Research Project Proposal:
Using data visualization to analyze fraudulent reviewing behavior on the Google Play store to inform machine learning

George Alexandru Ciobanita
MSc Software Engineering

Supervised by: Dr Radu Jianu
Submitted on: 30/04/2019
Academic Year: 2018-2019

1. Introduction

The project's aim is to provide a better understanding of the online reviewing phenomenon, using effective visualizations of app-review data from Google Play in order to discover features that help classify review data. Current machine learning algorithms, created with the aim to detect fraudulent reviews, classify reviews into either genuine or false, ambiguous in how the different aspects of a review, such as length, pattern, occurrence period, contribute to the output of the algorithm.

In a perfect world reviews would appear in a uniform manner during a product's life, with the purpose of helping potential customers in their decision-making process. However, the reality is that reviews can be manipulated. We hope to shed some light into this by using visualization to analyze data and acquire an understanding of the bias present in reviews, and use this knowledge to improve how machine learning algorithms detect fraud in reviews, moving away from a simple classification of good or bad to a classification that would analyze the legitimacy, and to a certain extent the intent of the one writing the review.

2. Aims and objectives

In this context the project seeks to answer 2 research questions:
RQ1: What visualizations of review data are effective in showing the different features of reviews in order to better classify them?
RQ2: Can existing machine learning algorithms be adapted to account for review legitimacy and intent? (features illustrated in the data visualization of reviews)

To answer these questions the project will pursue the following 2 objectives:
RO1: To design visualizations of fraudulent reviews on the Google play market that are effective in uncovering the different features of a review that are omitted in existing machine learning algorithms.
RO2: To analyze existing machine learning algorithms, and how they learn from review data sets, and compare their results against the classification obtained through visual means.

The aim is the improvement of algorithms, for better review classification, by using data visualization to drive machine learning algorithms, revealing the intent with which a review has been made, moving away from classifications of reviews as either fake or genuine.

3. Background

The project's aim is to provide a better understanding of the reviewing phenomenon, using effective visualizations of review data from Google Play in order to discover features that help classify review data.

Andy Kirk (2016), states that data visualization is: "the representation and presentation of data to facilitate understanding" with a process of understanding that involves three stages: perceiving, interpreting and understanding. The three stages correlate with our intentions to better perceive the review dataset, revealing features and patterns for interpretation, allowing us to understand reviews and the intention with which they have been created. This justifies our need to identify data visualizations to maximize the results each stage has to offer.

Current machine learning algorithms, created with the aim to detect fraudulent reviews, classify reviews into either genuine or false, ambiguous in how the different aspects of a review, such as length, pattern, occurrence period, contribute in the development of the algorithm.

ML is a subset of Artificial Intelligence, allowing for computer systems to perform specific tasks, centered around pattern recognition and decision making without being explicitly programmed for the task. For this to be possible, ML algorithms develop a mathematical model, commonly in three different stages: training, validation and testing, requiring multiple datasets, one for each stage. Focus is on training and validation stages, as this is where the algorithm learns and is tuned to detect fraud in a review.

In a perfect world reviews would appear in a uniform manner during a product's life, with the purpose of helping potential customers in their decision-making process. However, the reality is that reviews can be manipulated.

A survey conducted by PowerReview(2015) shows that "95% of consumers use reviews and 86% say they are essential when making purchase decisions", making it attractive to manipulate reviews. As the market keeps growing, so does the number of reviews created, making it impossible for humans to keep up and manually analyze suspicious reviews. As an example, in 2018, the Play Store had a total of 2.6 millions reviews(Statistica 2019), and in the same year, Google had removed millions of reviews and ratings detected to be fraudulent(Ye et al., 2018).

We hope to shed some light into this by using visualization to analyze data and acquire an understanding of the bias present in reviews, and use this knowledge to improve how machine learning algorithms detect fraud in reviews, moving away from a simple classification of good or bad to a classification that would analyze the legitimacy, and to a certain extent the intent of the one writing the review.

The main characteristic of a review is the intent with which it was written. It is impossible to determine with 100% certainty that a review is genuine or fake(Shivagangadhar et al., 2015).

There are ML algorithms that specialize in sentiment analysis to "automatically identifying whether a user-generated text expresses positive, negative or neutral opinion about an entity"(Jain et al.,2016), and through data analysis and data visualizations we will display how they determine intent. With this knowledge we hope to add to or improve to the parameters present in the training and validation stages of the ML algorithm, possibly influencing the outcomes of the decision phase.

