NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

IT3010 – Assignment 4: Revised Paper

RESEARCH-BASED INNOVATION METHODOLOGIES IN COMPUTER AND INFORMATION SCIENCE

Comparison of humans' and machine learning algorithms' accuracy in differentiating fake reviews from real ones

Group members
Asma FAROOQ
Trym GRANDE
Mathias HEGGELUND
Eirik PLAHTE
Anthony VU

Group number: 18

Spring 2022
Word count: 2 890

https://www.ntnu.edu/studies/courses/IT3010

NTNU
Norwegian University of
Science and Technology

Comparison of humans' and machine learning algorithms' accuracy in differentiating fake reviews from real ones

Norwegian University of Science and Technology
Department of Computer Science

Trondheim, Norway

Asma Farooq asmaf@stud.ntnu.no
Trym Grande trymg@stud.ntnu.no
Mathias Heggelund mathiaoh@stud.ntnu.no
Eirik Plahte eirikpl@stud.ntnu.no
Anthony Vu thvu@stud.ntnu.no

***Abstract*—Companies seek reviews from their customers because it serves as a feedback mechanism for them to improve their product or service quality and customer relationship. These reviews can help customers as well, because they often anticipate and evaluate products and services' quality based on reviews. Positive reviews help to increase the trust of customers. So, it can be said that many businesses flourish based on positive reviews. Some companies try to create false impressions by purchasing fake reviews of their products or services and evidently, they succeed in gaining customers. This scam can be prevented by detecting fake reviews. This research is conducted to determine the accuracy of humans in detecting fake reviews compared to machine learning algorithms. To do this we used the survey research strategy as our data generation method. We used a questionnaire containing real and fake reviews, and quantitatively analyzed the results and compared them with previously done studies. Our analysis shows that the mean human accuracy of detecting fake reviews is at 49%, considerably lower than the machine learning algorithm having an accuracy of 98%. This is a significant gap, and it indicates that there might be a need to apply these algorithms to real-life situations.**

***Index Terms*—machine learning algorithm, fake reviews, authentic reviews, accuracy**

I. Introduction

User reviews of products being sold online have a significant influence on customers' willingness to purchase said products. According to J. Pitman [1], who published a paper earlier this year, 93% of consumers say that online reviews influenced their purchase decisions, and 91% of 18-34 years old trust online reviews as much as personal recommendations. However, not all online reviews are genuine. Reviews can be purchased from large private groups on Facebook and other sites [2] and can be used to artificially boost a product's perceived quality. This also applies to app stores, where an illegal market for fake reviews has emerged [3] which threatens the platform's integrity and can mislead and manipulate users. To counteract this, it is important for people to be able to discern fake reviews from real ones to the greatest possible extent. Machine learning algorithms being trained on large data sets containing thousands of app store reviews (both real and fake) have achieved high accuracy in detecting fake reviews [3].The objective of our research is to compare humans' ability to detect fake app store reviews to machine learning algorithm's ability to do the same.

RQ: *How accurate are humans in detecting fake reviews compared to a machine learning algorithm?*

We will research this by analyzing data collected from a survey containing both real and fake reviews compiled in a data set previously used in another paper [3]. In the study, seven machine learning algorithms were used to train models using this data set. However, we will limit our scope to comparing human abilities to one algorithm, the Random Forest classifier. If the results show that there is a large discrepancy in the accuracy of humans compared to the accuracy of machine learning algorithms, there might be a need to apply these algorithms more towards consumers, so that they can reap the benefits of not having their purchases being influenced by fake reviews.

