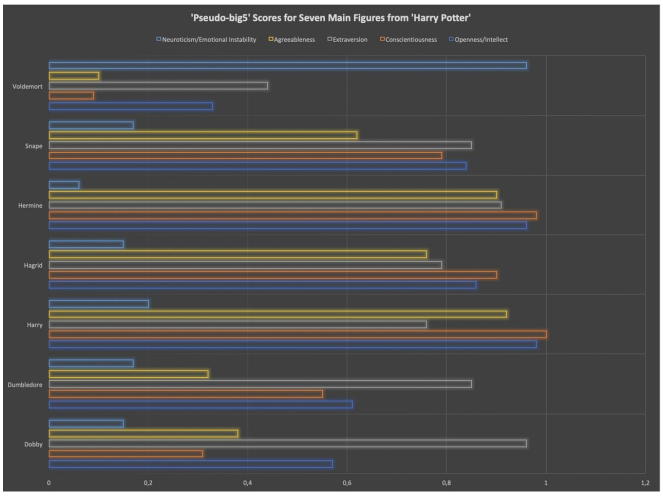
SentiArt: a sentiment analysis tool for profiling characters from world literature texts

15 July 2019, by Ingrid Fadelli



Pseudo-big 5 scores for seven main figures in the Harry Potter books. These scores are percentiles based on a sample of 100 figures appearing in the book series. Credit: Arthur M. Jacobs.

Arthur Jacobs, a professor and researcher at Freie Universität Berlin, has recently developed SentiArt, a new machine learning technique to carry out sentiment analyses of literary texts, as well as both fictional and non-fictional figures. In his paper, set to be published by *Frontiers in Robotics and AI*, he applied this tool to passages and characters from the Harry Potter books.

Jacobs has a background in neurolinguistics, a branch of linguistics that explores the neural mechanisms associated with language acquisition, comprehension and expression. In his previous work, he has often investigated how machine learning tools could be used to analyze and better understand human language. He is particularly interested in what he calls computational poetics, an area of study that focuses on the use of computational tools to understand literary content.

"In 2011, I wrote a book with Austrian poet Raoul Schrott called 'Brain and Poetry,' where we speculated that it would help to develop sentiment analysis tools for literary texts and poetry, not only for movie reviews or Trump tweets, which appears to be the gold standard in classical sentiment analysis," Jacobs told TechXplore. "We also wanted to develop a tool that can predict human neuronal and behavioral data, not only self-reports collected via Amazon Turk."

In his new study, Jacobs tried to put some of the ideas introduced in his previous work into practice by developing a tool for analyzing sentiment in literary texts. The technique he proposed, called SentiArt, uses vector space models and theory-guided, empirically validated lists of labels to compute the valence of individual words in a text. Vector space models are representations of text documents as vectors of identifiers, which are often used to filter, retrieve or organize information.

"SentiArt is a very simplistic tool that can be used by non-experts to simply compare the words in their test text (i.e., the text they want to do a sentiment analysis on) with an excel sheet that they can download from my homepage for free," Jacobs explained. "In principle, the tool should work in any language for which you can download Facebook's so called vector space models, on the fastText webpage. While my study focuses on English and German, you could also use it in Malaysian, Farsi or a Chinese dialect, and a multitude of other languages, as fastText has vector space models for over 290 languages."

Jacobs highlights that SentiArt is fairly easy to use, adding that he was able to teach 30 German literature students how to use it during a one-hour class. In his recent work, he tested the tool's accuracy using data gathered during a neurocognitive study and then used it to compute emotional and personality figure profiles for some of the main Harry Potter characters, including Voldemort, Snape, Hermione, Hagrid, Harry, Dumboldore and Dobby.

Interestingly, he computed these characters' emotional figures and personality profiles based on the 'big five' personality theory, an established construct in psychology research. The 'big five' theory is generally used to roughly measure people's personality traits based on five key dimensions, namely openness, conscientiousness, extraversion, agreeableness and emotional stability.

Jacobs carried out a series of analyses comparing the tool he developed with other machine learning classifiers for sentiment analysis, such as Vader and Hu-Liu. SentiArt performed remarkably well in predicting the emotion potential of text passages from the Harry Potter books, while also making plausible predictions about the emotional and personality profile of fictional characters. Finally, the tool attained a promising cross-validation accuracy in classifying 100 fictional figures into 'good' or 'bad' ones.

"The paper is on a few limited applications and in two languages (German/English), so before I can speculate on the application potential, being an experimental scientist, I would wish to have many more cross-validation studies using human data," Jacobs explained. "This is just how I am trained, although usually in natural language processing (NLP) or the machine learning community these are not the main priorities. But as a neurolinguists, we would always try to test the predictions of an algorithm with human data before we speculate on what it is really useful for."

Although Jacobs emphasizes the need for further studies to ascertain SentiArt's effectiveness and generalizability, the tool he developed could eventually have numerous interesting applications. For instance, it could be applied in fields such as computational linguistics, personality psychology, digital humanities and perhaps even in clinical settings. It can, in principle, also be applied to non-fictional characters appearing in Wikipedia or Wikinews, e.g. Winston Churchill, Marilyn Monroe or Angela Merkel.

"The model fit with a first set of empirical data, the Harry Potter ratings, is definitely encouraging," Jacobs added. "Also two of the most popular sentiment analysis tools I compared it with do not fare better in this context, so I think this is an achievement that deserves publication. I think that showing the emotional character profile for Voldemort or Harry Potter was a nice gimmick, but of course, the tool could also be applied to non-fictional characters too."

Jacobs is now planning to carry out further cross-validation studies testing his model's predictions with human data. He hopes that teams at other universities will do the same, either using data collected via Amazon Turk or neuroimaging data, as in the "Harry Potter' study carried out in his lab. In addition, he would like to explore ways to improve the performance of sentiment analysis tools in tasks using machine learning regressors instead of classifiers.

"Machine learning approaches are generally divided into two different types," Jacobs explained. "The first are classification approaches, which classify data into categories, such as positive or negative. This is where my algorithm does very well. The hard test is not classification, it's regression, which entails fitting an algorithm's predictions to continuous human data, such as ratings on a scale from one to 10. Few people in sentiment analysis use regressors, especially for literary texts, because accuracy tends to drop, for instance, from over 90 percent to about 30 percent to 50 percent. I would like to see more work testing this, and once more empirical data has been published, I will try to improve parts of the algorithm in agreement with this new data."

In addition to his research endeavors, Jacobs will soon start teaching natural language programming (NLP) and machine learning as part of a new data science course at Freie Universität Berlin. His hope is to train new generations of data scientists to value the collection of empirical human data related to reading literature and poetry just as much as publishing code or predicting particular things.

More information: Sentiment analysis for words and fiction characters from the perspective of computational (neuro-) poetics. DOI: 10.3389/frobt.2019.00053. www.frontiersin.org/articles/1 … .2019.00053/abstract

Computational Poetics blog. sentiart.de/blog/

APA citation: SentiArt: a sentiment analysis tool for profiling characters from world literature texts (2019, July 15) retrieved 21 June 2021 from https://techxplore.com/news/2019-07-sentiart-sentiment-analysis-tool-profiling.html