4. Methodology

4.1. Research Objective 1:

To design visualizations of fraudulent reviews on the Google play market that are effective in uncovering the different features of a review that are omitted in existing machine learning algorithms.

The creation process, of review data visualizations, involves four stages, shown in Figure 1, which will be explained in detail.

Figure 1: The Four Stages of the Visualization Workflow
Data Visualization a Handbook for Data Driven Design, Andy Kirk (2015)

4.1.1 Formulating the brief

With the brief we establish the objective's context: the curiosity which drives its development, the circumstances surrounding it and its purpose.

The **curiosity**, and the intrigue of the project audience, is the desire to uncover the different features of a review, focusing on fraudulent reviews, that help in determining the intentions with which it was written. The project audience includes markers and readers from different background, possibly without any knowledge on the subject, which will be discussed more in the circumstances of the objective.

The **circumstances** of the objective are shaped by multiple factors. Knowing our target audience, an explanatory visualization will be developed, meaning that it is our responsibility to deliver the data and its insights to the viewer, assisting the viewer in the three stages of understanding. Other circumstances include time constraints, discussed in the **Work Plan** section, and tools, which will be discussed in the **Methods and Tools** section.

The **purpose** of the objective is to aid research who develop machine learning algorithms, building the visualizations to help understand what features predict intent, to incorporate them in ML algorithms.

4.1.2 Working with data

The datasets which will be used for the visualizations were issued by the project supervisor. The data available is qualitative, composed of: user name, user id, review date, review title, rating value and application id.

Before entering the editorial and design stages of the workflow, data needs to be analyzed in order to see its characteristics. Examining data at this stage, we have 1000 reviews, and the data is mainly textual. In later stages of the project, when the project begins, data will be analyzed, transformed to correct any quality issues and explored for its qualities using visual and statistical techniques.

4.1.3. Establishing the editorial thinking

Editorial thinking is the decision of selecting which perspective offered by the data will be selected, after being shaped by the needs of the project. Editorial thinking is used when we attempt to convey something that is in the data. This will come into play after visual analysis of the data has finished, for visualizations aimed at the project audience.

4.1.4. Developing the design solution

This stage focuses on the creation of the visualization design, divided into the following: data representation, interactivity, annotation, color and composition. A major influencer of this stage, including all the previous stages, is the scope with which the visualization will be created.

The scope of the visualization is about the audience that will have access to it. If only the student and the marker have access to the visualization, the visualization needs to be clear yet the concern on some aspects of the design, such as interactivity and color is lowered. If the outcome of the visualization, and the visualization itself, were to be made available to the public, on the internet, the effort would increase, one example being interactivity on multiple platforms.

Further discussion on the development of the research objective in the **Methods and Tools** section.

4.2. Research Objective 2:

To analyze existing machine learning algorithms, and how they learn from review data sets, and compare their results against the classification obtained through visual means.

We will identify how ML algorithms learn and detect fraudulent reviews and verify the possibility of improvement using the classification of intent from our data visualization. The focus is on supervised and unsupervised ML. This decision was made to focus the project on a specific area of ML algorithms but also because it lines up with the mentality that no review can be verified with 100% certainty as "using either supervised or unsupervised method gives us only an indication of fraud probability...no stand alone statistical analysis can assure that a particular review is fraudulent one"(Shivagangadhar et al., 2015).

5. Methods and Tools

5.1.1 Research Objective I:

Oates (2006) suggests that a research strategy is required when attempting to answer our research questions. The main data type of the project is qualitative, textual to be specific, meaning that a qualitative data analysis approach is suitable for the project. Oates explains that this strategy "involves abstracting from the research data the verbal, visual or aural themes and patterns that you think are important to your research topic".

The intention of the project is to discover theories as to what exactly makes a review fraudulent, defined as grounded theories, which is explained by Oates as an approach where we "analyze the data to see what theory emerges, so that the theory is grounded in the field data", as theories developed prior to their discovery could cloud the results discovered through data analysis. To support this visualizations will follow have an exploratory design, creating multiple visualizations which will be gradually be reduced until we figure out which ones are more insightful.

The reviews datasets will be transformed into a form ready for analysis, as the shape of the data needs to be consistent across all datasets, requiring multiple duplicates, all unique in their purpose.

The next step is the analysis of the shaped datasets in order to identify key themes, relevant to our research questions. Key themes are labels based on concepts found in the data. Analysis is constant process and will happen throughout the project until a satisfying theory is developed.

