake reviews; the same is the case with words like “husband” and “vacation”. While behavioral features work well with “elite” users, who write more than 1 review, my personal experience at Trustpilot as well as an important study of resellerratings reviews has shown that close to 90% of users only write one review. So what about the singleton fake reviewers?

- slide 4: The hypothesis is semantic similarity measures should outperform vector based models because they should also capture more subtle deception behavior, meaning more paraphrase intent of the spammers. This would inherently also work on users who operate in groups, know one other and paraphrase or rephrase each other’s reviews. Besides proposing a solution to detect opinion spam of one-time reviewers using semantic similarity, the study also proposes to use topic modeling and in particular Latent Dirichlet Allocation (LDA), a technique used lately to extract product aspects from review. Another goal was to test this hypothesis on real-life reviews and make a comparison with the existing vectorial similarity models, which are based on cosine similarity.

- slide 5: For the semantic similarity approach we used Wordnet. For those of you who are not familiar with Wordnet, imagine a large word graph where similar words are closer to each other in terms of numeric distances. If you look at the neighborhood of the word transport, it is closer to move than it is to shipping.

- slide 6: I guess everyone is familiar with the cosine similarity formulation so I won’t explain in, I’ll only describe the semantic similarity measure Given a metric for word-to-word similarity(WordNet) and a measure of word specificity(idf)(no. of documents in corpus / no. of documents containing the word), we define the semantic similarity of two text segments T1 and T2 using a metric that combines the semantic similarities of each text segment in turn with respect to the other text segment. First, for each word w in the segment T1 we try to identify the word in the segment T2 that has the highest semantic similarity (maxSim(w,T2)). Next, the same process is applied to determine the most similar word in T1 starting with words in T2. The word similarities are then weighted with the corresponding word specificity, summed up, and normalized with the length of each text segment. Finally the resulting similarity scores are combined using a simple average.

- slide 7: Aspect mining is a new opinion mining technique used to extract product attributes, also called aspects from reviews. If you look at the review, you can see the <aspect, sentiment> pairs, also called opinion phrases. Things like <hotel, charming>, <staff, courteous>.Topic models are statistical models where each document is seen as a mixture of latent topics, each of the topics contributing with certain proportions to the document. So the similarity between two reviews can be translated to the similarity between their topics distributions.

- slide 8: This is the simple LDA graphical model representation which I won't cover now. The Kullback-Leibner (KL) measures the difference between two probability distributions P and Q as shown in equation. So it can be used to compute a value for the distance between the underlying topics distributions of two reviews. This measure has two drawbacks though. If Q(i) is zero, then the measure is undefined. It is also not symmetric, meaning the divergence from P to Q is not the same as that from Q to P. Translating this to the reviews context, it is not a suitable metric to use, because if a review R_1 is similar to R_2 then it would be expected that R_2 is similar with the same amount to R_1. The Jensen-Shannon (JS) measure is based on the KL divergence and it addresses these drawbacks: it is symmetric and always provides a finite value. It is also bounded by 1, which is more useful when comparing a similarity value for a review pair with a fixed threshold in order to classify the reviews as fake. We used the information radius measure which is basically the JS measure rescaled to be more suitable for topics similarity. Think of it of simply square root of the Jensen-Shannon value.

- slide 9: We crawled Yelp and built a dataset of 57K reviews from 660 New York restaurants and considered Yelp's recommended reviews (unfiltered) as truthful and the not recommended (filtered) ones as spam. Several well known studies have considered Yelp's filtered reviews as fake and unfiltered ones as truthful. The Trustpilot dataset of 9K labeled English reviews was kindly shared with us by the company. It contains 4 and 5 star reviews from 130 businesses, from one-time reviewers only. The company has been filtering away fake reviews for several years now, so we have assumed their detection mechanisms provide fairly good results. The Ott dataset contains 800 reviews, balanced between truthful and fake and is publicly available. The dataset was created through AMT crowdsourcing, by soliciting participants to pretend working in the marketing department of hotels and write fake reviews for their employers. The authors allowed only one submission and rejected short, illegible or plagiarized reviews.slide

- slide 10: We did some standard preprocessing of the review text, removed stop words, extracted nouns, verbs and adjectives and also lemmatized the extracted POSs. We used 2 extra variations of cosine similarity, one where we considered only nouns, verbs and adjectives and another where we added lemmatization to these type of words. The pairwise similarity model is very simple, for every review pair, compute the similarity and record which results go over a certain threshold. Over the threshold means both reviews are fake, otherwise they are truthful.

- slide 11: Alright, some results now. You can see the results for both Yelp and Trustpilot reviews, in terms of vectorial and semantic similarity, left plots show the precision, right plots show the F1 score. The semantic similarity measure is shown in black. The semantic similarity is close but below the vectorial ones for thresholds below 0.7, but it reached 90% at a threshold of 0.8. This happens for both datasets. You can see the F1 score is generally higher for Trustpilot than for Yelp and one possible explanation might be that the opinion spammers targeting Trustpilot are not that professional. They do not make the effort to write more elaborate reviews, mimicking the honest reviewers writing styles. They seem much more prone to reuse the same exact words or synonyms when writing new reviews.

- slide 12: Next we tried to see if there is a distributional difference, for both vectorial and semantic similarity measures, between deceptive and truthful reviews inside the in the well known dataset used by Ott. We plotted the cumulative distribution function for both truthful and fake reviews. The purpose of the CDF curves of truthful/fake reviews is to check if they overlap or there is a separation margin between the two curves. There are more details about this experiment in the paper, so what I will stop to show on this slide is there appears to be a larger gap between the two reviews types when looking at the semantic plot.

- slide 13: Here, you can see the results for the bag-of-words topic modeling approach using 10, 30, 50, 70 and 100 topics. You can see that as when the number of topics increases, the classifier precision decreases, which is a bit non-intuitive. Best overall performance was achieved for 30 topics, but the conclusion we got from this experiment was that using bag-of-words LDA doesn't perform well mainly because the reviewers (honest or fake) basically talk about the same aspects, and it's really hard to distinguish just from the topic distribution similarity.

- slide 14: For the bag-of-words approach, the classifier performed worse the more topics were used. However, for opinion phrases, using the Yelp dataset, there seems to be a smooth increase in performance as the similarity threshold increases, coupled with an increase in the number of topics. Intuitively, it makes more sense, since increasing the number of topics should create a better topics separation using opinion phrases. This causes reviews which mention the same aspects and sentiments to score higher in terms of their topic distributions similarity. The model performed really bad on the Trustpilot dataset, giving more or less a flat precision regardless of the number of topics, therefore we did not plot the results, even in the paper. The poor performance could be a consequence of the dataset being much smaller than Yelp and plus, Trustpilot reviews are generally much shorter. Also opinion phrases induce topic sparseness even more than individual words. The precision for 100 topics reaches 65% for a 0.85 similarity threshold .This definitely shows more promise than the bag-of-words approach.

- slide 15: So, a quick wrap-up of the things I presented.