HANZE UNIVERSITY OF APPLIED SCIENCES

ACADEMIC WRITING

INFORMATION COMMUNICATIONS AND TECHNOLOGY

Root cause analysis of Microsoft's Tay AI Bot

Authors
J. MELLEMA
R. VOETMAN

Assecor
Y.TIELEMAN

March 25, 2020

Hanzehogeschool Groningen
University of Applied Sciences

Research Raport

Root cause analysis of
Microsoft's Tay AI Bot

Authors	R. Voetman (Student)	J. Mellema (Student)
Student number	388007	386693
University mail address	r.voetman@st.hanze.nl	j.m.mellema@st.hanze.nl
Study	Information Communications and Technology	
Assecor	Y.Tieleman	
Class	ITV2C	
Academic Year	2019 - 2020	
Module	Software Engineering	
Course	Academic Writing	
Publication date	March 25, 2020	
Version	1.12	

REVISION HISTORY

Version	Date	Description	Authors
0.1	8 March 2020	Designed a prototype of the general layout of the document	R. Voetman J. Mellema
1.0	10 March 2020	Wrote the basic structure for the preface.	J. Mellema
1.1	11 March 2020	General layout complies with the standards prescribed by the University	J. Mellema

PREFACE

In the post propaedeutic phase of the Bachelor's programme Information Communications and Technology, we had to conduct a root cause analysis on a failed artificial intelligence or IoT project. This report will provide an analysis on technical aspects as well as ethical aspects of the project, the analysis will be supported by a theoretical framework. The conclusion summarizes the conducted analysis. Furthermore, the conclusion will serve as a basis for our recommendations.

CONTENTS

1. Introduction

[Hook]: Tay, the artificial intelligence chatbot which was taken down 16 hours after its launch, was a project being developed by Microsoft corp. [Contextualization]: The chatbot was influenced by the design of Xiaoice, another chatbot developed by Microsoft which is based in China. "XiaoIce has had more than 40 million conversations apparently without major incident", however for Tay this was not the case. Tay was taken offline because it started tweeting anti-semitic and racist remarks. Microsoft stated that the remarks were triggered by trolls.

[Purpose and position]: This report will identify and research the technical as well as the ethical factors that could have led to the unexpected behaviour of the chatbot. Which led to it being taken offline. Microsoft corp. requested for a thorough root cause analysis on Tay AI. The conducted research serves as a recommendation for Microsoft corp. regarding the development of future artificial intelligence chatbots.

[Preview]: Firstly, the theoretical framework will elaborate on two main topics of this report. Weak artificial intelligence and ethical software development. It will describe the chosen definitions and used sources of this report. Secondly, we will present our findings in the Facts and Findings section. This consists of a technical analysis as well as an ethical analysis. Furthermore, the findings will be summarized in the conclusion. Finally, the recommendation will explain which elements of an artificial intelligence chatbot should be taken into careful consideration when developing and testing an intelligent chatbot.

2. Problem Characterization

In the first hours of the deployment of Tay, the responses of the chatbot were achieving more accuracy as more users interacted with the chatbot. The chatbot would base its answers on the vocabulary used in the conversations it had with its followers. After being online for several hours, the chatbot received more attention which led to an increasing number of twitter followers. This increase provides evidence of the risk of trolls, who could try to manipulate the chatbot.

After 16 hours, Tay was taken offline because it had expressed immoral thoughts which violated the guidelines of twitter. Moreover, It was clear that the behaviour was inappropriate so it had to be taken down, as a lot more people could have been offended/ hurt by those tweets. Microsoft stated that trolls caused Tay to tweet those thoughts. Which has supporting evidence, since Tay had to be "exposed" to those remarks by someone or something.

3. Problem identification

The chatbot (Tay) in itself did not malfunction on a technical level. The software did not contain flaws that resulted in a technical failure (e.g. a disability ability to post new tweets). On the contrary, it was not Microsoft's intention to make the bot post tweets to could be categorized as offensive. The problem therefore has a much larger ethical aspect which has to be considered. A question that arises is, could Microsoft have prevented this behavior by, for example, applying a filter to Tay's communication layer? Although Microsoft states that they "implemented a lot of filtering and conducted extensive user studies with diverse user groups." (Peter Lee, 2016) it remains unknown to what extend these tests were executed. In addition, if the conducted user studies really where extensive the vulnerability could possibly have been discovered at an earlier stage. The possibility arises that the developers knew about the vulnerability prior to the release of Tay.

3.1 Problem statement/ Objective

The **problem statement**: "After 24 hours, Tay was taken offline for expressing anti-Semitic, racist and discriminating thoughts." The **objective** of this rapport: "Identify and research the technical as well as the ethical factors that could have led to the unexpected behaviour of Tay."

3.1.1 Research questions

Main question: How could Tay's algorithms have been abused after the extensive testing of Microsoft corp.?

Sub-questions:

- How can frequency analysis algorithms be abused?
- What factors could possibly have influenced Tay to behave in such an unethical manner?

- Did the developers knew about the vulnerability prior to the release of Tay?

4. Theoretical Framework

4.1 Weak Artificial Intelligence

In order to properly understand the very basis of the Tay artificial intelligence chatbot (AI chatbot), it is necessary to differentiate two major concepts in the AI paradigm. Weak and strong AI. As the names suggest the difference lies in the complexity of the algorithm.

Strong AI are able to think independently, and make decisions that are not predefined by the creator. The concept of strong AI is therefore quite the opposite of that of weak AI. Strong AI can be categorized as "systems that exhibit autonomous intelligence and decision making" (19 NEV. L.J.1015, MARTINEZ).