5.1.2 Research Objective I – Tools:

Microsoft Excel will be used in the early steps of the project, specifically in shaping the reviews datasets and initial analysis of themes, support for table formulas and early visualizations will aid the project.

Long (2017) informs us of the visualizations that are regularly used in qualitative research, some of which are useful to us: coding stripes, word clouds, charts, word trees, concept maps, mind maps, hierarchical charts, explore diagrams, comparison diagrams, project maps.

Vega-Lite is the tool that will be used to develop the above visualizations, and further analysis. Developed by University of Washington Interactive Data Lab, it is a free and open source tool that provides great visualizations, allowing for many programming languages to be used, programming language to be used is Python, as Vega-Lite charts use .json based syntax, while also being easy to host on websites(Linacre, 2018).

5.2.1 Research Objective II:

Document-based research will be conducted to achieve the aims of RO2. Target documents are academic literature, journal, books and conference papers, together with research data from and organizational research.

Documents will be evaluated to see if they are authentic, considering their context and purpose. Document context should be in the machine learning area of study, with the purpose of proposing or verifying machine learning algorithms. Focus should be on machine learning algorithms that promote the classification of textual data, such as twitter posts or customer reviews.

Documents should contain discussions about the proposed framework of analysis, preprocessing of data, classifier algorithms, evaluation and results, ideally also containing the dataset used.

