Mini-project Technical Report

Introduction to Data Science

Bengs, Pulli, Strang

25.10.2024

Online deliverables available at:
https://github.com/eevib/introDataMiniProject

In this mini project our group produced a pipeline to analyse user comment sentiment on Helsingin Sanomat news articles. Our goal was to improve understanding of commentary on online articles on a major news outlet for researchers, publishers and the general public. Mini-project material and canvas can be found on the deliverables GitHub-page (link above).

The pipeline collects article identifiers, fetches the articles' comments from an open HS.fi API, preprocesses and analyses the data. Key step in preprocessing is to anonymize the collected data to ensure GPRD compliance. The data analysis utilizes a native Finnish language machine learning model to assess user comment sentiment. Additionally, we analyse distribution of user gender among the commenters.

Data collection

Our group wanted to study specifically Finnish language comments. We chose Helsingin Sanomat (later HS), a prominent Finnish news outlet as our data source. On their online news platform HS.fi, some comment section requires users to use their real name to comment on articles. We decided to use comments from the Politics-section (https://www.hs.fi/politiikka/), since this is one of the sections that require real names for commenting. Additionally, we expected articles on political topics to generate more than average interaction.

Comments data was collected using an open API. Inspecting the requests sent by a user's browser revealed a publicly available API endpoint for article data, that included every comment for the article as well. To collect comments from the API, we needed an adequate number of article identifiers. We used Selenium with a Chrome WebDriver to load the website. This was required as the page loads dynamically and not all IDs are visible initially. For the data used by this project, full page loading was done manually. In the project finalization phase the system was improved to be capable of automatically loading the dynamic content.

HS.fi user agreement doesn’t allow scraping their website using automated tools without permission. At the start of this project, after we had come up with our topic, we sent an email to HS.fi customer service to request permission for gathering a small amount of data. Unfortunately, we haven’t received any answer from them so far. As a result, we decided to continue by using the manual approach, and kept our data gathering as small as possible as to not stress the servers or cause any other inconvenience to the newspaper.

When the entire page was loaded, BeatifulSoup [1] was used to extract article identifiers from the rendered page. We conducted this data collection two times, with a week interval to get enough article IDs for our analysis. With the acquired identifiers, we called the API and received comments and related metadata such as names, number of votes and time of commenting. This data was saved in JSON format for processing.

After gathering the comments, we started by removing the last names of the commenters to avoid saving personal data. We created two Pandas DataFrames from the JSON; one that contained article titles, their IDs, tags and total number of comments. The other DataFrame contained article IDs, comments, votes on the comments, creating time and first names.

Exploratory Data Analysis

We did several plots and graphs to visualize our data. The main idea was to get an overview of the data we had. We wanted to know if we can find any patterns in the comments. We did very basic histograms on the distribution of comments per article, distribution of votes on comments and distribution of comment lengths (in characters). The distribution of votes per comment revealed that most of the comments have only a few votes, however there are some comments with a huge number of votes.

Data preprocessing

To be able to divide the commenters by gender we used the Digital and Population Data Services Agency of Finland’s name statistics data provided by Avoindata.fi [2]. We combined the lists of male and female names into one list and labelled the commenters as male or female depending on which gender had more persons with that first name. The few names with a 50-50 distribution of name holders were just labelled as NaNs.

We had around 300 comments where gender couldn’t be specified, the most common reason was that the commenters had mixed up the first name and last name fields, another reason was that the name was so rare that it wasn’t included in the name list (under 5 with that name in Finland). Since this number was relatively small, we didn’t include any logic to figure out whether the order of the names was wrong. This could be implemented with relative ease if the proportion was significant.

Overall, we gathered 211 articles, where 208 had some comments. The total number of comments were 7973 and we used 7673 in our analysis. The difference is due to comments, where we didn't manage to label the gender of the commenter. Additionally, one comment was left out due to its length being too long for the classifier. The lack of gender identification also dropped the number of used articles to 205, which does not compromise our analysis.

Our initial preprocessing pipeline included lowercasing all the words, removing punctuation, <br> and <br/> tags, stop words, using Snowball stemmer for stemming. After conduction some testing with the classifier, we noticed that it was sensitive to the text format and thus all, but the tag removal was commented out in the final product. These steps were, however, left as an alternative pipeline.

Machine learning

We used FinBERT-model [3] for classification of the comments into neutral, positive and negative ones. FinBERT is a Finnish language sentiment analysis model published in March 2023, trained on the FinnSentiment [4] dataset. FinnSentiment introduced a 27,000-sentence dataset that was manually and independently annotated by three native annotators. The model labels each comment string input as positive, negative or neutral with a certain confidence value.

