University of Antwerp

Report on the Internship at CLiPS

Dimitrios Boumparis

MA "Digital Text Analysis"

s0211156@ed.ua.ac.be

University
of Antwerp

This internship report presents the main experiences during a smaller and a larger project that aimed to develop machine learning models to predict the gender and age group of authors of short texts. The project was carried out for a total amount of around 357 hours at the Centre for Computational Linguistics, Psycholinguistics and Sociolinguistics (CLiPS) at the University of Antwerp, under the supervision of Prof. Dr. Walter Daelemans and PhD candidate Jens Lemmens.

Introduction

The project aimed at building and evaluating machine learning models that could automatically predict basic demographic characteristics of authors of short and mid-length texts, based on their writing style. This task is generally known in the Natural Language Processing field as author profiling. Author profiling is typically approached by training models on a labelled dataset of texts written by authors of known genders and age groups. The training process of such a model involves using a learning algorithm to extract the patterns and characteristics that are present in the training data. The trained model can then be used to automatically predict the gender or age group of new texts that it has not seen before. Author profiling is applicable in a variety of tasks, such as identifying the gender or age group of anonymous online profiles, detecting fake news, or analysing the writing style of authors in different genres or even languages. It can also provide an insight into the ways in which different genders and age groups tend to express themselves in writing. These predictions can be used in broader pipelines with various applications, such as authorship attribution, namely the task of finding the author of an anonymous text.

There is a variety of different approaches that can be employed to build an author profiling model, depending on the type of data that is available and the specific characteristics that are being examined. The most common approaches are:

- **Supervised learning:** In this approach, the model is trained on a labelled dataset of texts, where the labels correspond to the known genders and age groups of the authors. The model learns to predict the labels of new texts by finding patterns in the training data that are associated with the labels.
- **Unsupervised learning:** In this approach, the model is trained on a dataset of texts without knowing about the labels. The model learns to predict the genders and or age groups of new texts by clustering (i.e., grouping) similar texts in the training data.

Other factors that can affect the performance of an author profiling model include the type of features that are used to represent the texts, the specific machine learning algorithm that is used, and the quality and size of the training dataset. These factors can all be varied in order to optimize the performance of the model and to achieve the best possible predictions.

To this end, we used two large datasets of short online texts written by authors of different genders and age groups to train and test the machine learning models. We experimented with a variety of different machine learning algorithms, textual features, and model hyperparameters, and evaluated their performance in terms of accuracy and F1-score.

In the following sections of this report, we will not present in a strictly academic manner, since this is not a scientific paper, but rather will give a general understanding of the thought process behind my choices, the experiences of training models in the wild, and the main takeaways of being an intern at CLiPS.

Sentiment Lexicon

At the start of the internship, all non-Dutch interns were asked to translate a sentiment lexicon into their mother tongue. As someone who has always been passionate about language and studied Translation Science in their bachelor’s, I was excited to take on this task and contribute to the project. However, even though the task seemed straightforward at first, I quickly realised that some words did not have an easy translation. This was particularly true for words that carried a strong emotional connotation. Some English words had two or sometimes three translations in Greek. Despite the challenges, I was able to complete the translation, which contributed to me remembering where I come from and my past studies and my academic journey overall.

The translated lexicon proved to be incredibly useful for many CLiPS sentiment analysis project. By providing a way to understand and analyse emotions expressed in many different languages (some of them were also morphologically different, such as Russian and Greek), the lexicon allowed for gaining deeper insights into the data the CLiPS students were and are going to be working with. Without it, the sentiment analysis would have been much less focused on the original language of the lexicon (English) and Dutch, in which it was previously translated, and less accurate and effective. Overall, our translation project was a valuable and enlightening experience.

