From Online Opinion Shifts to Market Reactions: A Data Pipeline for Messaging Platforms

Genady Kogan*, Natalia Vanetik*, Sven Nõmm**
*Shamoon College of Engineering, Beersheba, Israel
**Department of Software Science, Tallinn University of Technology, Tallinn, Estonia

Motivation

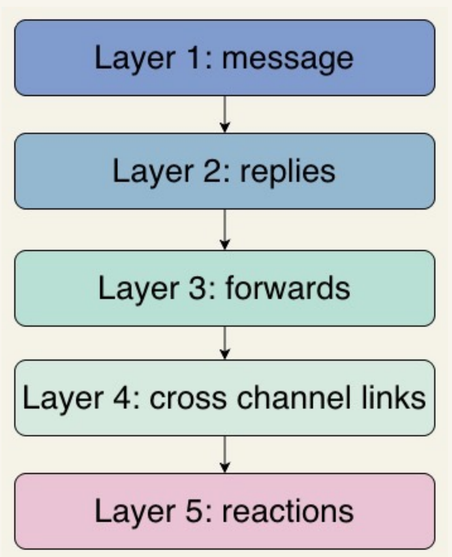
- Telegram is a major source of narrative spread and coordinated messaging.
- Existing datasets are small, fragmented or unstructured..
- We build a large scale multilingual dataset capturing message flows and interactions.
- The goal is to study how online communities form, evolve and react to information flows

Dataset overview

- More than 152,000 public Telegram channels collected since 2015.
- Over 200 GB of multilingual text: Arabic, English, Russian, Hebrew and others.
- Includes posts, replies, forwards, edits, reactions and message metadata..
- Preserves relational structure: threads, interactions and cross channel links.
- Covers broad regional and topical communities over a ten year span.

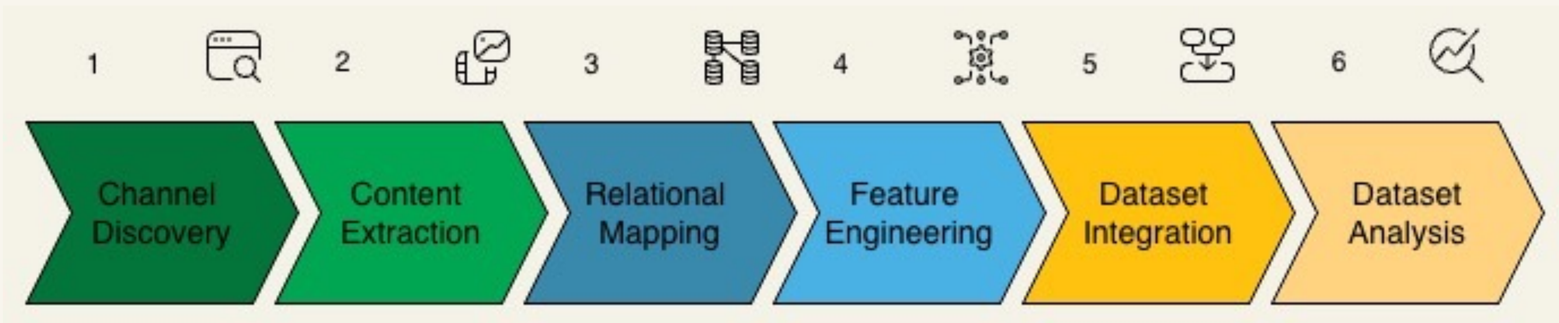
Multi layer message structure

```
extracted = {
    'id': message.id,
    'date': message.date.strftime('%d-%m-%Y %H:%M') if message.date else None,
    'message': message.message,
    'silent': message.silent,
    'pinned': message.pinned,
    'via_bot_id': message.via_bot_id,
    'reply_to': reply_to_data,
    'views': message.views,
    'forwards': message.forwards,
    'edit_date': message.edit_date.strftime('%d-%m-%Y %H:%M') if message.edit_date else None,
    'replies': {'summary': replies_data, 'comments_data': None},
    'reactions': reaction_data,
    'forwarded_from': forwarded_data,
}
```



Processing Pipeline

Data collection and analysis pipeline



Analytical subset

- The analysis is based on a selected subset of large channels with more than 1000 subscribers, containing 139 channels and more than 10 million messages.
- This focused subset offers stable activity patterns and clear community level behavior suitable for temporal and interaction analysis.