II. Background

A. Fake review concept

Reviews are an important aspect of an app, since it allows for user feedback to developers for further improvements of their application. In addition, reviews are used to share customer opinions and experiences with a product and has been shown in recent studies to have the ability to highly affect willingness to buy or use a product. A review that has been given by a user without an actual experience by the user is considered a fake review. Fake reviews can be divided into three categories: untruthful reviews, which consist of false information to either affect the product negatively or positively; branding, which focuses on branding the product without giving an experience review of a product; and non-reviews, which are not directly related to a product but rather as advertisement. This has created an illegal market to buy fake reviews to affect products negatively or positively [2]. These

fake reviews can either be human or computer-generated. Purchasing fake reviews is highly discouraged for developers in the App store and is strictly prohibited. Apple states: "If you attempt to cheat the system (for example, by trying to trick the review process, steal user data, copy another developer's work, manipulate ratings or App Store discovery) your apps will be removed from the store and you will be expelled from the Apple Developer Program." [4] To protect customers from the scam of fake reviews, identification of fake reviews can be helpful.

B. Fake review detection

There are three important variables that are used for assessing an ML algorithm's accuracy in detecting fake reviews. The first is precision, which is the proportion of guesses for fake that are correct. The second is recall, which is the proportion of the fake reviews that were correctly classified as such. F-score is a representation of an overall accuracy by calculating the harmonic mean between the two former two values.

As for prior research on the topic, Wang et al. [5] merged multi-feature fusion and rolling collaborative training for detecting fake reviews. In addition to being identified by key features, reviews can also be identified by the behavior of reviewers. Extraction of behaviours by features engineering is the subject in the paper *Fake Reviews Detection using Supervised Machine Learning* [6]. This study compared a Yelp data set with and without feature extraction with respect to user behavior. The performance was compared with K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine, Logistic Regression and Random Forest. It concluded that KNN works best among chosen classifiers while considering behavioural features of reviewers. In a study [7], Barbado et al. proposed a framework named Fake Feature Framework(F3). They have classified features into user centric and review centric, where user centric features study the behavior of the user in social networks and review centric features study the text of the review. They concluded that the proposed framework has a 82% F-score by using Random Forest or Ada Boost algorithms and user centric give finer results than review centric features. Research by Salminen et al.[8] has shown that machine learning algorithms are highly efficient for fake reviews detection with good precision. Amazon´s data set has been used in the research to train their text generation model to generate fake reviews, which they have tested on different machine learning algorithms, achieving 98% accuracy at highest compared to humans achieving 57% using a sample of their generated data set. A language model used in text generators generally has no understanding of the text itself but will try to find common patterns from the data set and mimic them to the greatest extent possible. This could lead to a bias towards the machine learning algorithms that were tested as they could utilize and discover these patterns better than humans yielding a higher accuracy than humans. We have found no study that tested the accuracy of humans in detecting fake reviews using a data set that is not computer-generated. Therefore, we will try to fill this research gap by doing so, using a data set provided by Martens et al. [3], removing the factor where results could be affected by bias. In addition, a gap has been addressed in Salminen et al.'s paper [8] where they claim that different review platforms (e.g., Amazon, Apple Store and Yelp) and their data sets should be further examined due to how reviews can be given related to its platform.

III. Methods

A. Data generation

This study uses a survey as the research strategy and a questionnaire as the primary data generation method. The survey research strategy was chosen for its allowance of collecting an adequate amount of data per the time-frame for the research. An experiment where the researchers had tested out one or several different hypothesis related to review classification by humans, by observing users in real-world situations instead could have been an alternative as a research strategy. But this would have limited the amount of data that one would have been able to collect because of a much more extensive and time-consuming work. The data set used to sample reviews for the survey was the same as used by Martens et al. [3]

To be able to do some proper analysis and answer the research objective and question, data from a larger amount of people was seen beneficial. This was the main reason for the choice of a questionnaire as the data generation method. The questionnaire allowed for the collection of adequate amounts of quantitative data from a larger group of people in an organized and low-cost manner. Compared to other alternative data generation methods such as observations or interviews, a questionnaire was more feasible to do within the time-frame for the project. These strategies could have generated high quality and valuable data. However in the scope of the research and the time-frame for this study, not enough data could have been collected to properly answer the research question and objective. Therefore, the survey strategy with the questionnaire data generation method was therefore seen as the best alternative and chosen.