An unexpected turn of events

On 6 July 2022, a fire broke out at the Stadtcampus of the University of Antwerp (UA), which, according to UA's official statement, started due to human error. Many offices of professors, teaching assistants, and personnel of the Faculty of Arts, among others, were located at the building that the fire started from (Building B). Although the fire spread to surrounding buildings, no lives were lost. One could argue that the timing was to thank, since no classes were taking place in early July and thus no large groups of. No one would attempt to image what the evacuation could have looked like if more people of the UA community were at these buildings that day or what could have happened should the fire have started in a lower floor.

This all of a sudden new reality resulted in some UA teaching and administrative staff losing some personal belongings or their entire offices to the fire. Despite the fact that the University started the restoration procedures as soon as they possibly could, Buildings B and D, the upper floor of which was also caught on fire, are yet to become fully operational again (last update was given on 17 August 2022). To this day, the erected scaffolds that surround the whole facade of these buildings are there to remind us of the awful moments the UA and its staff went through in the middle of the summer.

It wasn't more than a week since I was back home, in Nafplion, Greece, to enjoy a – short – month of the summer with friends and family until I had to come back to study fulltime for my September retake exams. A friend, who is also Greek but was studying in the Faculty of Business at the time, called me to deliver the news to me. I was shocked. We both were. A few moments later, we received the first of a plethora of emails to come from UA about what has just happened. When heard that no lives were risked and lost, a feeling of relief washed over us. But we were both far from realising the implications of such an event on the university life as a whole.

A few weeks later, and after the submission of the three assignments I had to retake in the September examination period and, arguably, the hardest and at the same time most rewarding 30 days of studying so far in my academic life, it was finally time for me to focus on finishing up with the internship. A few models were yet to be run and evaluated. "Shouldn't take long", I said to myself. The problem was that, given the fire and the consequences brought along with it, many from the Faculty of Arts staff were relocated to Building L where CLiPS, among a couple other research centres, holds its offices. I am sure that this was a heavy burden for the CLiPS staff to carry. They had to restructure their offices, find new spots for people to work in, and coordinate multiple of staff in a not-so-large space. Although the COVID-19 pandemic had already taught all of us the possibility of working from home, it was evident that certain procedures (I'd personally say "traditions" – why not?), such as the Lab meetings on Tuesdays could not take place entirely online.

As it was to be expected, priority was given to the UA staff. This meant for us, interns, that we would need to move to working from home entirely. Both my supervisors, Prof. Dr. Daelemans, and Jens, were already very flexible when it came to working from home. If I let them know on time, I could do my research, write my scripts, and train my models from the comfort of my dorm. They were also very eager to help me in every step of the way, especially when I asked them for an extension due to some mental and – minor – health challenges.

Apart from all these, my supervisors (or *mentors*, as the university calls them, and it's accurate) were willing to give me access to the CLiPS hardware, one of its servers, in particular. This would enable my models to run faster and take a heavy load from my laptop computer. Prior to this, I would leave my laptop running a Grid Search Cross-Validation for hours, even days. This not only made my laptop unusable while training, as the CPU was constantly hitting 100% usage, but it would also make it very noisy, to the point that I started sleeping in my earbuds. Although the connection via SSH solved the problem, it was only temporary. For some reason, inexplainable to this very day, I'm not able to establish an SSH connection since the early days of October. I tried connecting through another laptop, increased the timeout of the connection, checked the credentials, but nothing worked out. Even the person responsible for the remote connections and the maintenance of the servers, Maxime, couldn't provide any help – he had tried everything. It goes without saying that this was a huge setback and is the main reason that put off the finishing touches on the internship.

The positive side

On a more positive note, however, it made me be more cautious when writing code for training models. Since I could not afford running the same model multiple times, for example in cases that the preprocessing was not complete or had forgotten to add the values for hyperparameters in the parameter grid for Cross-Validation, I started thinking ahead and managed to develop a 'sixth sense' on how whether a model would overfit or underfit.

