Mimicking human fake review detection on Trustpilot

[DTU Compute, special course, 2015]

Ulf Aslak Jensen
Master student, DTU
Copenhagen, Denmark

Ole Winther
Associate professor, DTU
Copenhagen, Denmark

Rasmus Hentze
Product manager, Trustpilot
Copenhagen, Denmark

1. INTRODUCTION

Trustpilot is an open review platform that facilitates trust and transparency between consumers and businesses, by allowing users to provide public feedback on experiences with companies. Their service has aggregated approximately 13 million reviews on over 100,000 businesses across 65 countries[3]. Due to their rapid growth, fraud detection is becoming an increasingly important task calling for a high degree of automation in order to secure the long-term validity of their value proposition. At present the methods used for removing fake reviews requires a high degree of human aid, and are not scalable.

Defining which kinds of behavior classify as fraud on the Trustpilot platform is still open for discussion and is continuously reevaluated within the company. On a very general level and with some overlap, fraud experienced by Trustpilot can be sectioned into three categories.

1. Singleton deceptive review spamming.
2. Professional deceptive review spamming.
3. "Cherry Picking".

Category 1 spans every type of single instance deceptive reviewing. This could e.g. be cases where company associates review their own company positively or their competitors negatively. While it violates the usage guidelines[5], it is hard to identify and often requires asking the user for proof of purchase documentation (POP). Previous efforts to address this problem involves analysis of temporal patterns[14] and semantic similarities in review text[12]. Category 2 contains reviews placed on a company by a third party, paid by either the company itself or a competitor. This is considered the most damaging category of review spam and is frequently discussed in related literature[10, 8, 7, 9]. Category 3 contains genuine reviews placed by real customers, that were unknowingly "cherry picked" by a company to provide feedback after a successful transaction. This creates a skewed distribution of opinions and is against the guidelines[4]. Catching this involves analysing the time series data for each company in order to identify bursts of irregularly positive reviews. Research aiming at identifying spam by considering bursts in time-series data has been conducted[6], none, however, directly addressing the issue of "Cherry Picking".

There is a huge diversity in the types of problems addressed in literature considering spam detection. By narrowing the scope to only consider review spam the diversity decreases, but retains significance due to the nature of the different reviewing platforms. Trustpilot remains unique due to its focus on company reviews with its closest neighbors being services like Yelp.com and Google Reviews that focus on local businesses. Most studies concerning review fraud attempt to classify individual reviews as fraudulent based on either linguistic features[11] or behavioral features[7], while some consider groups of reviews[10] e.g. by individual users or directed towards specific brands/items.

This work addresses the challenge of catching category 2 fraudulent reviews using machine learning classification on a dataset labeled by employees hired to filter reviews, hence the title. It relates to the research conducted by V. Sandulescu in 2014[13], which considers detection offake Trustpilot reviews in the same category using semantic similarity in review-text, based on topic modeling. In contrast, this study considers only behavioral features as it has been argued that they, in many cases, work better than linguistic features for predicting review authenticity[7].

The goal of this project is first of all to explore what opportunities exist for automation in review detection. Secondly, no previous data-driven efforts have been made to understand the behavioral patterns in Trustpilot reviews, and as such, one may consider this report a "first look".

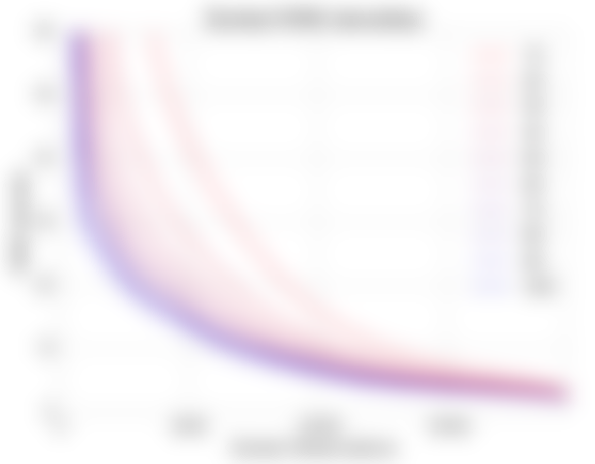
The work is presented very much in order of execution. It is written as an industrial research report which puts the strongest emphasis on analysis. It first offers a thorough explanation of the data used, such as to promote a clear understanding of the basis of the results. The analysis considers outliers and data point distribution using K-Nearest Neighbor (KNN) density estimation and Principle Component Analysis (PCA), and compares a number of different classifiers using Receiver Operating Characteristic (ROC) curves and Precision-Recall (PR) curves. Based on findings in the analysis a few of the most important features are investigated. The discussion evaluates the validity of the results and regards problems in the data as well as issues associated with scalability. A few comments are also attached to aspects of preprocessing and analysis which could have been done differently. After the conclusion a number of suggestions for future work are made, based on findings in this report as well as ideas that were picked up along the way. It is the hope of this author that the results of this study may be of value to future research addressing the problem of catching fake reviews on Trustpilot.

