Project Report: Whatsapp Lens

Yahya Jabary, 11912007

Code: `https://github.com/sueszli/whatsapp-lens/`

Contents

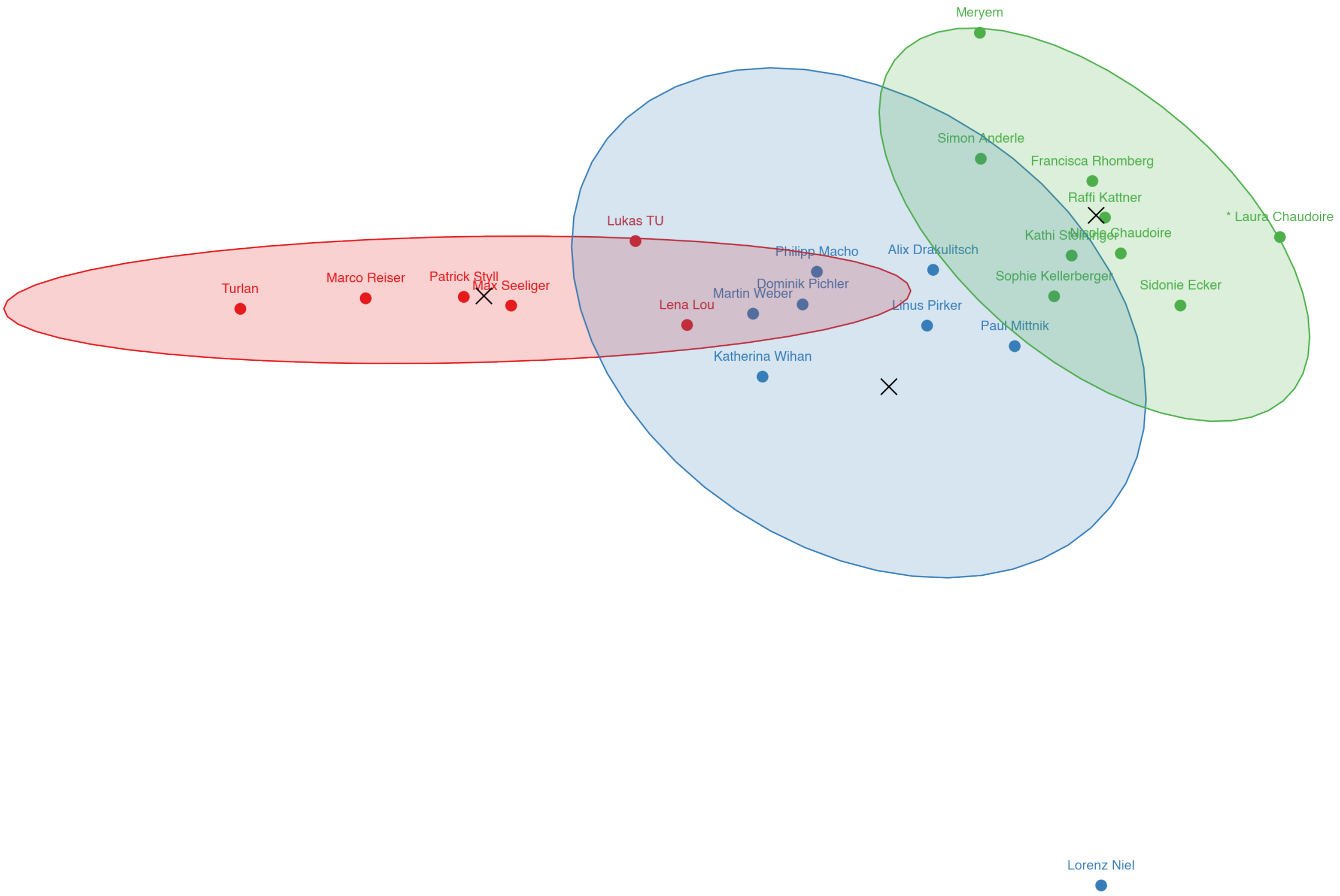

Figure 1: Latent Representation Clustering of WhatsApp Conversations via PCA

Proposal

Assignment 1 - Initiate

In this chapter we discuss the state-of-the-art in WhatsApp analytics, the motivation behind the project, the contribution of the project and a detailed plan of execution with a time estimate for each task.

Motivation

With over 2 billion users WhatsApp has reached the highest user penetration in the European Union[1] [2] [3] and has become the primary means of Internet communication in regions including the Americas, the Indian subcontinent, and large parts of Europe and Africa[4].

This ubuiquity has made WhatsApp a rich source of data for social and behavioral analysis. For instance, Kaushal and Chadha[5] have surveyed various sentiment analysis techniques of WhatsApp, while Ngaleka, Uys[6] have studied m-learning and recently, as of 2023 Jannah [7] has explored the utilization of "WhatsApp Business", a Facebook-Business driven platform for customer service and marketing purposes.