The target audience for this study were people with some experience using a smartphone and their related app store, such as the App Store on iPhone or Google Play Store on Android phones, as this is where the origin of the reviews in the questionnaire is. The reason for selecting this audience is that they would have some context as to what a mobile app is, how they work, and therefore can relate to the reviews and assess their legitimacy even though what app is written for is unknown.

B. Data analysis

The data collected from the questionnaire, along with the result from the previous study are visualised with graphs/tables. By analyzing this data quantitatively, it will be possible to determine trends and correlation in data points in the generated data and between the generated data and the previous results we compare with. This analysis will help determine to

what extent humans compare to artificial intelligence (AI) in detecting and differentiating fake reviews from real ones.

C. Questionnaire

The questionnaire was designed to generate quantitative data with the participants' opinions. The questionnaire consisted of review itself, title, username and rating with only two answers, 'fake' or 'real', and asks the participant to evaluate the presented review and classify it.

IV. Results

The results have been exported from the questionnaire created on Nettskjema. This is referenced along with a report [9]

The score distribution for each individual question shown in figure 1, have a total of 41 individual submissions [10]. This shows that a large proportion of the questions has an even distribution between fake and real answers indicating participants tend to guess when labeling the reviews. The average classification precision achieved was 49%, while the recall was 45%. This further lead to an F-score of 47%. This is summarized in table I. As the questionnaire had only two options to choose from, true or false, the chance of randomly choosing the correct label is 50%. The results for the survey therefore place humans just below the random chance, at an accuracy of 49%.

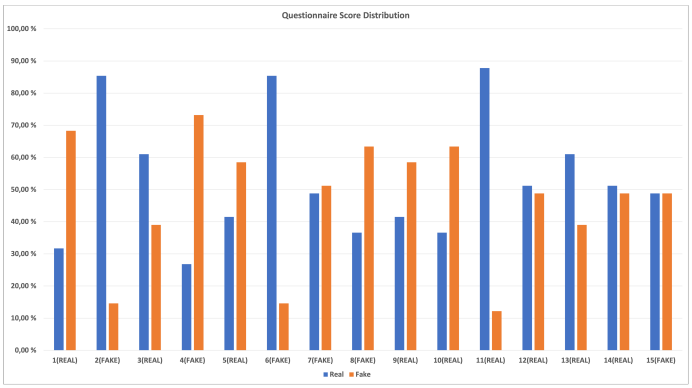

Fig. 1. Questionnaire Score distribution

Classification measure	Value
Precision/accuracy	49%
Recall	45%
F-score	47%

TABLE I
Classification results

For comparison, results from Martens and W. Maalej [3] are shown in Table II.

The precision value shows that out of the 15 App Store reviews presented to the respondents, they were able to label 7,35 of them correctly on average. Secondly, the recall tells us that on average, 45% of the fake reviews were classified as

Classification measure	Value
AUC/ROC (accuracy)	98%
Recall	91%
F-score	undefined

TABLE II
Classification results from Martens and W. Maalej

Score measure	Value (%)
Mean	49%
Median	47%
Standard deviation	13%
Average participation duration	4 minutes 24 seconds

TABLE III
Score measures

fake. The F-score lands in between these two values. This is shown in table III.

The score values have a mean of 49%, and a standard deviation calculated to 13%. This is compared to the normal distribution and shown in figure 2.

