

Social Media als Quellengattung: Crawl, Archivierung und Vorverarbeitung von Telegram-Channels

Epistemische und kuratorische Perspektiven

Simon Donig, Ole Meiners, Cristina Vertan

Frankfurt, Main 19./20. März 2024

DOI: [10.5281/zenodo.10885634](https://doi.org/10.5281/zenodo.10885634)



HERDER-INSTITUT
für historische Ostmitteleuropaforschung
INSTITUT DER LEIBNIZ-GEMEINSCHAFT

Aufbau der heutigen Präsentation

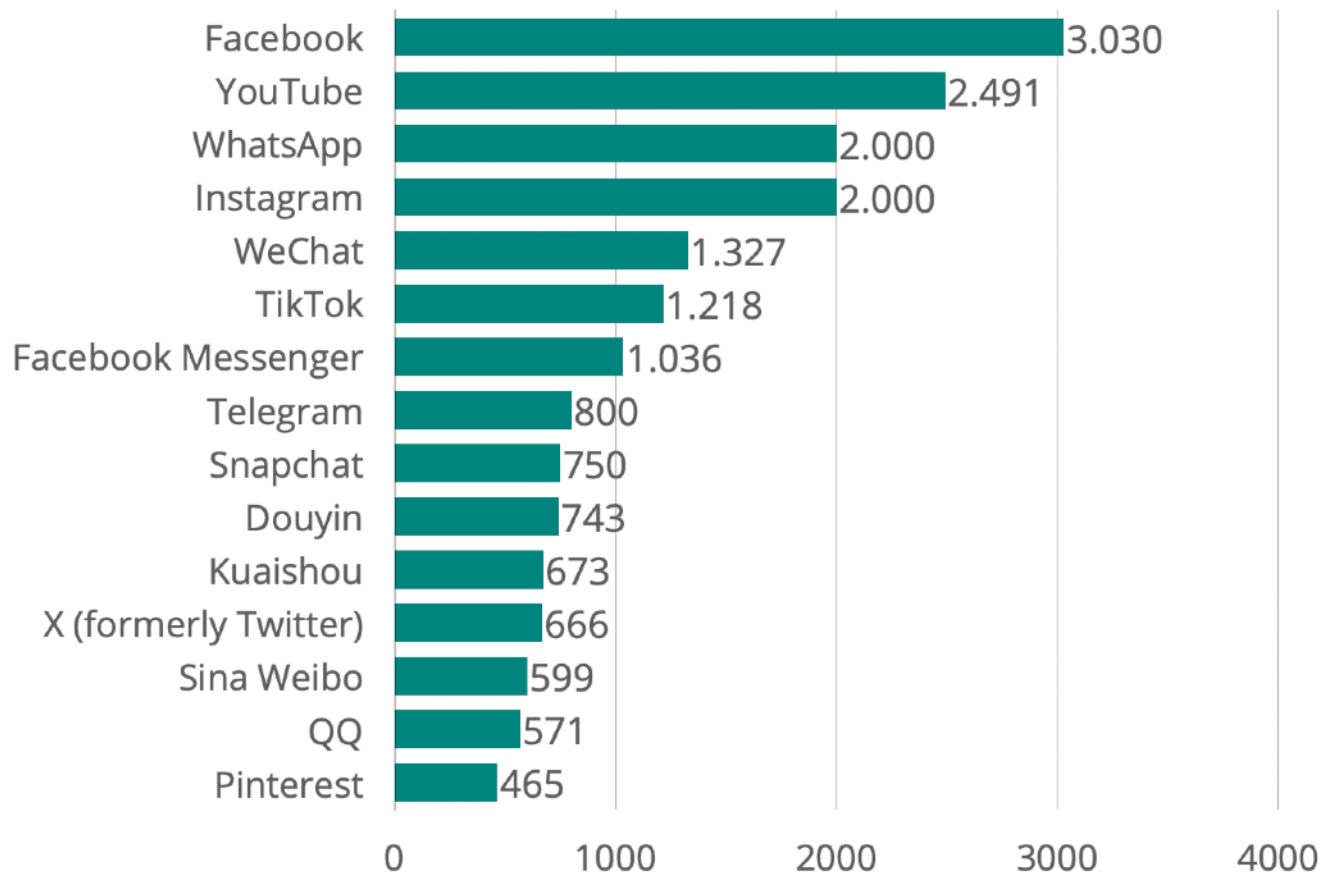
- Motivation: Eine datengetriebene Geschichte von Verschwörungserzählungen online
- Struktur und Funktionalität des Messengers Telegram
- Telegram-Inhalte als Quellengattung
- Datenakquise: Eigener Crawler und Datenformat(e)
- Untersuchungsmethode: Textmining & Netzwerkanalyse
- Implikationen für Daten-Kuratierung und Archivierung
- Datenethische Fragen / epistemologische Grenzen