We tested the FinnSentiment model and realized, as mentioned before, it worked better on less pre-processed comments, so in the end we only removed the <br> tags from our comments, before giving the comments to the model to be analysed. For example, the usage of many exclamation marks (!) and upper-cased words made the feeling stronger, and the classifier was more certain about the classification being negative.

Results

We wanted to keep our visualization of the results simple and clear. Therefore, we used only basic histograms, pie chart and a boxplot. We focused on labelling the data and making the graphs as clear and informative as possible.

Chart of the gender distribution of comments reveals somewhat surprisingly, that 77.40% of the commenters were male and only 22.60% were female. It would be very interesting to get some confirmed data on the gender of the readers.

Histogram on the distribution of comment sentiments shows that most of the comments are neutral (70 %), some are negative (25 %) and only a small fraction is positive (5 %). Boxplot on distribution of votes in different sentiments indicates that neutral comments tend to gain least interaction and negative have on average slightly more than positive. Interestingly, neutral and negative comments have several extreme outliers.

Communication of the results

A blog post seemed as the most suitable forum to communicate our results for our target group (media houses, researchers and the general public). We added the most important graphs and results to the blog post without going into too much technical details. Since the code and the results were all in the same place, we thought that the verification of results and further research for any interested parties is going to be easy.

Thoughts on the process, what worked and what didn't?

We found it challenging to decide the topic for our project. We had a lot of different ideas from a very wide range of topics. After thinking back and forth about different topics we were all happy with what we chose. Analysing news article comments was something that none of us had done and we were sure we would learn something new. After maybe a bit slow start we got everything rolling and rest of the project went very well. We valued face to face meetings highly and had one to two meetings every week throughout the course.

One challenge was working with Finnish text. There weren't that many tools available, and it required quite a lot of research to find suitable solutions for Finnish. We considered translating the text to English before conduction analysis on it. Luckily, we found a native tool for Finnish language and didn't have to take that route, as it could have been problematic and could have decreased the reliability of our results, since the translation algorithms might struggle to preserve sentiment over meaning.

We should have made our topic and desired output clearer from the start and we could have planned the work a bit better at the start. One reason we didn't do this was we didn't really know how to do the things we had planned to do and therefore the scope changed a bit during the project.

Adjusting the scope of the project during the process worked very well. We were able to entertain options for further developments, choose ones that were possible to complete in the time available and let go of those that were interesting but tangential to our main line of inquiry. At the end, the workload and progress felt to be in a good balance, and we are very happy with the results and outcome of the project.

The group work worked very well. We found it easy to work together and could utilise each other's strengths.

Learning outcomes

No one of us had experience with string processing or working with texts, except for the weekly assignments in this course, so we learned a lot about that. We found it a good task to plan and execute a small-scale data science project. It was interesting that the project scope was very wide, and we could basically do anything. Thus, it was also a good opportunity to learn about decision making in the context of a group project. We were able to explore many alternatives and settle on an interesting topic without getting lost in the number of choices available.

We also had very little experience in extracting data from online sources and transforming it into a machine learning pipeline. In person meetings were particularly useful for figuring out how to do each of the steps along the way, from loading dynamic content with Selenium, extract data from the page, gather and preprocess the raw data used later on in the process. Same was true for finding out solutions for our machine learning system, exploring combinations of translation, analysis and native solutions.

Finally, it was fun and rewarding to see that the sentiment analyser worked surprisingly well.

Future steps

It would be very interesting to conduct a similar analysis using a larger and more complete set of data from HS.fi. Are users similarly active commenters in other topics as well? Do they have similar phenomenon in comment and vote distribution? Are some topics dominated by either male or female audiences? Are men more active commenters in general or only when talking about politics? Comparison between different news outlets would be interesting as well, although this might be a more difficult task since many sites don't require the commenters to use their real names.

It would be nice to do a website that would show recent statistics of the comments, however, this would need permission from Helsingin Sanomat.

References

[1] BeatifulSoup 4.12.0 Documentation. Available at: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[2] Avoindata.fi. *Etunimitilasto 2024-08-05*. Available at: https://www.avoindata.fi/data/fi/dataset/none/resource/08c89936-a230-42e9-a9fc-288632e234f5.

[3] FinBERT fine-tuned with the FinnSentiment dataset. Available at: https://huggingface.co/fergusq/finbert-finnsentiment.

[4]: K. Lindén, T. Jauhiainen, S. Harwick. *FinnSentiment – A Finnish Social Media Corpus for Sentiment Polarity Annotation*, 2023. Available at: https://arxiv.org/pdf/2012.02613.