Web-Based Annotation Tool for Instant Messaging Conversations

Jaromír Plhák, Michaela Lebedíková, Michał Tkaczyk, and David Šmahel

Faculty of Informatics, Masaryk University, Brno, Czech Republic
{xplhak, x450458, x245062, davs}@fi.muni.cz

Abstract. This paper presents a customized web-based annotation tool that allows users to annotate utterances in data from instant messaging applications. Efficient annotations are provided by a well-arranged user interface, operating using key-pressing and integration of an interactive annotation manual. Moreover, the interface for supervisors allows them to determine which utterances belong to the gold standard. We also provide information on two accomplished annotation tasks: annotating online risk phenomena with sparse occurrence (0.85% to 1.98%) and annotating social support that can be used to generate efficient detection models.

Keywords: Text annotation tool, online risky behavior, social support.

1 Introduction

Annotation tools are beneficial when developers create deep-learning NLP models, as they require a lot of high-quality data to make accurate predictions. Labeling of this training data is usually provided by human users who have good expertise in the target domain. This task is frequently very effort-intensive and time-consuming. Therefore, annotation tools should allow users to simplify the annotation process to improve their productivity and ensure data coherence and inter-annotator agreement (IAA).

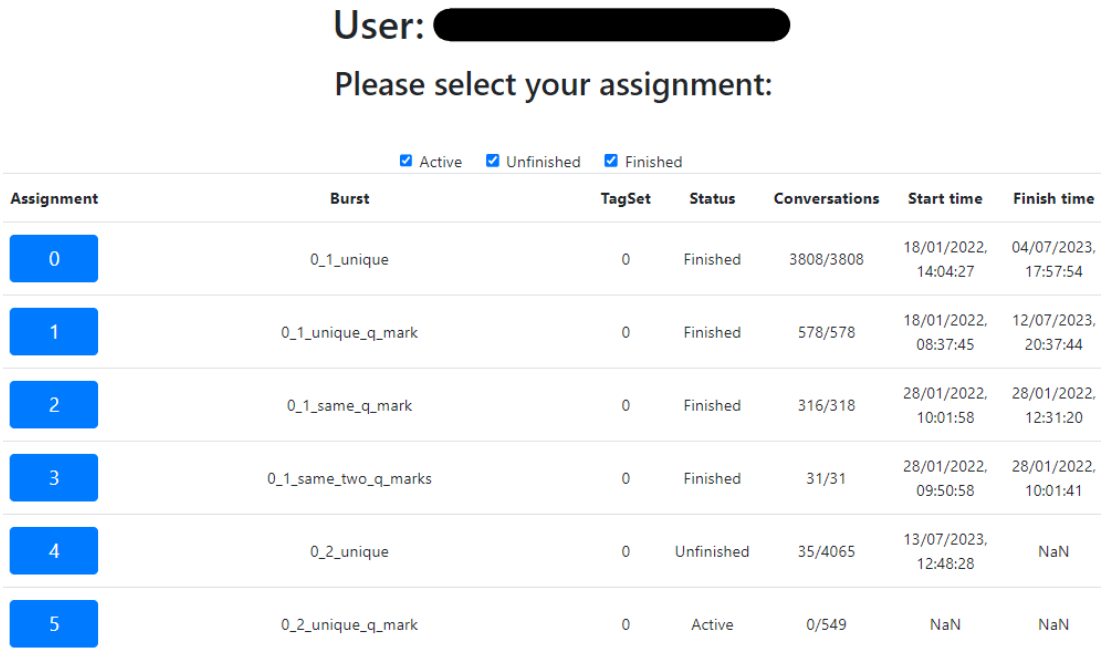
Many text annotation tools are available for various text annotation tasks [1,3,4,5,8,9,10,11,17]. They allow the users to annotate words, sentences, and other text parts using specified tags and also label their relations and dependencies. Such tools also usually provide work distribution among team members and different user roles like annotators or supervisors. They also offer various levels of security, including role-based access, zero data sharing, or multi-factor authentication. Moreover, they provide functionalities like integration with external resources, annotation comparison, IAA calculation, or AI-assisted annotation.

However, as each annotation tool is developed for a given task or set of tasks, it is hard to use them for processing specific types of data. Within the project Modeling the future: Understanding the impact of technology on adolescent's well-being (FUTURE) [2], we acquired data from adolescents' Messenger and

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023, pp. 3–10, 2023.