Forschungsinteresse: Transnationale Verschwörungsnarrative in Massenquellen

- Epistemologische Frage nach der Bedeutung neuer Archive (**Realtime & Living Archives** innerhalb d. Zeithorizonts). **Data-driven/-centered History**?
- Fokus auf Untersuchung von **Verschwörungserzählungen** in transnationalen Online-Öffentlichkeiten
- “Biolabs-in-der-Ukraine” Narrativ - **Ursprünge** und **Konjunktoren** auf der Online-Plattform Telegram?
- Welche **Akteure** sind zentral für die Verbreitung des Narrativs?
- Wo überschreitet das Narrativ **Sprachgrenzen** und welche Online-**Communities** sind seine hauptsächlichen Träger?

Weltweit beliebteste Social Media Plattformen

Oktober 2023, nach Statista, gerankt nach monthly active users (MAUs) in Mio.



<https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>

* WhatsApp has not published data in the past year, so this figure is less reliable

** Figure uses daily active users—monthly active user count is likely higher

Struktur und Funktionalität der Messenger App Telegram

Was ist Telegram?

- Messenger App ähnlich WhatsApp, Signal, Threema
- **Nutzung vorrangig über Smartphones** (Webapp ebenfalls verfügbar)
- **offene, dokumentierte API**. Erstellung von Apps/Bots unterstützt
- Funktionsweise:
 - Einzelne **Konversationen** und geschlossene Gruppen.
 - Sog. **Supergroups**, (bereits 2015 eingeführt, inzwischen bis zu 200.000 User) selten öffentlich
 - **Channels** (Broadcasting-Gruppen: Posten nur für engen Userkreis, reguläre User (= Subscriber) können mit Emojis reagieren oder in eigenen **Comment-Bäume** posten. Beschränkungen hinsichtlich Anzahl Subscriber, oftmals öffentlich. Ähnliche Nutzung wie andere Social Media-Plattformen möglich (v. a. Twitter). Erreichbarkeit großer Zielgruppen.

Struktur und Funktionalität der Messenger App Telegram

Wer nutzt Telegram und wie?