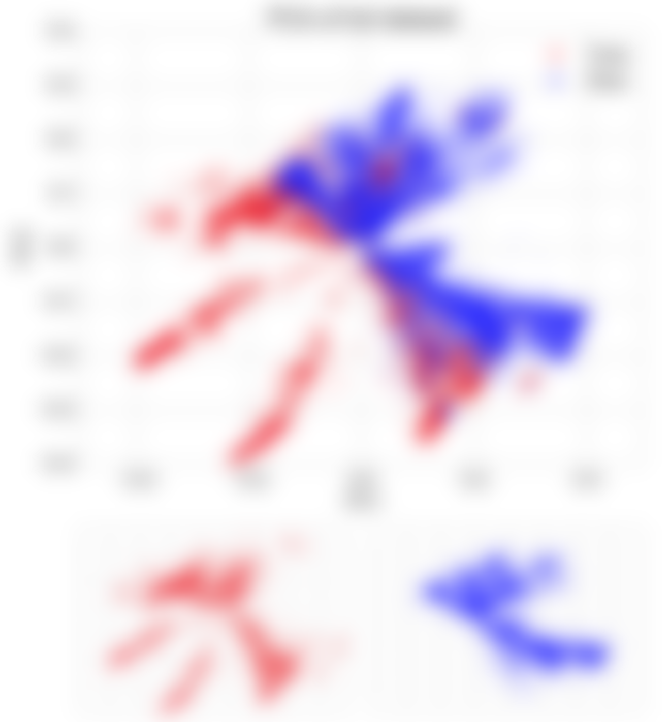
3 Analytics

Figure 1: Programmatic workflow. [illegible]

3.1 Labelling

3.2 Programmatic





5. DISCUSSION

When achieving such a high classification accuracy, skepticism is very important. Sec. 4.4 serves as an important sanity check, and makes clear indications that a number of the features in the dataset are highly correlated with the target variable, which explains the high accuracy. To fully understand how it comes to be that observations in a dataset can be so easily partitioned, it is wise to consider the labeling process. Q&C agents consider suspicious reviews and recognize patterns that they have been trained to know indicate fraud. While performing this repetitive task, one can easily imagine that it becomes attractive for even highly skilled employees to generalise on certain parameters and ignore others for most reviews, for reasons such as time pressure and various kinds of bias. Recall now that the classifier does not actually classify user behavior; realistically speak-

ing it classifies Q&C agent behavior, and finding this to be highly predictable is arguable not that exceptional. Albeit so, the obtained results still mean that the kind of human fraud detection investigated in this report can be accurately mimicked, and inspires a future where the development of automated fraud detection tools is incrementally designed, such that humans first define the patterns that characterize fraud in a certain category (or subcategory) through manual labeling, after which machines learn the patterns using the data which has been generated in the labeling process.

A general limit to the findings of this study is unarguably the capabilities of the humans that labeled the data. As pointed out in Sec. 4.4, a number of features that one would expect to be highly significant was not attributed any significance by the classifier. Investigation into the labeling process reveals that Q&C agents are in fact not able to even consider them due to the design of the interface they use. This is hugely problematic, but serves to emphasize the point that a classifier which attempts to score high accuracy in a dataset labeled by humans, will never become more accurate than the humans that labeled it.

The issue of model robustness may become an issue as time passes and Trustpilot scales. Right now the model relies on features that may not be available or simply mean something else in the future. For example the model discriminates against users from Bangladesh, but it may well be that the genuine user base in this part of the world increases at some point, leaving the model obsolete. A solution to this could be to periodically train the classification model based on rolling month data gathered by Q&C agent working together with the classifier. This however raises an issue in terms of scalability, as it requires perpetual human supervision even in the long run. Consequently this calls for investigations into which general and timeless features define fraud.

In regards to problems in the analysis there are a number of things which could have been done differently, had time restriction not been an issue. Feature selection could have been done more gently, e.g. by using variance thresholding more conservatively and performing backwards/forwards feature selection on a larger subset. This would likely have done a better job at reducing the amount of correlated features and allowed for more informative feature analysis.