6. Work Plan

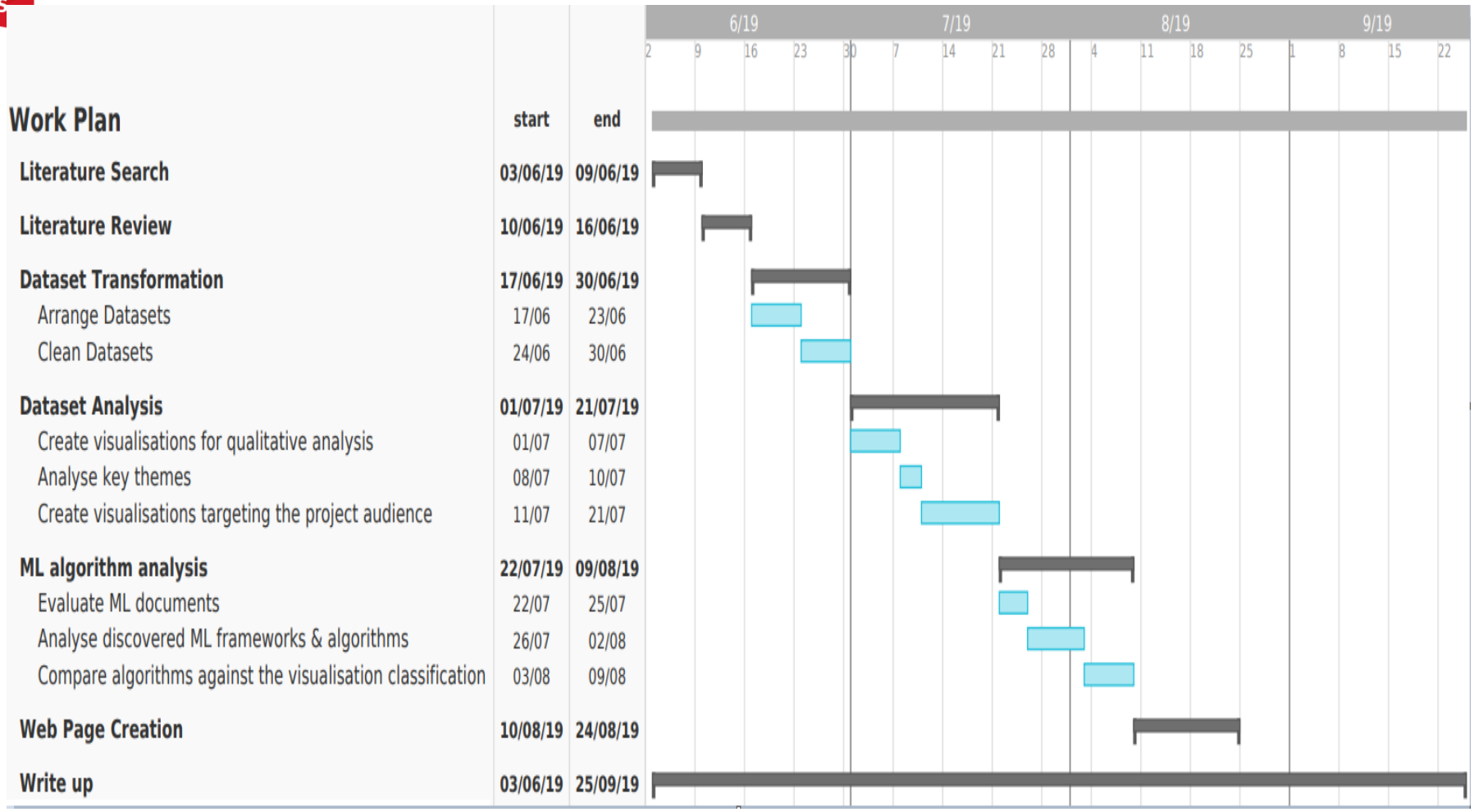

Figure 2: Project Work Plan

7. Risks

Risks	Risk Triggers	Likelihood	Impact	Mitigate
Loss of project work.	Corrupted files. Missing files. Hard drive failure.	3	5	Having various storage locations, both physical, e.g. external hard drives, and cloud, e.g. Google Drive.
Change of requirements.	Change of project aims and objectives. Research leads to new questions.	2	5	Extensive research will be carried out on current objectives and aims. Consult the supervisor whether a new project direction is feasible while also meeting the project deadline.
Visualizations fail to aid in the analysis of fraudulent reviews.	Selection of visualizations fails to reveal review features.	2	5	A multitude of data visualizations will be created to aid in data analysis.
Visualizations aid in the analysis of fraudulent reviews but lack details.	Attention hasn't been given to aspects of the visualization such as colour, interactivity, composition. No attempts have been made to enhance the visual appeal of the visualizations.	3	3	Time has been allocated to the creation of visualizations that appeal to the target audience of the project.
Insufficient time to develop a website hosting the created visualizations.	Unexpected bugs when developing the website. Lack of time to properly the website.	4	3	I have picked Vega-Lite for it's flexibility. Through the Altair API, it allows me to use Python for its development and has access to a multitude of repositories and documentations aimed at the development. Adding to this is the usage of Vega-Lite during the Data Visualization module at City University of London.
Inadequate or insufficient documentation on ML detection of fraudulent reviews	Lack of ML applications on review analysis Existing applications lack in documentation.	3	4	Research is available that focuses on the sentimental analysis of other types of qualitative, textual, data such as Twitter, movie reviews or new articles, which can still be used to understand intentions behind the qualitative dataset available to the project.

8. References

1. Dandannavar P., Jain P., (2016), "Sentiment Classification using Machine Learning Techniques", *International Journal of Science and Research(IJSR), Pages 819-821*
2. Kirk, A. (2016) *Data Visualization: A Handbook for Data Driven Design*. Los Angeles: SAGE.
3. Linacre, R. (2018) *Why I'm backing Vega-Lite as our default tool for data visualization.*Available at: https://medium.com/@robin.linacre/why-im-backing-vega-lite-as-our-default-tool-for-data-visualisation-51c20970df39 (Accessed: 30th of March 2019)
4. Long, A. (2017) *Popular Techniques for Visualizing Qualitative Data*. Available at:https://www.linkedin.com/pulse/popular-techniques-visualizing-qualitative-data-adam-long/ (Accessed: 30th of March)
5. Nagayama K., Ye F., (2018), "*In reviews we trust – Making Google Play ratings and reviews more trustworthy*". Available at: https://android-developers.googleblog.com/2018/12/in-reviews-we-trust-making-google-play.html (Accessed: 30th of March 2019)
6. Oates, B. J. (2006) *Researching information systems and computing*. London: SAGE.
7. PowerReviews(13th of February 2015), *Survey Confirms the Value of Reviews, Provides New Insights*. Available at: https://www.powerreviews.com/blog/survey-confirms-the-value-of-reviews/ (Accessed: 27th of March 2019)
8. Shivagangadhar K., Sagar H., Sathyan S., Vanipriya C.H.(2015) "Fraud Detection in Online Reviews using Machine Learning Techniques", *International Journal of Computational Engineering Research, Volume 5, Issue 5*. Available at:http://docplayer.net/15165970-Fraud-detection-in-online-reviews-using-machine-learning-techniques.html
9. Statistica(2019), "*Number of available applications in the Google Play Store from December 2009 to December 2018".* Available from: https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/ (Accessed: 29th of March 2019)

9. Ethics checklist

Research Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC)

http://www.city.ac.uk/department-computer-science/research-ethics

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/*	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/*	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**

2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/*	**NO**
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London?	**NO**

	For example, students studying on a particular course or module. *If yes, then approval is also required from the Head of Department or Programme Director.*	
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**NO**