WhatsApp conversations to generate efficient online risk and social support detection models [12,14,15]. This task presented several challenges that we needed to overcome: (1) we needed to anonymize the data, (2) we needed to parse exported data to suitable units of text that would be large enough to convey the meaning of the conversation and its topic, (3) the task of annotating complete utterances representing one user prompt. Taken together, the current solutions provided by other annotation tools have proven ineffective, time-consuming, and unable to meet our demands due to their complex GUI and general functionality.

Therefore, we designed and implemented a new web-based tool for annotating instant messaging (IM) conversations called **IRTIS Annotation Tool (IRTIS AT)**. Unlike most existing solutions, our tool allows for annotating complete utterances representing one user prompt, regardless of whether this prompt is one word or multiple sentences. Also, our solution relies on using key-pressing instead of a computer mouse to fasten the annotation process.



Assignment	Burst	TagSet	Status	Conversations	Start time	Finish time
0	0_1_unique	0	Finished	3808/3808	18/01/2022, 14:04:27	04/07/2023, 17:57:54
1	0_1_unique_q_mark	0	Finished	578/578	18/01/2022, 08:37:45	12/07/2023, 20:37:44
2	0_1_same_q_mark	0	Finished	316/318	28/01/2022, 10:01:58	28/01/2022, 12:31:20
3	0_1_same_two_q_marks	0	Finished	31/31	28/01/2022, 09:50:58	28/01/2022, 10:01:41
4	0_2_unique	0	Unfinished	35/4065	13/07/2023, 12:48:28	NaN
5	0_2_unique_q_mark	0	Active	0/549	NaN	NaN

Fig. 1: Interface for annotation burst selection

2 IRTIS AT

The IRTIS AT allows users to annotate texts from online communicators like Messenger or WhatsApp. It processes files that can be manually exported from these apps via their export functionality [6,7]. During import to IRTIS AT, data are anonymized based on the algorithm presented in [16]. Subsequently, it

provides users with two basic interfaces: for annotators and for supervisors, where they check annotation disagreements. Annotator's interface allows the users to annotate utterances using specific tags from the tagset (predefined set of applicable tags). The example of the interface for annotation burst selection can be seen in Figure 1 and the interface for supervisor's annotations revision in Figure 2.

Potential users of the IRTIS AT are researchers who want to annotate data from online messaging applications in a well-arranged and intuitive way. It supports the annotators and supervisors, who can examine the amount of finished work and time spent on annotations.

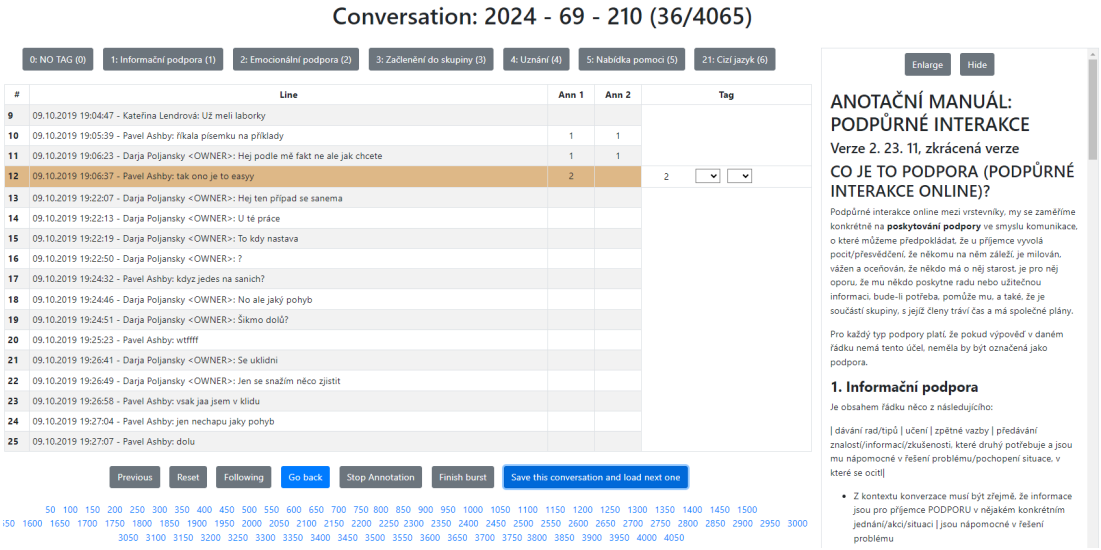


Fig. 2: Interface for supervisor's annotations revision

2.1 Functionality and Development Process

Initially, the following functionality for the IRTIS AT was discussed with supervisors and fine-tuned after testing the prototype:

- Data can be uploaded to the server, anonymized, and divided into conversations spaced by periods of non-communication longer than 60 minutes (currently done outside the tool).
- Tagsets can be specified and include selected tags (currently done manually using .json configuration files).
- Conversations are grouped in a series of messages called bursts. Bursts contain a limited number of conversations to prevent annotator fatigue (currently done outside the tool).
- Annotators have accounts and can be assigned to selected bursts (currently done manually using a .json configuration file).

- Annotators can choose the burst (from the assigned bursts) they want to work on.
- Conversations can be annotated by annotators using tags from tagset:
  - Tags can be selected using key pressing.
  - Importantly, annotators can select up to three additional tags, if applicable. For example, adding a T+ symbol means that the annotator assumes another online risk or social support category can be used for a given utterance. Moreover, as our annotation task was very complex, annotators sometimes were unsure whether a tag should be applied, especially in the training phase. Our tool allows them to express this uncertainty by using a question mark, and these can be later viewed by the annotation supervisors and fine-tuned.
  - Previous and subsequent conversations can be loaded on demand to assess the context of the conversations.
  - Annotation manual is integrated into GUI and allows annotators to access requested parts interactively.
- Supervisor mode allows supervisors to decide ambiguous cases where the annotators disagree with the tag.
- Annotated data can be exported into MS Excel sheet or .csv files for further processing and IAA calculation.
- Statistics about work progress will be displayed in the GUI (number of finished conversations, total number of conversations in the burst, tagset, starting and finishing times).
- Time spent on the annotation is logged for each annotator and burst.

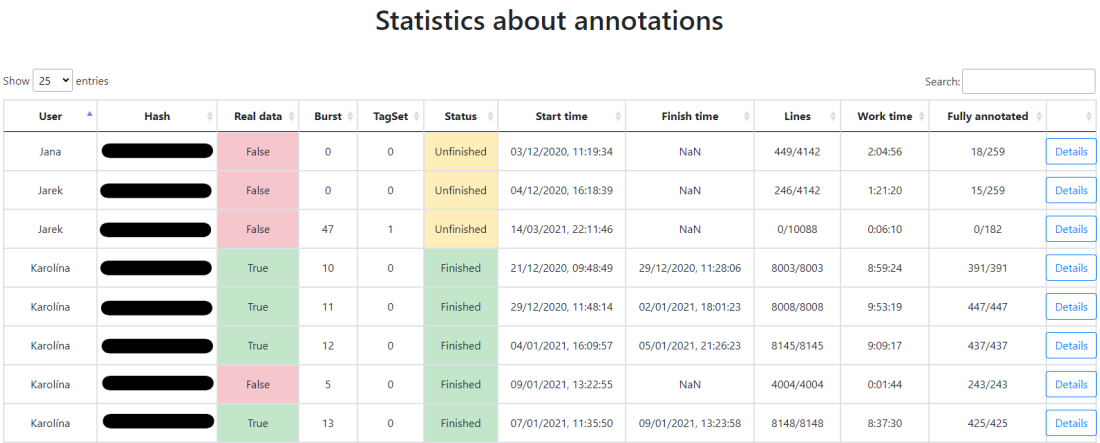


Fig. 3: Annotation statistics for supervisors

IRTIS AT was implemented using Python/flask technology and deployed on the server within Masaryk University. The source code is available in the Gitlab repository [13].