6. CONCLUSION

It has been shown that it is possible to create classifiers that mimic the decision making process of humans searching for fake reviews on the Trustpilot site, with an accuracy of over 96%. The Random Forest classification model performs best in terms of accuracy, AUC (for ROC and PR curves) as well as computational demands measured in testing time, but the Logistic Regression classifier performs extremely well too and offers easy deployment on Amazon Machine Learning. Several traits in the data were highlighted such as high point similarity (Sec. 4.1) and spacial clustering into respective classes (Sec. 4.2). Feature analysis (Sec. 4.4) reveals high correlation between the classifier's most important features and the target variable, explaining the high accuracy. A number of important arguments, relating to the validity of the findings as well as model robustness and scalability, were established in Sec. 5.

The ability to classify reviews automatically only solves a small fraction of the general problem with fake reviews. As pointed out in Sec. 1 many categories (and arguable subcategories) of review fraud exists and in that respect this study serves as a useful stepping stone giving light to previously unexplored aspects of the problem.

7. FUTURE PERSPECTIVES

An important field of research which needs to be addressed is that of gaining a better understanding of what constitutes fraud. This would e.g. require use of unsupervised classification in order to discover new classes in the data, or for that matter a better dialogue between the engineers attempting to automate fraud detection and the Q&C agents doing it manually. A proper feature exploration study to determine the true indicators of fraud in a given category would unarguable be extremely valuable, initially to the engineers designing the interface that the Q&C agents use and in turn to the company as a whole.

Recalling the work of V. Sandulescu described in Sec. 1[12], the effectiveness of a classifier using both behavioral features and linguistic features remains unexplored and should be investigated, to improve the robustness of the classifier.

8. REFERENCES

[1] Amazon machine learning developer guide. `http://docs.aws.amazon.com/machine-learning/latest/dg/transforming_data.html`. Accessed: 2015-26-06.

[2] Scikit-learn website. `http://scikit-learn.org/stable/`. Accessed: 2015-25-06.

[3] Trustpilot business website. `http://business.trustpilot.com/`. Accessed: 2015-05-06.

[4] Trustpilot company usage guidelines, dec 2013. `http://legal.trustpilot.com/company-guidelines`. Accessed: 2015-25-06.

[5] Trustpilot user usage guidelines, may 2014. `http://legal.trustpilot.com/user-guidelines`. Accessed: 2015-25-06.

[6] Exploiting burstiness in reviews for review spammer detection. *Proceedings of the 7th International Conference on Weblogs and Social Media, Icwsm 2013, Int. Conf. Weblogs Soc. Media, Icwsm*, pages 175–184, 2013.

[7] What yelp fake review filter might be doing? *Proceedings of the 7th International Conference on Weblogs and Social Media, Icwsm 2013, Int. Conf. Weblogs Soc. Media, Icwsm*, pages 409–418, 2013.

[8] N. Jindal and B. Liu. Opinion spam and analysis. 2014.

[9] F. Li, M. Huang, Y. Yang, and X. Zhu. Learning to identify review spam. *Ijcai International Joint Conference on Artificial Intelligence, Ijcai Int. Joint Conf. Artif. Intell*, pages 2488–2493, 2011.

[10] A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal. Detecting group review spam. 2011.

[11] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. 2012.

[12] V. Sandulescu and M. Ester. Detecting singleton review spammers using semantic similarity. In *Proceedings of the 24th International Conference on World Wide Web Companion*, WWW '15 Companion, pages 971–976, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[13] V. O. Sandulescu. Opinion spam detection through semantic similarity, 2014.

[14] S. Xie, G. Wang, S. Lin, and P. S. Yu. Review spam detection via temporal pattern discovery. *Proceedings of the Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, Proc. Acm Sigkdd Int. Conf. Knowl. Discov. Data Min*, pages 823–831, 2012.