However, these studies are mostly of academic interest and lack both practicality and accessibility for the average user in order to gain data driven insights from their own WhatsApp conversations. This is particularly relevant in the context of WhatsApp Business which is increasingly used for customer service and marketing purposes and where customer interactions are a valuable source of feedback and could be quantified and analyzed to improve business processes.

Analyzing WhatsApp conversations can provide valuable insights into:

- Sentiment patterns: The emotional tone of conversations by using sentiment classifiers.
- Temporal patterns: Message frequency, response times, interaction patterns and engagement metrics between participants as an indicator of relationship strength.
- Topic modeling: Extracting underlying themes in conversations to understand how discussion topics evolve over time using techniques such as Latent Dirichlet Allocation (LDA) or more modern models capable of capturing context for Q&A tasks.

Especially given recent advancements in natural language processing with the advent of transformer models, downstream tasks could also use latent representations of the chat data to train models to both classify and cluster relationships based on conversational patterns, as well as predict future interactions with high accuracy.

Contribution

One of the main challenges in applying the mentioned state-of-the-art NLP techniques to this domain however is the lack of both open datasets and accessible parsing tools for WhatsApp chat exports, given that the data comes in an unstructured and proprietary format, increasing the barrier to entry for researchers and practitioners alike.

The project category "Bring your own data" in this course lend themselves itself nicely to bridging this gap by providing a ground for the development of an end-to-end prototype for WhatsApp chat data analysis:

- (1) Data Generator: implementing synthetic data generation tools, a robustness testing dataset and a personal dataset for demonstrative purposes.
  - (1.1) Synthetic dataset: a synthetic conversational data generation tool, based on role-playing large language model agents.
  - (1.2) Robustness dataset: a custom dataset, demonstrating all edge cases and possible media to be handled by the tool.
  - (1.3) Personal dataset: a private dataset consisting of the author's own WhatsApp chat data for demonstrative purposes, not to be shared in the open-source repository.
- (2) Parser: parsing the proprietary format into a standardized CSV format using regular expressions.
- (3) Feature Engineering: validating, preprocessing and extracting key features from the data.
  - (3.1) Validation: ensuring the data is correctly parsed and standardized, including message timestamp normalization, emoji encoding, media message handling, language detection and processing.

[1] Shan, S. The battle between social giants: WeChat and WhatsApp's influence in digital marketing.

[2] WhatsApp-Website. http://blog.whatsapp.com/. Accessed 30 Oct 2024.

[3] Montag, C., Błaszkiewicz, K., Sariyska, R., Lachmann, B., Andone, I., Trendafilov, B. & Markowetz, A. (2015). Smartphone usage in the 21st century: who is active on WhatsApp?. BMC research notes, 8, 1-6.

[4] Metz, Cade (April 5, 2016). "Forget Apple vs. the FBI: WhatsApp Just Switched on Encryption for a Billion People". Wired. ISSN 1059-1028. Archived from the original on April 9, 2017. Retrieved May 13, 2016.

[5] Kaushal, R., & Chadha, R. (2023, March). A Survey of Various Sentiment Analysis Techniques of Whatsapp. In 2023 2nd International Conference for Innovation in Technology (INOCON) (pp. 1-6). IEEE.

[6] Ngaleka, A., & Uys, W. (2013, June). M-learning with whatsapp: A conversation analysis. In International Conference on e-Learning (p. 282). Academic Conferences International Limited.

[7] Jannah, R. (2023). Utilization of whatsapp business in marketing strategy to increase the number of sales through direct interaction with customers. Syntax Idea, 5(4), 488-495.

  - (3.2) Feature Extraction: extracting key features such as message frequency metrics, response time patterns, sentiment scores, topic distributions and user interaction patterns.

- (4) Analytics: performing exploratory data analysis and applying state-of-the-art NLP techniques to the data.
  - (4.1) Sentiment Analysis: training a sentiment classifier on the data to predict the emotional tone of conversations.
  - (4.2) Topic Modeling: applying LDA or more modern models to extract underlying themes in the data.
  - (4.3) Relationship Clustering: clustering relationships based on conversational patterns using latent representations of the chat data.
  - (4.4) Frequency Analysis: analyzing message frequency, response times, interaction patterns and engagement metrics between participants.
  - (4.5) Demographic Analysis: predicting user demographics based on conversational patterns.
- (5) Deployment: deploying the analytics results in a user-friendly and interactive web application.
- (6) Presentation: preparing a final report and a video presentation on YouTube.

Given our time constraint of 3 ECTS credits, equivalent to 75-81 hours of work (25-27 hours per ECTS credit), excluding the time spent on optional lectures and exercises, we constrain each task to a maximum of 15 working hours, ensuring that the project is completed within the time frame.

Execution

Assignment 2 - Hacking

Results

Assignment 3 - Deliver