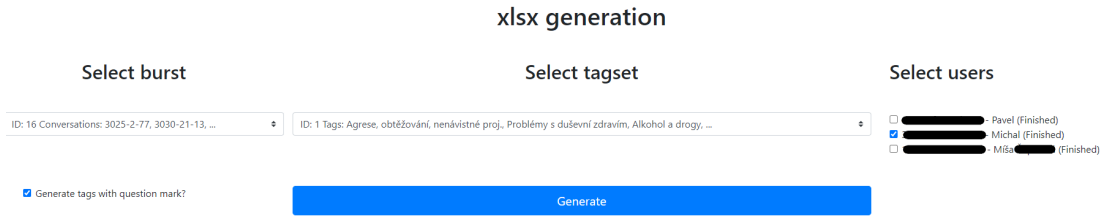


Fig. 4: Interface for data export to MS Excel

2.2 Annotation Process

Before the IRTIS AT's implementation, annotation manuals were developed for **Online risky behavior** (aggression, harassment, hate; mental health; use of alcohol and drugs; sexual content and sexting), and provision of **Social support** (informational support; emotional support; social companionship; appraisal; and instrumental support). For each task, two annotators were trained for two months, and the manual was gradually refined based on their and supervisors' feedback. Finally, the annotators started to code randomly generated bursts of data. The occurrence of the category online risky behavior in our corpus was sparse (see Table 1). Therefore, as we need as many positive examples as possible, we developed a preliminary classifier to identify conversations with a higher chance of containing utterances. It was then used to generate bursts for annotation of this specific category.

In the next step, the gold standard was generated. A dedicated supervisor interface was designed and developed as one of the tool components. Using this, supervisors could solve disagreements between annotators and utterances with ambiguous tagging (see Figure 2).

The annotation component of the tool comprises two windows. The initial window offers an overview of batches that were assigned to the annotator. The following information is provided to each batch: ID number, name, ID of specific tagset, status, number of annotated units/number of all units in batch, starting time, and finish time (see Figure 1). After selecting a batch, the window for annotation opens (see Figure 2). Information about the annotated unit is displayed in the upper part of the annotation window. Below this information, virtual keys for each tag are displayed. Each key contains the name of the tag and key bindings. The annotation manual is displayed on the right side of the screen; it can be scrolled down or enlarged. The annotation user interface is displayed in the center part of the window. It comprises lines to be annotated and tags assigned to each line by the annotator. In the bottom part of the window, virtual functional keys are located. Those keys enable moving backward and forward across larger units (conversations), enlarging the number of rows displayed in the annotation user interface if a broader context is needed to decide, finish, and save the annotation.

The annotation process resulted in 272,465 utterances with online risky behavior tagset (ORB) and 196,772 with social support tagset (SocS). To detect the difficulty of annotating each tagset, we compare the time spent by every

Table 1: Overview of the number of annotated utterances and IAA (Cohen's κ)

Category	**Tagset**	**Annotated by at least one annotator**	κ
Aggression, harassment, hate	ORB	5393 (1.98%)	.470
Mental health problems	ORB	3101 (1.14%)	.460
Alcohol, drugs	ORB	2301 (0.85%)	.609
Sexual content	ORB	3550 (1.30%)	.485
Informational support	SocS	9967 (5.07%)	.685
Emotional support	SocS	9669 (4.92%)	.639
Social companionship	SocS	3331 (1.70%)	.604
Appraisal	SocS	2524 (1.28%)	.650
Instrumental support	SocS	5317 (2.70%)	.599

annotator. Table 2 shows that the annotation time differs between two annotators by approximately 15%, and annotating social support was approximately three times slower than annotating online risky behavior. This is based on a higher density of social support categories in corpora, and, therefore, it requires more cognitive effort to evaluate that the social support category does not occur on a given line.

Table 2: Times of annotations for selected bursts with 64,452 utterances.

Annotator's id	**Tagset**	**Total time (h:mm:ss)**	**Per utterance**
1	SocS	68:47:33	3.84 sec.
2	SocS	77:13:49	4.31 sec.
3	ORB	25:51:13	1.44 sec.
4	ORB	22:27:41	1.25 sec.

3 Limitations and Future Work

IRTIS AT comes with its limitations. First, similarly to other tools, we developed IRTIS AT for a specific task that has arisen. While our tool can be adapted to other tasks that include instant messaging or conversational data, our solution may be unsuitable for other tasks, such as annotating medical records.

Second, in the current version, some functionality has to be solved manually or outside the tool (uploading and anonymizing data, generating bursts for annotation, specifying tags and tagset, detecting disagreements between annotators). In addition, the application lacks more sophisticated authentication. Therefore, future versions should include user accounting with roles (annotator, supervisor, and data manager), store data directly in the database instead of .json files, and allow users to automatically divide and anonymize uploaded data.