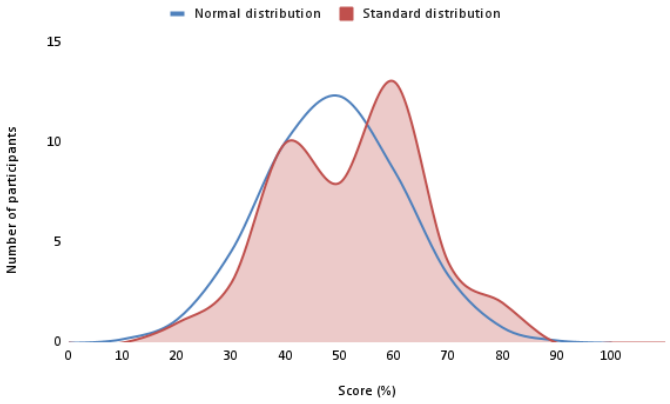

Fig. 2. Score distribution

V. Discussion

The machine learning algorithm from the paper by Martens et al. [3] that we decided to compare with reached an accuracy (AUC/ROC value) of 98%, on the data set containing the same reviews sampled from for this study. Similarly, the Random Forest classifier algorithm achieved a recall of 91%, while the respondents in this survey had a recall of 45%. In the paper three features are used to classify a review as fake or real: 1) total number of reviews the app received and 2) the user provided, as well as the 3) frequency in which the user provides reviews. Therefore, the algorithm can detect a user frequently providing reviews and classify if the user and its reviews are real or fake. A previous study by A. Mukherjee [11] shows that word distribution of fake reviews does not significantly differ from real reviews, achieving an accuracy of 67.8% using a textual classifier. Our results show that classifying a review is difficult for humans based on text, username and rating.

From the score distribution one could see that the score variation is relatively normally distributed for the respondents when comparing it to the normal distribution. This again can be linked back to a 50% random chance, as we see the highest peaks in the standard distribution slightly below and slightly over the 50% score, averaging out at 49% in the end. Additionally, we can see that the average participation duration for the survey was at around 4 and a half minutes, which arguably is a relatively normal and expected time. This means that the results will be somewhat thought through.

The research question introduced in section I raises the question of how accurate humans are in detecting fake reviews compared to machine learning algorithms. From the results presented and discussed above in this section, we can clearly see a huge difference in the accuracy of human detection and machine learning algorithms, with a difference in the accuracy of 49 percentage points. Looking further into question 4 and 11 where the distribution is more one-sided and classified correctly in 1, it can be seen that the username and rating [9] highly differ from the other questions in the survey and therefore can make it easier to classify a review correctly. This can also lead to a misclassification as seen on question 1,2 and 6 where reviews have a perceiving effect on the participants instead. This indicates that a common pattern for humans when classifying a review is to look for titles, username or rating that stands out from other reviews resulting in a one-sided distribution of fake or real regardless of being correct or not. When these standouts can´t be found, it leads to guesswork, creating an even distribution between fake or real for a review. This may be caused by how reviews are provided, as there is no general format of how reviews should be written, and will vary.

Our findings in this paper support what was previously shown in another study [8] as well. It is important to mention that the circumstances in that study were a bit different than in our case, as the paper compared humans' and machine learning's accuracy on another machine learning model's generated reviews and not reviews collected from main sources such as in our case. But apart from the origin of the reviews, the results were much the same. This study reached a human accuracy of 57,1%, slightly better than the result from our survey. This shows that previous results concerning machine generated reviews also applied to reviews collected from real applications.

VI. Conclusion

Based on the findings in the research, it is hard for humans to detect fake reviews, with our respondents achieving an average accuracy of 49%. The study shows that there is no general common agreement between humans on what a fake review is, but in some cases there is an agreement whether a review is fake or real regardless of being correct.

Our research is limited to the number of questionnaire respondents in our survey. Thus, we cannot generalize the accuracy of our respondents to a bigger group of respondents or humans as a whole.

Another limitation of our research is the construction of the survey. The survey that was given to our respondents contained 15 questions, which is a very small sample of the data set. This has an impact on the validity of the results that were discovered in the research. In addition, our respondents were only asked to determine if a review was real or fake based on their own assumptions. These individual assumptions were not reflected through the survey and proper insight could not be determined other than accuracy.

To further determine humans' capability of detecting fake reviews, it would be interesting to conduct a similar study by creating a survey where classifying if a user is real or fake using multiple reviews provided by same user, similarly to how the Random forest algorithm [3] does it.

References

[1] J. Pitman, "Local consumer review survey 2022," 2022. URL.

