Catch Them If You Can: Malicious Behavior Simulation in Deep Question Answering

Nikita Galinkin
Supervisors: Zoltán Szlávik[1], Lora Aroyo[2] and Benjamin Timmermans[1]

[1] *IBM Benelux Center for Advanced Studies, Amsterdam*
[2] *Vrije Universiteit, Amsterdam*

1 Introduction

Recent advances in artificial intelligence and machine learning have allowed question answering systems to become much more prominent for retrieving information to handle day-to-day tasks. In the study, we investigate the impact of malicious user behavior in question answering systems specifically deployed in the cultural heritage domain. We need to be prepared for some users being 'malicious', trying to render a learning system useless by misusing it. To prepare, first we need to estimate the impact of malicious actions, and study ways to deal with this issue.

Recently, the problem of malicious attacks has been broadly studied in the machine learning community [3]. It has been shown that in some cases 20% of malicious data can lead to a ten-fold increase in classification errors [5]. The main trend in these studies is to investigate the worst-case scenario of the machine learning model with the assumption that the attacker has full knowledge of the system and its actions are optimal. This assumption rarely holds in a system used in real-life. In this study, we propose a user model to simulate users attacking the system to gain a better indication of what might happen.

2 SQALPEL: A Deep Art History Question Answering System

As a use case to study the impact of malicious feedback on question answering systems we use the SQALPEL system, developed during the like named project. The project is a collaboration between the IBM Benelux Center for Advanced Studies in Amsterdam, the Mauritshuis museum in Den Haag, and the Vrije Universiteit Amsterdam. The objective of the project is to make use of new techniques to answer and raise new questions about the subjects of a painting, relevant historical context and changing interpretations over the course of time in the art historical literature.

During the project, we have developed SQALPEL, a deep art history question answering system that is designed to answer questions about Rembrandt's famous painting "The Anatomy Lesson of Dr. Nicolaes Tulp". The painting, owned by the Mauritshuis, was chosen for this project by the museum as it has been widely studied in an art historical context over the course of time (see, e.g., [1]). The SQALPEL system is built on IBM Watson technology[1], relying on services such as the Natural Language Classifier, Retrieve and Rank, and Conversation, to engage the user by answering questions in conversational manner. The system relies on previously created question-answer pairs (i.e., via crowdsourcing) as well as on automatically extracted document passages retrievable by the system. End user input for training the system is taken into account in two ways: primarily, we employ a thumbs up/down functionality when users receive answers, but we also process never before seen questions after expert validation. This paper considers the former kind of user input.

[1] `https://www.ibm.com/watson/products-services/`

3 Methodology for Malicious User Behavior

The data for this study contains 4037 question-answer pairs, consisting of 3797 unique questions and 554 unique answers. Of the answers, 131 were collected from interviews with museum visitors, the rest we have extracted from literature about the painting. Answers from both sources were enriched with questions through the crowdsourcing platform called CrowdFlower, where human workers were given the answer, and they had to come up with a diverse set of questions that match the answer within the task. Questions collected from the interviews are more frequently asked, thus, on average, there are four times more questions to each interview answer.

We simulate two weeks of operation with 200 users daily, each user simulated according to user models with various percentages of maliciousness expressed as providing the right or wrong feedback to answers given to questions with known answers in the knowledge base. The classifier we built was inspired by the question answering evaluation metric POURPRE that measures the co-occurrence of words between a given answer and an answer nugget. An answer nugget is defined as a string containing a fact for which the assessor could make a binary decision as to whether a response contained that nugget [4]. In the crowdsourcing task aiming to collect questions, users were asked to highlight the part of the answer that justifies the answer to the question. These annotations we use as the answer nugget.

4 Results and Conclusion

We found that with up until 20% of malicious feedback the system continues to learn, and system performance does not drop below the level of the initial system. With over 30% of the users being malicious, system performance will gradually decrease. With 60% of malicious feedback it drops by a factor of two after only two weeks. When a peak of low quality feedback appears, we investigated whether the timing of the peak makes a difference. The results indicate that a peak in malicious activity, in addition to negatively influencing system performance, creates a long term effect. The timing of the peak defines its influence on system performance. The later the peak occurs, the less effect it will have, as by now the system is expected to be more robust.

For somebody who plans to deploy a question answering system with limited technical staff in-house, the results of our study indicate that user behavior and misbehavior have to be watched closely in the first weeks of system operation, as the potentially negative effect on overall system performance can be severe. When the system is retrained with new data, it might indicate overall level of maliciousness. It can then be decided on when to restore the system to a previous state, discarding data with too high level of low quality input. For user-level filtering, we might consider employing various agreement-based metrics such as those, for example, offered by CrowdTruth [2].

References

[1] Arnon Afek, Tal Friedman, Chen Kugel, Iris Barshack, and Doron J Lurie. Dr. tulp's anatomy lesson by rembrandt: the third day hypothesis. 2009.

[2] Lora Aroyo and Chris Welty. The three sides of crowdtruth. *Journal of Human Computation*, 1:31–34, 2014.

[3] Marco Barreno, Peter L. Bartlett, Fuching Jack Chi, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, Udam Saini, and J. D. Tygar. Open problems in the security of learning. In *Proceedings of the 1st ACM Workshop on Workshop on AISec*, AISec '08, pages 19–26, New York, NY, USA, 2008. ACM.

[4] Jimmy J. Lin and Dina Demner-Fushman. Methods for automatically evaluating answers to complex questions. *Information Retrieval*, 9(5):565–587, 2006.

[5] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In *ICML - Volume 37*, ICML'15, pages 1689–1698. JMLR.org, 2015.