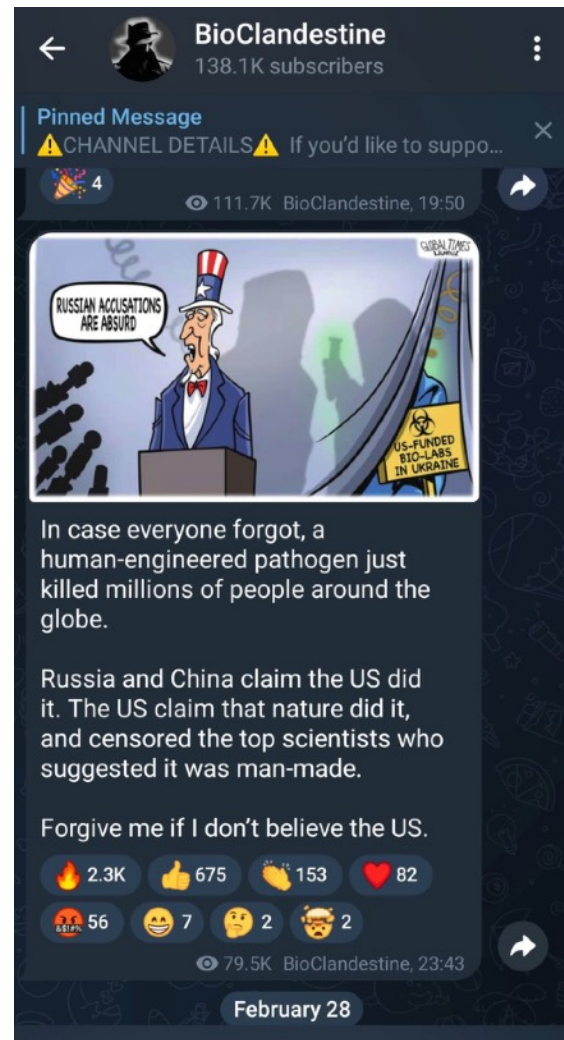
- In einigen **Ländern/Regionen/Communities stark verbreitet**. Ähnliche Rolle wie Twitter für viele Institutionen im Westen, z. B. in Russland ("Militärblogger") und Ukraine
- **Keine Moderation** der Inhalte durch Telegram, daher attraktiv für Akteure, die von stärker regulierten Diensten wechseln.
- Möglichkeit eigene **Werbung** auf Kanälen zu schalten und Einkommen zu generieren. (Telegram schaltet ab einer gewissen Gruppengröße eigene Werbung).
- Anders als bis vor kurzem bei Twitter **keine öffentliche Zugänglichkeit** der Channels im WWW.
- Private Bereiche (z.B. geschlossene Gruppen) die wir aus ethischen Gründen nicht crawlen (allenfalls Datenspenden vorstellbar).

Struktur und Funktionalität der Messenger App Telegram

Mediale Charakteristik

- **Keine Beschränkung bei Nachrichtenlänge.** Medien (Audio, Video, Websites, Bilder, Dokumente) können leicht eingebunden werden
- Messages können in andere Channel/Chats geteilt werden; Channel/Chats sind über URLs in Messages verlinkbar
- **User-Interface eines Messengers** -> Kommunikation ähnelt Chats (viele Emojis, eher kurze, stark durch Absätze strukturierte Texte, viele Verweise auf andere Telegram-Channels/Groups und andere Webressourcen)

Screenshot Telegram-Channel
"Bioclandestine", 27.02.2024,
aufgenommen am, 14.03.2024.



Telegram als Quellengattung / theoretischer Zugang

- Einblick in Öffentlichkeiten mit oftmals **transnationalen Bezügen**. Ideal, um die **Viralität von Themen und Inhalten** zu verfolgen.
Kommunikationstheoretisch: spezifische Ausprägungen von **Narrativen**.
- Hochgradig **multimodal**.
- **Situative Kommunikation**, aber durch das Message/Kommentar Schema nur bedingt dialogisch. Theoretisch: **Second Orality** (Walter J. Ong)?
- Verschwörungsnarrative als Fenster zu Teilöffentlichkeiten, die sehr homogen und selbstverstärkend sind (z.B. kaum dissentierende Emojis wie thumbs down). **Dekonstruktion statt Pathologisierung**.
- Erinnerungstheoretisch: Medialisierung von **Gedächtnis** in der **hypervernetzen Welt** (Andrew Hoskins)?

Crawler

Motivation & Ansatz

- **Motivation:** Verbreitung von Narrativen über Telegram-Channels nachverfolgen
- Ansatz: Crawler startet mit 1 bis n **Seed Channels**. Vorab identifiziert, müssen hohe Wahrscheinlichkeit relevanter Inhalte aufweisen.
- Crawler erhält eine Liste mit **Suchwörtern**, durchsucht Seed Channel.
- Bei Match wird Message mit zugehörigen Metadaten ausgegeben, zudem wird auf zwei Arten von **Verknüpfungen zu anderen Channels** geprüft:
 - **forwards:** ist die message aus einem anderen Chat/Channel etc. in den aktuellen Channel weitergeleitet worden?
 - **mentions:** Erwähnungen von Chats/Channels in der Message
- Crawler legt für den weiteren Crawl **Liste der identifizierten Channels** an; zudem werden **Verknüpfungen** gesondert ausgegeben zur Nachverfolgung der Informationsflüsse via Netzwerkanalyse
- Erfolgreich gecrawlte Channel werden mit charakteristischen **Metadaten** gespeichert, u. a. um doppelte Crawls zu vermeiden