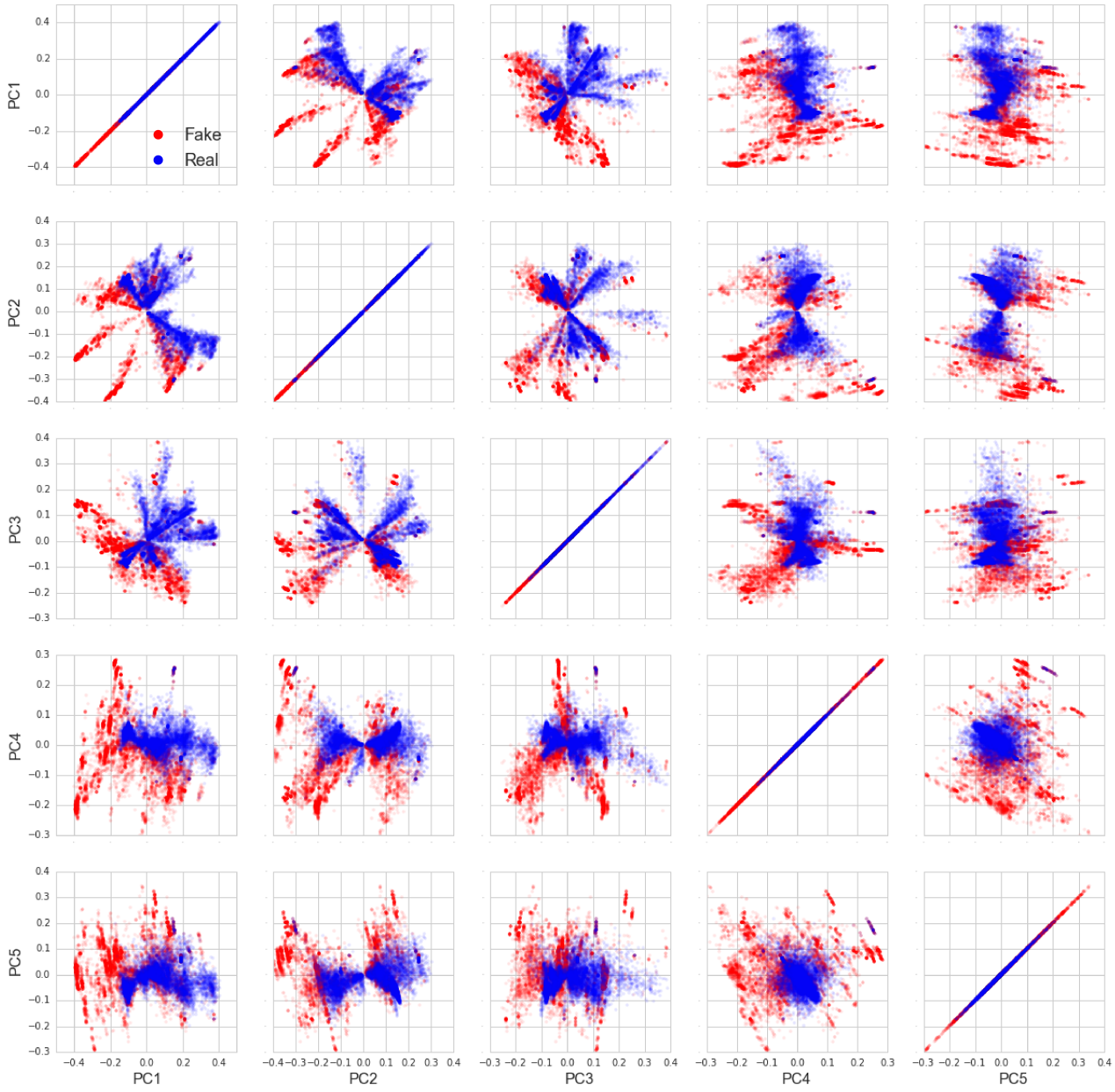
APPENDIX

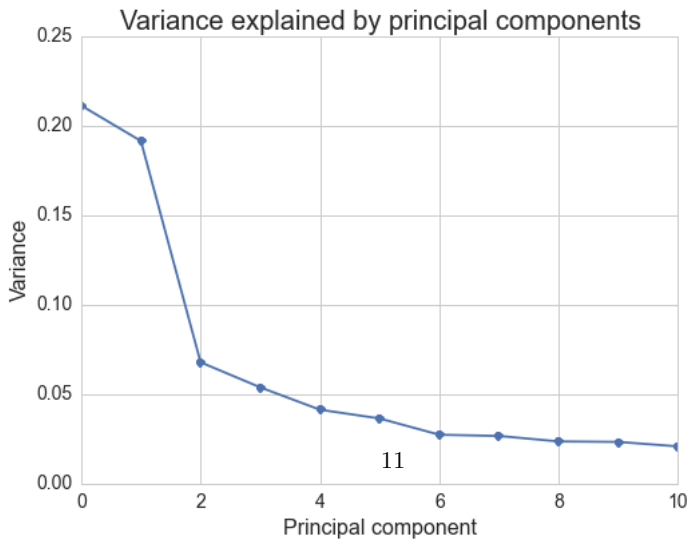
A. USERMETRICS RAW DATASET SAMPLE

B. USERMETRICSPROCESSED RAW DATASET SAMPLE

C. FEATURES

D. PRINCIPLE COMPONENT ANALYSIS





E. PC POINTS INTERPRETATION

F. FEATURE IMPORTANCE

G. AMAZON MACHINE LEARNING