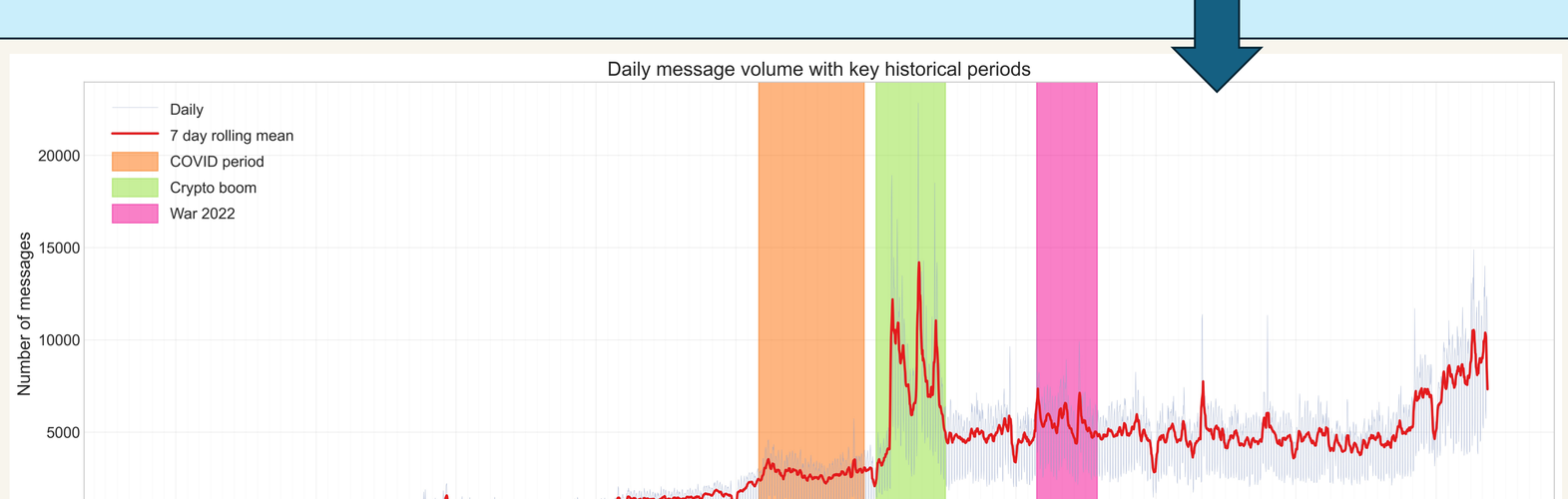
9.3 million posts	335 million word tokens
1.08 million comments	Average message length of 32.3 tokens
134,352 unique users	time span 2015 to 2025

Text analysis

We compute several standard lexical richness measures across all messages:

- Type Token Ratio (**TTR**): 0.0075
- Corrected Type Token Ratio (**CTTR**): 97.56
- Root Type Token Ratio (**RTTR**): 137.97
- Measure of Textual Lexical Diversity (**MLTD**): 87.24

The combination of low TTR with high MTLD indicates a large and varied vocabulary across the corpus, consistent with long time periods and many contributors.



Zero Shot Classification

	category	mean_score	count	percent
1	news	0.90	7245	50.66%
2	technology	0.87	1943	13.59%
3	finance	0.83	1414	9.89%
4	crypto	0.94	1385	9.68%
5	other	0.74	643	4.50%
6	politics	0.90	584	4.08%
7	promotion	0.81	573	4.01%
8	scam	0.90	266	1.86%
9	war	0.88	204	1.43%
10	propaganda	0.91	45	0.31%

- Data consists of the first 100 messages from 139 Telegram channels with over 1000 subscribers.
- Zero-shot classification was performed using the facebook/bart-large-mnli transformer model.

Key findings

- Large channels display stable lexical patterns across years.
- Strong bursts of replies correlate with narrative coordination.
- Forward chains reveal hidden cross community bridges.
- Multilingual channels act as hubs for cross regional opinion flow

Conclusion

- Large scale multilingual dataset from 152k public Telegram channels.
- Structured layers: posts, replies, forwards, interactions.
- Pipeline enables NLP, behavioral and event study analysis.
- Supports testing links between online opinion shifts and market reactions.

Contact Information

G. Kogan genadko@ac.sce.ac.il
N. Vanetik natalyav@sce.ac.il
S. Nõmm sven.nomm@taltech.ee