Crawler

Technische Umsetzung

- Python Script unter Verwendung der Pyrogram Library für **Zugriffe auf Telegram API** (<https://pyrogram.org/>)
- **Benötigt:** API-ID von Telegram (Mobilfunknummer und Telegram-Account erforderlich). **Kein vollständig 'anonymer' Crawl** möglich. Inwiefern Channel-Owner über API-Zugriff informiert werden oder sich informieren können, ist nicht bekannt.
- **Depth-first vs. breadth-first Strategie:**
 - zunächst rekursiver depth-first-Ansatz verfolgt mit vorgegebener Maximaltiefe.
 - problematisch, da API unvorhersehbar intern Fehler verursacht; Abbruch des rekursiven Crawls ging mit Verlust des gesamten Crawls einher
 - bei größer werdendem Channel-Netzwerk zudem exponentielles Wachstum der pro Run zu crawlenden Channel (bei depth > 1)
 - alternativ breadth-first Ansatz implementiert: depth = Anzahl Runs.
 - **Zwiebel-Prinzip:** Pro Run wird ausgehend von den Seed Channels jeweils ein Verlinkungsschritt weiter gecrawlt
 - **Lückenloses Fortsetzen** abgebrochener Crawls.

Datenerzeugung

Vier Ausgabeformate:

- auf **Channel-Ebene:**

- Message-Objekte (beinhalten Chat-Objekte, etwa bei forwards) unbearbeitet gespeichert
- Zentrale Metadaten zu Chats und Messages
- Forwards & Mentions: Keys der jeweiligen Chats, Message-ID (zukünftig zudem Timestamp)

- auf **Crawl-Ebene:**

- Metadaten zu den erfolgreich gecrawleten Channels (id, username, typ, members_count, is_private)
- Zukünftig: Protokoll zum Crawlvorgang

Rohdatenausgabe (rechts), json (links)

