Project Report: Whatsapp Lens

Jabary Yahya (11912007)

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

Proposal

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).

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 lends itself nicely to bridging this gap by providing both:

- (1) an open dataset consisting of organic chat data from volunteers, as well as synthetic data which encorporates all possible message format for robustness testing, and
- (2) a parsing tool capable of extracting and standardizing the data from the proprietary format into a structured format that can be used for further analysis
- (3) a simple demo of potential downstream tasks that can be performed on the data for demonstrative purposes.

The data preprocessing will encompass the standardization of the data, including message timestamp normalization, emoji encoding, media message handling, language detection and processing. The feature engineering can include extracting key features such as message frequency metrics, response time patterns, sentiment scores, topic distributions and user interaction

[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.

patterns. And finally, the demo will showcase the metrics and a few selected downstream tasks that can be performed on the data.

Given the explorative nature of the project, the extent to which the optional tasks will be implemented will depend on the time and complexity constraints of the project. However, the deliverables will include a well-documented codebase, a detailed report as well as a presentation of the findings in less than 100 hours of work.