Related Readings

Note: While we didn't end up including a related works / review section in our final paper, we did do reading about the general practice of fraud detection systems and these are just notes we extracted about the more relevant or interesting parts we noticed in papers on the topic.

Not sure if this is really worth submitting or not, we took the notes so we figured it can't hurt to throw this in with the rest of our work as it guides what we've done and where we hope to go with this, and reflects our engagement with the material.

Nico & McCoy

Apparently this man is a bit of an NLP guru: https://myleott.com/

(not an academic paper) -- Textblob Sentiment - Calculating Polarity & Subjectivity
word2vec illustrated

Fake reviews detection based on LDA

- we only have data of reviews themselves, so we can't really work on reviewer behavior modeling like other research has
- *Snehasish Banerjee and Alton YK Chua extract linguistic features to distinguish fake reviews by* ***word n-gram, psycholinguistic deception words, part-of-speech distributions****, ~~readability of reviews and review writing style~~*
    - look into what psycholinguistic features are
- *Heydari focuses on systematically analyzing and categorizing models that detect review spam*
- using an 80-20 split
    - SVM was their worst, LDA+SVM slightly better,
    - **best accuracy was from both LDA+LogisticReg and LDA+MLP (81.3%)**
        - LDA had slightly higher F1 score than MLP
- "discussion" -- *"LDA can extract topic-words from one document, and to some extent, topic-words can represent whole document. Thus, we use LDA to respectively extract topic-words from fake reviews and non-fake reviews, it is more reflected the features of fake or non-fake reviews. Then when we counts the term frequency of each word, the import words to reflect the features of fake or non-fake reviews will have a higher term frequency, and then increase the accuracy of classification models. But due to the quantity of data is enormous, the quantity of topic-words is far less than that. Therefore, the accuracy with LDA slightly higher than the accuracy without LDA."*

Detection of review spam: A survey

- positive/negative fake review, different intentionality potential

- things we don't have that other literature suggests useful (can put in final report as potential areas for theoretical improvement // current limitations)
    - metadata (IP, MAC address, location, etc.)
    - account behavior (posting negatively about tons of other brands?)
    - text body of item being reviewed (we only have title)
- spammers && duplicate detection
    - often times, more amateur // cheaper fake reviewers will use templates, so when you shingle and compare for similarity, it can create situations where their templates are more easily detectable
    - cosine similarity another metric used
    - issue of when spammers copy legit reviews and template out of those
- N=2 shingling and logistic regression seems like a solid option
- "*In addition, a probabilistic language model was proposed in Lai, Xu, Lau, Li and Li (2010) that generates a similarity score between two reviews. The model calculates the likelihood that one review was generated from another. The model compares a pair of reviews using Kullback–Leibler divergence (a measure used to estimate the distance between two probability distributions) to detect content similarity. To convert the Kullback–Leibler divergence measure to a spam score for each review, they used a linear normalization method. Finally, they used SVM for text categorization to classify spam and genuine reviews.*"
- 4.1.2 content-based methods
    - genre identification
        - refers to two papers, POS distributions, a paper by **ott**
    - psycholinguistic deception
        - "*Psycholinguistic deception detection is a technique to assign psycholinguistic meanings to the keywords used in a text. A well-known text analysis instrument is the Linguistic Inquiry and Word Count (**LIWC**) software (Pennebaker, Chung, Ireland, Gonzales & Booth, 2007), which assigns 80 psycholinguistic meanings to 4500 keywords.*"
    - text categorization
        - *"Based on a literature review, the authors assumed that readability of a review (complexity and reading difficulty), review genre (distribution of POS tags) and review writing style (positive cues, perceptual words, and future tense) are different between deceptive and genuine reviews."*
        - a lot of really cool shit we can't use for 671 realistically in this subsection
        - "The proposed system would examine a review text with the rating it gives using five defined criteria: ~~rating consistency~~, **questions in review**, **all capital letters review**, **comparative sentences**, **link spamming**."
            - refers to a model that did really well

Negative Deceptive Opinion Spam (ott 2013)