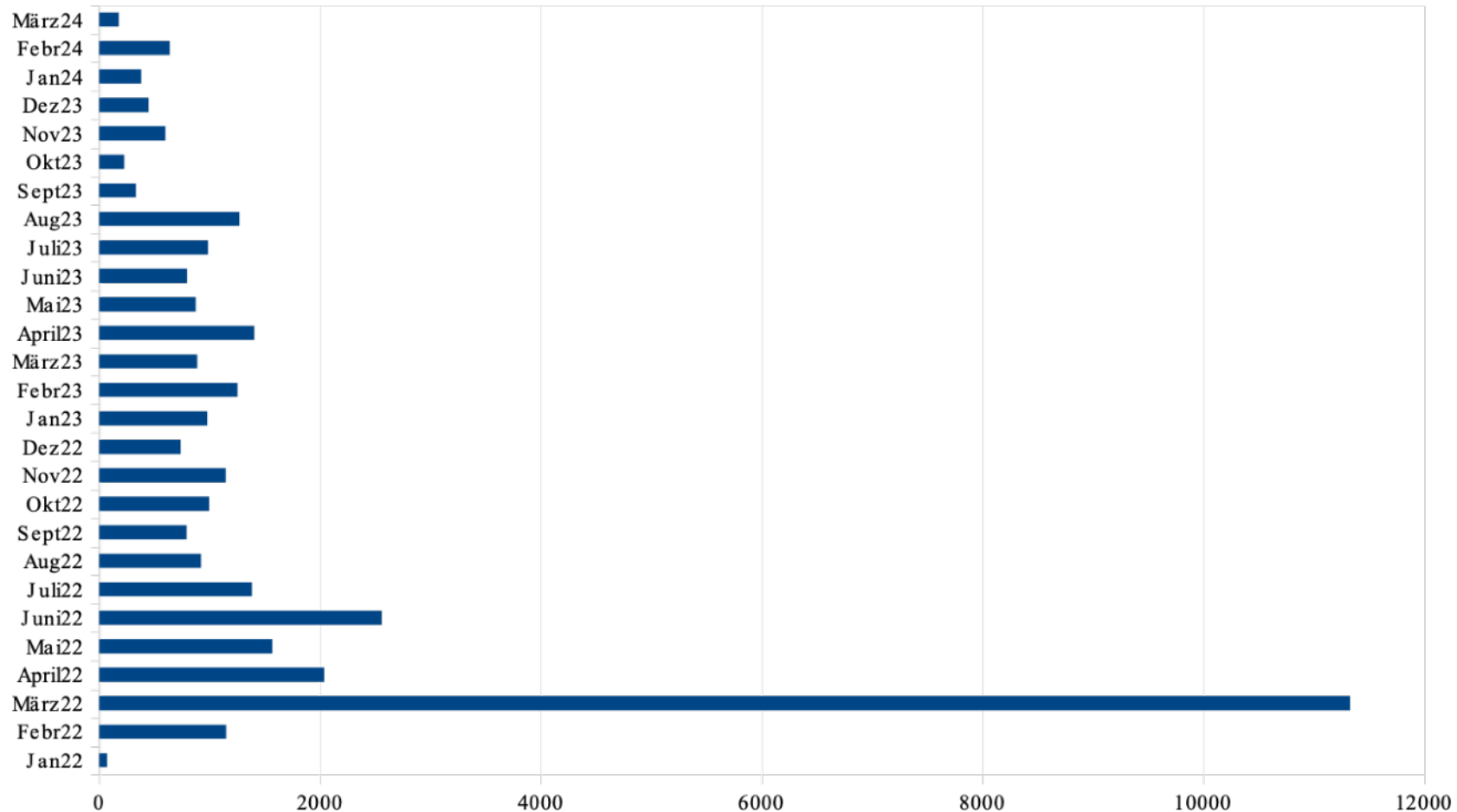
```
"messages": {
  "count": 1
},
{
  "id": 59331,
  "date": "2022-02-27T12:11:11",
  "chat": {
    "id": -1001135812070,
    "username": "unzensiert",
    "type": "ChatType.CHANNEL",
    "members_count": null
  },
  "media": "MessageMediaType.PHOTO",
  "edit_date": "2022-02-27T12:12:19",
  "caption": "@unzensiert_infoseite\\n\\nUkraine - Russische Raketenangriff",
  "text": null,
  "reactions": [
    {
      "emoji": "👍",
      "count": 211
    },
    {
      "emoji": "👎",
      "count": 15
    },
    {
      "emoji": "😬",
      "count": 5
    },
    {
      "emoji": "👊",
      "count": 2
    },
    {
      "emoji": "👏",
      "count": 1
    },
    {
      "emoji": "👉",
      "count": 1
    }
  ]
},
]
```

Message als json-file (gefiltert auf ausgewählte Properties)

```
{
  " ": "Message",
  "id": 59331,
  "sender_chat": {
    " ": "Chat",
    "id": -1001135812070,
    "type": "ChatType.CHANNEL",
    "is_verified": false,
    "is_restricted": true,
    "is_creator": false,
    "is_scam": false,
    "is_fake": false,
    "title": "👉Unzensiert👉",
    "username": "unzensiert",
    "photo": {
      " ": "ChatPhoto",
      "small_file_id": "AQADAgADHtExGx5fCEgAEIAAxAxrSp-cW____cnkXVRdTuKEABB4E",
      "small_photo_unique_id": "AgADHtExGx5fCEg",
      "big_file_id": "AQADAgADHtExGx5fCEgAEIAAxAxrSp-cW____cnkXVRdTuKEABB4E",
      "big_photo_unique_id": "AgADHtExGx5fCEg"
    }
  },
  "dc_id": 2,
  "has_protected_content": false,
  "restrictions": [
    {
      " ": "Restriction",
      "platform": "ios",
      "reason": "appleviolence",
      "text": "Unfortunately, this channel couldn't be displayed on your device."
    },
    {
      " ": "Restriction",
      "platform": "android",
      "reason": "androidterms",
      "text": "Unfortunately, this channel can't be displayed on Telegram apps d"
    }
  ]
},
  "date": "2022-02-27 12:11:11",
  "chat": {
    " ": "Chat",
    "id": -1001135812070,
    "type": "ChatType.CHANNEL",
    "is_verified": false,
    "is_restricted": true,
    "is_creator": false,
    "is_scam": false,
    "is_fake": false
  }
}
```