Additional functionality that has to be included is automatic IAA calculation and creating bursts of conversations for annotation based on specified rules (e.g., according to the number of lines or conversations) using GUI. Also, the specific burst for supervisors could be available directly in the tool (e.g., all conversations where the concrete annotator puts a question mark as an additional tag).

4 Conclusions

In this paper, we presented a user-friendly annotation tool that allows users to annotate texts from online communicators like Messenger or WhatsApp efficiently. The effectivity is ensured, e.g., due to a well-arranged user interface or involving key pressing to fast annotation of utterances. It also gives annotators and supervisors valuable feedback about how many annotations were done and left, as well as the time spent annotating given bursts of data. The tool has been used in practice for annotating social support and risky behavior in anonymized data of adolescents with sufficient results [14]. Such data can be practically usable in many applications like chat-bots or parental control applications provided by social networking sites.

Acknowledgements. This work has received funding from the Czech Science Foundation, project no. 19-27828X.

References

1. Cejuela, J.M., McQuilton, P., Ponting, L., Marygold, S., Stefancsik, R., Millburn, G., Rost, B.: tagtog: interactive and text-mining-assisted annotation of gene mentions in plos full-text articles. Database : the journal of biological databases and curation **2014**, bau033 (01 2014). https://doi.org/10.1093/database/bau033
2. Elavsky, S., Blahošová, J., Lebedíková, M., Tkaczyk, M., Tancos, M., Plhák, J., Sotolář, O., Šmahel, D., et al.: Researching the links between smartphone behavior and adolescent well-being with the future-wp4 (modeling the future: understanding the impact of technology on adolescent's well-being work package 4) project: protocol for an ecological momentary assessment study. JMIR Research Protocols **11**(3), e35984 (2022)
3. HumanSignal, Inc.: Label and annotate data. `https://labelstud.io/` (2023), [Online; accessed 25-October-2023]
4. John Snow Labs Inc.: NLP Annotation Lab: Free No Code AI Platform. `https://nlp.johnsnowlabs.com/` (2023), [Online; accessed 25-October-2023]
5. Labelbox, Inc.: Labelbox | Data-centric AI Platform for Building & Using AI. `https://labelbox.com/` (2023), [Online; accessed 25-October-2023]
6. Meta Platforms, Inc.: Facebook: Download a copy of your information on Facebook. `https://www.facebook.com/help/212802592074644` (2023), [Online; accessed 25-October-2023]
7. Meta Platforms, Inc.: WhatsApp: How to save your chat history - Help center. `https://faq.whatsapp.com/1180414079177245` (2023), [Online; accessed 25-October-2023]

8. Montani, I., Honnibal, M.: Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models, https://prodi.gy/
9. Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., Liang, X.: doccano: Text annotation tool for human (2018), https://github.com/doccano/doccano, software available from https://github.com/doccano/doccano
10. Neves, M., Seva, J.: Annotationsaurus: a searchable directory of annotation tools. arXiv preprint arXiv:2010.06251 (2020)
11. Perry, T.: Lighttag: Text annotation platform (2021)
12. Plhák, J., Tkaczyk, M., Lebedíková, M.: Detecting online risks and supportive interaction in instant messenger conversations using czech transformers. (2021)
13. Plhák, J.: Irtis annotation tool. https://gitlab.fi.muni.cz/irtis/irtis-annotator (2023)
14. Plhák, J., Sotolář, O., Lebedíková, M., Šmahel, D.: Classification of adolescents' risky behavior in instant messaging conversations. In: International Conference on Artificial Intelligence and Statistics. pp. 2390–2404. PMLR (2023)
15. Sotolář, O., Plhák, J., Lebedíková, M., Tkaczyk, M., Šmahel, D.: Constructing datasets from dialogue data. RASLAN 2022 Recent Advances in Slavonic Natural Language Processing p. 131 (2022)
16. Sotolář, O., Plhák, J., Šmahel, D.: Towards personal data anonymization for social messaging. In: International Conference on Text, Speech, and Dialogue. pp. 281–292. Springer (2021)
17. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: brat: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107. Association for Computational Linguistics, Avignon, France (Apr 2012), https://aclanthology.org/E12-2021