This could not have been possible without other research and studying I was conducting during these past months. I set an objective of spending at least 20 minutes per day in anything related to Machine Learning, NLP and the like. What has given me a great boost in theoretical knowledge is the daily questions posted by bnomial[1]. Its creator, Santiago, is a machine learning enthusiast who was inspired by Wordle but has taken its project to a whole different direction. He posts a machine-learning-related question every day along with four answers. The number of correct answers varies. It is possible that only one answer is right, whereas it's been times that all four are correct (yes, he's evil sometimes). To top it all off, bnomial provides not deep explanations on why these answers are correct but also why the wrong answers the user selected are not! This makes learning engaging and the inclusion of references to quality websites such as TowardsDataScience, MachineLearningMastery, and his own blog and YouTube channel (funnily called "Underfitted") as well as relevant literature – albeit rarer – has been a top-notch source of inspiration, knowledge, and fun.

Another aspect that I set as an objective to achieve given the chance of this internship is to master git. "Get good at git" is an underestimated skill of today's software engineers and programmers in general. Git is a great tool that allows us to keep track of the changes we make in our codebases in small ("atomic") incremental steps. GitHub, which is my git service of choice, not only provides an easy way of sharing code with colleagues and the general public but is also a great way to build a programmer portfolio for the future. I use GitHub every day for personal and university projects, as well as for (my part) at work. Not only that, but I signed up for its

[1] https://today.bnomial.com.

Education package[2] that comes with all sorts of perks, such as 3-month free access to Datacamp, Pro features on GitHub, access to Microsoft Learn and Azure, and so much more. In these past months I dedicated a lot of my free time to diving deeper and deeper into the machine learning world and the outcomes are only positive and encouraging for the future. In addition, I started experimenting with Python for scripting. I have built a handful of projects – mostly personal – with type-annotated code and testing using **mypy**[3] and **pytest**[4], respectively. I also started adhering to Python's PEP8 code style and built a pre-commit plugin that walks in the literature directory of my thesis repo, identifies all references in my Markdown notes and progressively replaces in-text citations with inline links to the literature PDFs! Finally, I got into Streamlit[5], a library for developing web applications that run Python code and are hosted on the cloud for free (with some basic limitations), and, in a matter of just a few hours, I was able to make a Confusion Matrix Generator[6]! If you were to ask me whether I could manage this a few months back, let alone 14 months earlier when I started this Master's programme, I would have probably laughed and made fun of you.

Takeaways

All these past months were a total rollercoaster for me. With ups and downs, failures and successes, exit codes of 1s and 0s, it was, in a nutshell, a wild ride. In the next few lines, I will try to summarise the essence of what this internship taught me and where I could improve upon.

It is true that had I been more organized in terms of my daily schedule and focused less on work, I could have done even more for CLiPS. I take solace in thinking that I learned so much more than I was expecting, and this internship helped me not only score my highest grades ever during the September examination period, but it also helped me find what I truly want to do after this master's degree: where most of my fellow students want to either change careers or look for employment, I cannot look over my passion for research. After my graduation, I would

[2] https://education.github.com.
[3] http://mypy-lang.org.
[4] https://docs.pytest.org/en.
[5] https://streamlit.io.
[6] https://confusion-matrix-generator.streamlit.app.

like to follow a PhD and contribute to the ever-growing NLP field. Seeing AI application pop up everywhere every day, and the need for people who not only understand and are creative but also ethical about it, I hope that I can play my part in shaping the future of how we understand, measure, and generate language, our most powerful code as human beings to date. With the lessons I learned during this internship, and of course this master's programme overall, I found out things about myself; how I work best, how I do research, and primarily how to overcome challenges, both those that require addressing an error in code and those that require taking a step back, appreciating life and moving forward with optimism and esteem rather than fear.

I would like to personally thank Prof. Dr. Walter Daelemans and Jens Lemmes for their immense support during both the internship and this master's, for being there when I needed a break and hence more time to finish up my tasks due to personal matters. I am sincerely looking forward to working with them in the not too distant future!