Message als txt-file (Ausschnitt; alle Properties vorhanden)

Anzahl der Posts pro Monat Jan. 2022 bis März 2024



Korpusbeschreibung

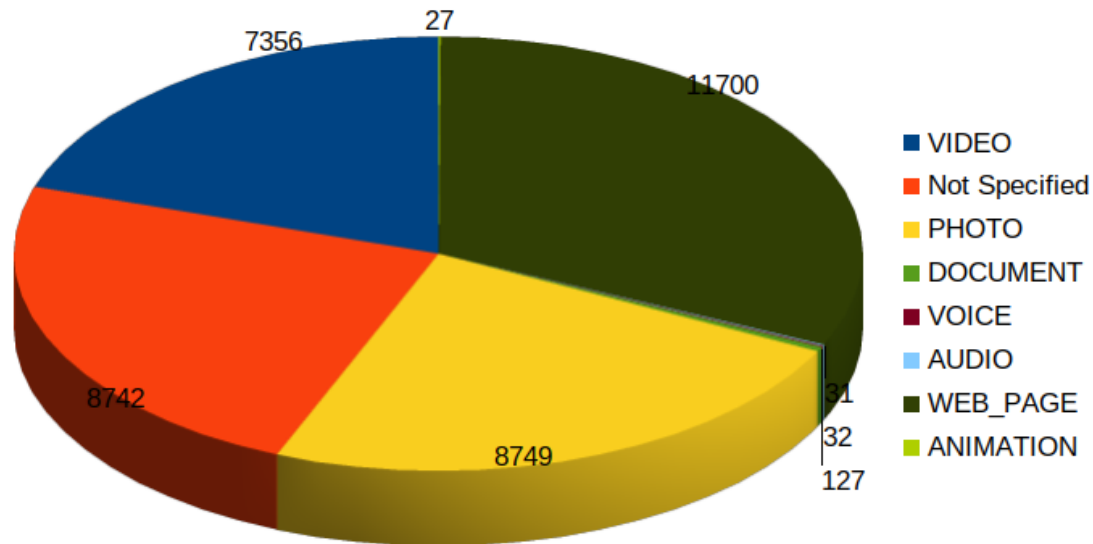
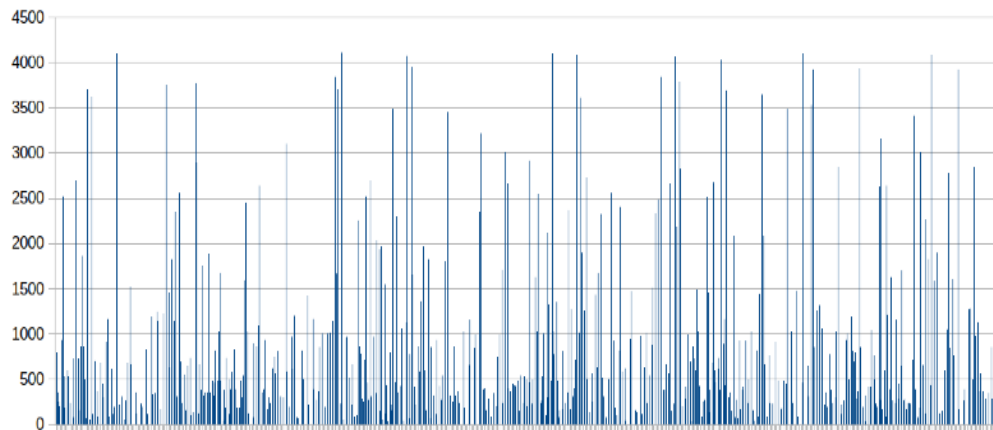
in Zahlen

1031 - channels

36764 - messages

- 5146666 - tokens

- 4791.0 - Average token/messages



- 20 442 Texts (3.924.090 tokens)
- 16 322 Captions (1.222.573 tokens)

