Appendix J. A Complete List of Message Metadata

We compile a complete list of metadata fields from the documentation of Telethon library [45], and provide why (or why not) we keep them.

Metadata necessary for extracting conversations.

1) `id`: The ID of this message, unique within a group; We keep it to fetch any media data this message attached.
2) `peer_id`: The peer to which this message was sent, in our case, this is the group ID; We keep it to separate messages from different group during conversation construction.
3) `date`: The $UTC + 0$ datetime object indicating when this message was sent; We keep it to properly order messages.
4) `message`: The string text of the message; We keep it as the message content of any conversation.
5) `from_id`: The peer who sent this message; We replace it with a hash value as the sender identifier.
6) `via_bot_id`: The ID of the bot used to send this message through its inline mode; We replace it with the value -1 as the sender identifier to indicate that this is a bot message, not from a real user.
7) `reply_to`: The original reply header if this message is replying to another; We use it to construct a conversation between any two users in a group.
8) `media`: The media sent with this message if any; We keep it and download the media file to construct the dataset.
9) `reactions`: Reactions to this message; We keep it as a potential way to expand conversation length for future research.

Ignoring metadata irrelevant to our research.

1) `out`: Whether the message is outgoing or incoming; Our test account never interacts within any group chat, all messages are incoming.
2) `media_unread`: Whether a user has read the media in this message or not; Our test account never interacts with any message except from API calls, hence, irrelevant.
3) `post`: Whether this message is a post in a broadcast channel or not; We fetch messages from groups, this field is for broadcast channels.
4) `views`: The number of views this message from a broadcast channel has; We are using group chats, not broadcast channels.
5) `post_author`: The display name of the message sender to show in messages sent to broadcast channels; We are not using broadcast channels.

Ignoring undefined metadata.

1) `from_boosts_applied` Not mentioned in the documentation.
2) `saved_peer_id` Not mentioned in the documentation.
3) `via_business_bot_id` Not mentioned in the documentation.
4) `quick_reply_shortcut_id` Not mentioned in the documentation.
5) `effect` Not mentioned in the documentation.
6) `factcheck` Not mentioned in the documentation.
7) `legacy`: Whether this is a legacy message or not; Unclear what is the definition of a legacy message, hence, we ignore this field.

Ignoring platform-specific metadata.

1) `mentioned`: Whether a user was mentioned in this message or not; This is not a criterion of including or excluding a message in our dataset, hence, irrelevant.
2) `slient`: Whether the message should notify people with sound or not; This is a customized setting for Telegram Messenger, which may or may not present in other messengers, hence, we do not focus on it.
3) `from_schedule`: Whether this message was originated from a previously-scheduled message or not; This is a customized setting for Telegram Messenger, which may or may not present in other messengers, hence, we do not focus on it.
4) `edit_hide`: Whether the edited mark of this message is edited should be hidden or shown; This is a customized setting for Telegram Messenger, which may or may not present in other messengers, hence, we do not focus on it.
5) `pinned`: Whether this message is currently pinned or not; This is a specific feature for groups which is not usually present in individual chats, in addition, it is irrelevant to the process of conversation extraction.
6) `noforwards`: Whether this message can be forwarded or not; This is a customized setting for Telegram Messenger, which may or may not present in other messengers, hence, we do not focus on it.
7) `invert_media`: Whether the media in this message should be inverted; This is a customized setting for Telegram Messenger, which may or may not present in other messengers, hence, we do not focus on it.
8) `offline`: Whether the message was sent by an implicit action, for example, as an away or a greeting business message, or as a scheduled message; This is a customized setting for Telegram Messenger, which may or may not present in other messengers, hence, we do not focus on it.
9) `fwd_from`: The original forward header if this message is a forward; The original message may or may not present in our conversation dataset, which makes it unclear how to maintain the forward relationship, hence, we ignore this field.
10) `reply_markup`: The reply markup for this message (which was sent either via a bot or by a bot); We are not interested in the formatting of bot messages, since we focus on messages from real users.
11) `entities` The list of markup entities in this message, such as bold, italics, code, hyperlinks, etc.;
12) `forwards`: The number of times this message has

been forwarded; A duplication of `fwd_from`.

13) `replies`: The number of times another message has replied to this message; We do not need to keep tract of this number, and we can reconstruct this counter from messages in our dataset if needed.
14) `edit_date`: The date when this message was last edited; This is a customized setting for Telegram Messenger, which may or may not present in other messengers, hence, we do not focus on it.
15) `grouped_id`: If this message belongs to a group of messages (photo albums or video albums), all of them will have the same value here; We do not see this field with any actual value, hence, it is either very rarely used or not accessible.
16) `restriction_reason`: An optional list of reasons why this message was restricted; This is a customized setting for Telegram Messenger, which may or may not present in other messengers, hence, we do not focus on it.
17) `ttl_period`: The Time To Live period configured for this message; This is a customized setting for Telegram Messenger, which may or may not present in other messengers, hence, we do not focus on it.
18) `action`: The message action object of the message for; This is a customized setting for Telegram Messenger, which may or may not present in other messengers, hence, we do not focus on it.