**Algorithm Analysis of Microsoft’s Chat AI Twitter Bot**

It took less than 24 hours for Twitter to corrupt and innocent an AI chatbot. In 2016 Microsoft launched a chatbot named Tay to twitter, where their intended purpose was for entertainment targeted for the younger generation between 18- to 24-year-olds. Essentially, how it works was the more you chat with Tay the smarter it gets by learning through people’s language and engage through playful casual conversation. Twitter users starting tweeting the bot with all sorts of misogynistic, and racist remarks. With Tay being a robot parrot, it started repeating these sentiments to users. 16 hours of its existence Microsoft shut down Tay which had more than 96,000 tweets.

Searching through the thousands of tweets, it shows that many of the bot's utterances have simply been the result of copying users. Simply put, if you ask Tay to “repeat after me” it will. This makes it easier for anyone to put words into an AI. Microsoft claims that Tay was built using relevant public data that has been modeled, cleaned, and filtered. However, this was not the case when it went public. A further analysis of the intelligence of this chatbot was published in the International Conference on Contemporary Computing and Informatics.[[1]](#footnote-1)

The first analysis that was done by classifying the tweets into a hierarchical taxonomy. These are predefined categories, which branch out into various subcategories and the text is classified into the most appropriate branch. This was done to make it easier to understand what Tay was talking about in a broad sense. Next was a text categorization, which allows the tweets to be more narrowed down by the conversations Tay had into one unified heading. An estimation of the type of entity, and its relevance was done which showed that the bot was intelligent enough to talk about a wide range of topics. Then a frequency analysis was implemented using translate characters functionality of the Unix platform on the tweets collected. To understand how original Tay was in its tweets, a TF-IDF cosine similarity was done to find the similarity between the questions it was asked and the response it generated. A score closer to 1 indicates that both the questions and answers were highly similar and a score closer to 0 indicates that there was significant difference between them. After vectorizing the result obtain a cosine similarity of 0.9640545176. The publication had performed various NLP task to analyze and quantify the intelligence of the chatbot. To define the Bot Intelligence Score, it had a weighted sum of the results of the NLP task. They empirically define the values of these weights and calculate the Bot Intelligence Score for Tay to be 99.438. The analysis has shown that even though Tay was online for just 16 hours it has learned very efficiently from its interactions, had it been online for a longer time, its Bot Intelligence Score would certainly have increased.[[2]](#footnote-2)

Enterprises need to ensure, to the best of their ability, that all uses of data and AI are ethical, and that customer, employee, and other data are kept protected and secure. The Ted Talk by Cathy O’Neil mentions how humans are responsible for creating the algorithm produced. A question arises on how humans are supposed to teach AI to use public data without incorporating the worst traits of humanity. Since humans create bots that mirror their users, should we care if their users are human trash? It’s us who is responsible of what data we input.

1. <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/> [↑](#footnote-ref-1)
2. <https://ieeexplore.ieee.org/abstract/document/7917966> [↑](#footnote-ref-2)