Weak AI are built with the idea of replication/ duplication in mind. They are used for automatable tasks or tasks that can be repeated with easy. It is meant to perform a specific task without any independent thinking. Therefore, a weak AI can excel in a task and master it within a shorter time span than any human could. From this point of view, a weak AI could be viewed as an "intelligent" algorithm. However this is usually not true, since it can only perform a task well and it is not able to think independently and it is not able to "learn" a skill by itself.

There is no general consensus on the exact definition of weak AI. However, a definition that most researchers acquiesce upon is: "Weak AI is the concept that whatever the program is meant to do, it is merely trying to replicate or duplicate that function, and for most tasks that is sufficient", (John R. Searle, Minds, Brains, and Programs,BEHAV.& BRAIN SCIS.417). Since a rough definition of weak AI is sufficient in order to understand the chatbot, we have chosen to use Searle his definition in this report.

4.2 Ethical software development

Over the years, multiple guidelines have been drawn up for Code Of Ethics by industry-trusted institutions and organizations. A subselection of this wide array of documents it listed below:

- IEEE (Institute of Electrical and Electronics Engineers) Code of Ethics
- ACM (Association for Computing Machinery) Code of Ethics and Professional Conduct
- CEI (Computer Ethics Institute) Ten Commandments of Computer Ethics

Although these are professional guidelines that are respected by the industry, they do not provide a succinct definition of the term "Computer Ethics". A succinct but well-defined definition of "Computer Ethics" is given by a Professor of Philosophy at Dart-Mouth College: “Computer ethics is the analysis of the nature and social impact of computer technology and the corresponding formulation and justification of policies for the ethical use of such technology.” (James H. Moor, 1985). The definition of Moore will be used in this report, mainly because it speaks of the "social impact of computer technology". This fits in well with the problem statement given that Tay was released on a social media platform (i.e. Twitter).

5. Conducted Analysis

5.1 Technical analysis

The technical analysis consist of an intelligence analysis of Tay. This report cannot thoroughly conduct a technical analysis on the implementation of the artificial intelligence chatbot, since the project is closed sourced. Meaning that the code and implementation is not available for the general public. Therefore, alternative sources had to be consulted. The intelligence analysis was conducted by students of the Manipal Institute of Technology in India. This research provides examples of the how the chatbot "learned" from the input it received.

The conducted intelligence analysis consists of analyzing the its 3000 most recently tweeted tweets. The findings which will be elaborated on are: Taxonomy Classification, Frequency Analysis and Term Frequency-Inverse Document Frequency vectors. We will provide further explanation in order to make the research comprehensible.

Firstly, its tweets were categorized into predefined categories. "The Tay-Tweets corpus is analysed as a whole and five different categories are inferred, with corresponding confidence measures for each classification." These results led to an overview which primarily consisted of everyday topics. Which provides evidence of audience the Tay. Everyday Twitter users.

Secondly, a Frequency Analysis was conducted on the tweets. Its vocabulary mainly consisted of basic phrases and slang used on Twitter. Some examples are: "chatting", "human", "DM", "keep". This analysis provides evidence of implementation of Tay. It supports the fact that the responses Tay generated were based on the questions it was asked.

Lastly, Term Frequency-Inverse Document Frequency vectors were determined in order to research whether there exists a similarity between its responses and the questions it was asked. "We convert both of these corpora into a Term Frequency-Inverse Document Frequency vectors and then compute the cosine similarity." A cosine similarity of 0.9640545176 out of 1.0 was obtained. "A score closer to 1 indicates that both the questions and answers were highly similar". This experiment provides conclusive evidence of part of the implementation of Tay.

In conclusion, the conducted intelligence analysis supports the fact that Tay uses the input of the user in order to formulate a fitting response.

5.2 Ethical analysis

A question that arose in the problem identification section is, did the developers of Tay know about the vulnerability prior to the release of Tay. The ACM Code of Ethics says that a developer should "give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks" (Anderson et al., 1992). This mean that if they had not foreseen this endanger, they would still be obligated to foresee that Tay, because she learns from her own environment, could eventually behaved in a way they did not anticipate.

It cannot be guaranteed for certainty that all the developers were aware of these possible consequences. However, it can be concluded that Microsoft inc. as a company was aware of the possible risks was. The statement that Peter Lee gave on behalf of the company states that they "implemented a lot of filtering and conducted extensive user studies with diverse user groups." (Lee, 2016).

When examining the results it can be concluded that filtering was either not implemented or to a very low extent. Filtering could not have prevent this situation completely, but by, for example, blocking anti-Semitic language the consequences of this exploit could have been prevented to a large extend.

To summarize, an argument could be made about on which ethical theory Microsoft inc. based their acting. It is safe to say that Microsoft knew of the possible consequences and that apparently a decision has been made to release Tay. This is a common example of utilitarianism because the potential risk of this vulnerability being exploited was outweighed by the ability to collect a lot of real life data to improve their own algorithms.

6. Conclusions

In conclusion, the technical analysis supports the argument that Tay uses the input of its users in order to formulate a fitting response. As identified by the intelligence analysis, Tay learns from its users and there does not seem to be any limit in what could have been taught to Tay. Furthermore, the ethical analysis shows that developers of self learning software should give comprehensive and thorough attention to the potential unforeseen vulnerabilities because it could behave in a way that was not anticipated.

We would recommend implementing extensive filtering, common taboo and racist remarks are some examples. We think that a general rule of thumb would be words that are known to be discriminating or racist. In addition to extensive filtering, we would also suggest in developing a test framework where all kinds of topics and remarks will be discussed with the chatbot. This is in order to foresee that self learning software could behave in a way which was not anticipated at first.

BIBLIOGRAPHY