- "*While previous related work (Ott et al., 2011; Ott et al., 2012) has explored characteristics of positive deceptive opinion spam, the complementary problem of negative deceptive opinion spam remains largely unstudied*"
- "*However, because the truthful reviews are on average longer than our deceptive reviews, we sample the truthful reviews according to a log-normal distribution fit to the lengths of our deceptive reviews, similarly to Ott et al. (2011)*"
- **4) Interaction of Sentiment and Deception**
    - "[*...], fake positive reviews included less spatial language (e.g., floor, small, location, etc.) because individuals who had not actually experienced the hotel simply had less spatial detail available for their review (Johnson and Raye, 1981). This was also the case for our negative reviews, with less spatial language observed for fake negative reviews relative to truthful.*"
    - *"Likewise, our fake negative reviews had more verbs relative to nouns than truthful, suggesting a more narrative style that is indicative of imaginative writing (Biber et al., 1999; Rayson et al., 2001), a pattern also observed by Ott et al. (2011)"*
    - **differences between fake positive and fake negative**
    - observed difference from positive fake reviews: negative ones tend to overdo it on the negative words
    - *"[...], while first person singular pronouns were produced more frequently in fake reviews than truthful, consistent with the case for positive reviews, the increase was diminished in the negative reviews examined here."*

Journal of Big Data -- Survey of review spam detection using machine learning techniques [2015]

- crtl+f 'ott' #6 result -- discusses their work as a broader overview
    - his major NLP hotel review fraud study was **also a balanced 50-50 fake real set**, 800 reviews total, 400 legit, 400 illegit
        - best results -- 89.8% -- with SVM trained on bigrams and LIWC features
        - human judges averaged about 61% accuracy as a baseline
    - the second study, on just detecting fraudulent negative reviews, they use **twice the data**, but humans scored about 65% and the model still topped out at 88.4%
        - "This suggests that separating spam review detection into positive sentiment spam review detection and negative sentiment spam review detection is beneficial."
- li et al used SAGE // sparse additive generative model // basically a blended method and applied it to a corpus of real//fake reviews **across sectors** and discovered that intra-domain classification had different optimal approaches
    - *"This indicates that different linguistic features may appear in different domains, and more robust cues of deceptive opinion spam need to be identified if a cross domain classifier is to be created. Of note was that the classifier exhibited particular difficulty when trained using the restaurant and hotel reviews and evaluated against the doctor reviews. Using SAGE, accuracies of 64.7 and 63.4 % were achieved using LIWC and POS tag features respectively but only 52.0 % when using Unigram features."*

Jay Kumar's Dissertation -- Fake Review Detection Using Behavioral and Contextual Features

- "(Jindal & Liu, 2008, 2007b). They considered 5.8 million reviews from **Amazon** [1] and used feature from product and reviewer meta-data on four categories of product to identify

fake reviews. The research work comprises identification of untruthful reviews, brand reviews, non-reviews and spammer groups. They discovered spamming activities including identifying duplicate or near duplicate reviews using shingle method. Untruthful reviews were identified by calculating content similarity between all reviews of a reviewer to highlight duplicate reviews. For identifying brand reviews and non-reviews, dissimilarity between product meta data and review content were used. Spammer groups were identified by calculating content similarity of reviews of different reviewers."
    - differentiates categorically 4 different types of spam, interesting
- *"[...] AMT Turkers are not good at faking a review. The reason is that AMT Turkers have limited knowledge about the domain. Word distribution of posted reviews by AMT turkers' is different from true reviewer."*
- *"Various behavioral features related with reviews and reviewer were exploited by (D. Zhang et al., 2016) . The importance of selected features were also investigated for identification of fake reviews. They exploited features including 24 behavioral and 16 contextual features. Exploration of behavioral features were based on Interpersonal Deception Theory (IDT) which posits that "deceivers display both strategic behaviors (e.g., information manipulation) and nonstrategic behaviors during deception" (Buller & Burgoon, 1996). Experiments were conducted on review dataset of Yelp. Reported classification results show 87.8% accuracy using all features. Highly correlated features were identified using Pearson Correlation before feature pruning. Twelve most important features were identified after feature pruning to achieve 90% accuracy by training RF classifier. Moreover, accuracy of SVM, NB, Decision Tree and RF were also compared."*

4.2.1 Text Preprocessing

Text preprocessing include data mining techniques used to transform unstructured text. Few text preprocessing techniques on our selected dataset are defined as follows:

- **Tokenization:** Tokenization is task of splitting-up the review text into words (tokens). i.e. Review content is tokenized into tokens. For calculating RCS and capital diversity, tokenization is vital step to separate each word in review.
- **Lemmatization:** The task of lemmatizer is to transform word with respect to morphological root word e.g. 'bought' lemmatized into 'buy'.
    - **interesting new features mentioned here (seemingly not super useful in Yelp's case, though)**
        - capital diversity (how many tokens beginning with capital)
        - noun ratio