[2] S. He, B. Hollenbeck, and D. Proserpio, "The market for fake reviews," *Marketing Science*, vol. 0, no. 0, p. null, 0. URL.

[3] D. Martens and W. Maalej, "Towards understanding and detecting fake reviews in app stores," *Empirical Software Engineering*, vol. 24, pp. 3316–3355, Dec 2019. URL.

[4] Apple, "App store review guidelines," 2022. URL.

[5] J. Wang, H. Kan, F. Meng, Q. Mu, G. Shi, and X. Xiao, "Fake review detection based on multiple feature fusion and rolling collaborative training," *IEEE Access*, vol. 8, pp. 182625–182639, 2020.

[6] A. Elmogy, U. Tariq, A. Mohammed, and A. Ibrahim, "Fake reviews detection using supervised machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 12, 01 2021. URL.

[7] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *Information Processing Management*, vol. 56, no. 4, pp. 1234–1244, 2019. URL.

[8] J. Salminen, C. Kandpal, A. M. Kamel, S. gyo Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," *Journal of Retailing and Consumer Services*, vol. 64, p. 102771, 2022. URL.

[9] T. Grande, "Questionaire questions with report from nettskjema," 2022. URL.

[10] T. Grande, "Questionaire results and analysis," 2022. URL.

[11] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What yelp fake review filter might be doing?," in *ICWSM*, 2013. URL.

Revision Table

#	Reviewer comment	Response	Changes done to
1	Improve consistency between RQ, contribution, and gap in previous research	The RQ is already responded to in the discussion part. In the next paragraph we explain that a similar study produced similar results, but that a difference between that study and ours is that the other study used reviews generated by an ML model instead of genuine reviews, thus making our contribution somewhat unique. The research gap will be more clearly defined.	Reformulated the research gap in the last paragraph of II.B.
2	Try to tidy up the results part, and make it easier to decipher	We agree with this point, and have improved the section referred to.	Added cross references for each of the tables/figures in the results section. Also added a new table for comparison to the Martens and W. Maalej paper.
3	Only 4 references and 2 of them are from peer reviewed venues.	Thank you for the valuable feedback. We agree with the fact and took possible measures	Exchanged reference paper with journal paper and added more references.
4	What's the motivation and need for comparing the accuracy of ML algorithms with humans? This is a big question mark on framing of your research. Also, the research gap is not clearly presented.	Regarding the first part: we feel like this has already been explained in the introduction (i.e. if the gap is big there might be a need to apply these algorithms to consumers), although it can be made more explicit. As for the research gap, see #1.	Added a sentence about our motivation to the abstract.
5	Analysis variables are not very descriptive. What does recall and f-score and other variables tell?	This is a good point. The variables will be described.	Made it clearer which description belongs to which variable and added definition of the variables at the start of II.B.
6	How observing users in real world could be labelled as a case study?	The alternative research strategy has been reviewed and changed to an experiment instead as this seems more fitting.	See III.Methods A. Data generation first section.

7	You should provide a scientific reason for these methods were chosen.	The reasoning for the methods chosen has been rewritten to a more scientific language to improve what it earlier lacked.	See III.Methods A. Data generation first and second section.
8	At the beginning of subsection III.C, you claim your nominal quantitative data.	The phrasing of the section has been changed.	See subsection III.C
9	After reading the paper its very hard to see the contribution of the paper. Its providing obvious results that ML algorithms are better than humans.	It may have been obvious that ML algorithms are better than humans, but it is not obvious exactly *how* much better they are. If there is a very significant gap there might be a need to apply these algorithms to consumers, as specified in the introduction.	No changes done.
10	“As your current findings lack novelty and contribution, a suggestion for you to improve you research paper is to spend some more time with your data and possibly do a little more data collection. A possible direction is to see how human are different than ML algorithms. You should look into your data and look for any patterns in the responses; are there any specific type of questions that the humans cannot classify? ….”	Added questionnaire score distribution figure. Analyze and discussing the figure	Figure Section 4 Results. See section 5 Discussion