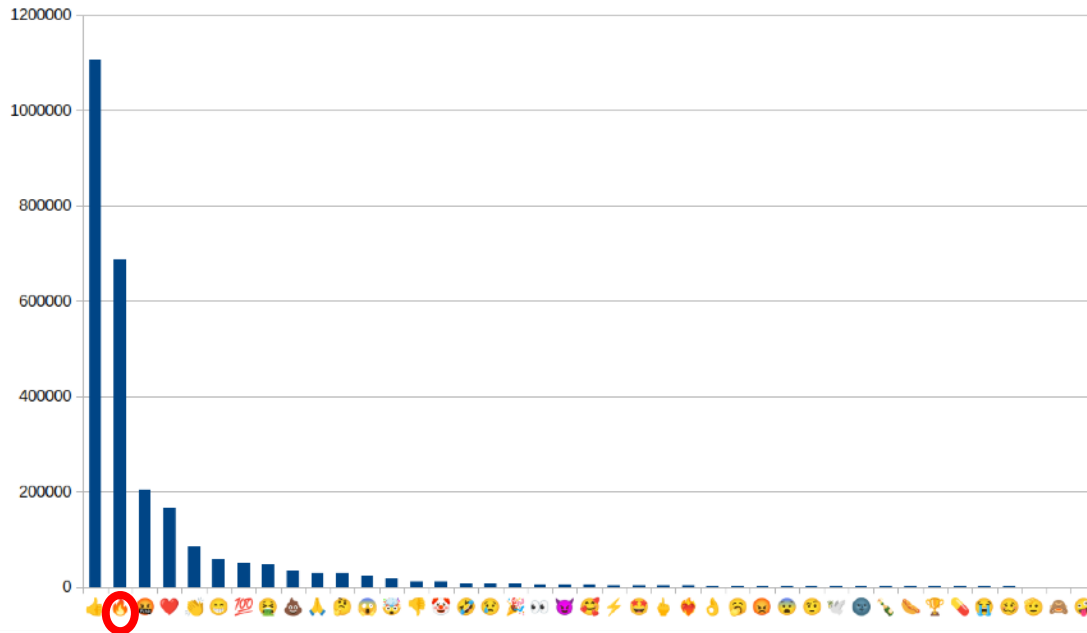
viel Inhalt bleibt verborgen

Korpusanalyse

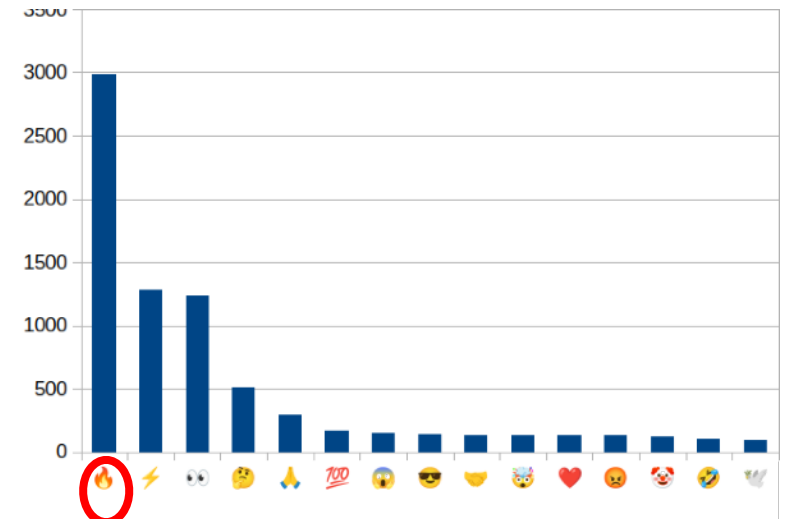
Reaktionen als Hinweis für Nachrichten-Botschaft und Community -Beteiligung

72 unterschiedliche Emojis als direkte Reaktionen

58 davon werden innerhalb von Texten erneut benutzt



HERDER-INSTITUT für historische Ostmitteleuropaforschung - INSTITUT DER LEIBNIZ GEMEINSCHAFT



Korpusanalyse

Vorverarbeitung -Herausforderungen

18.275 Zeilen mit Charakteren
ausserhalb des Lateinisches
Alphabets; meistens Piktogramme
und Emojis

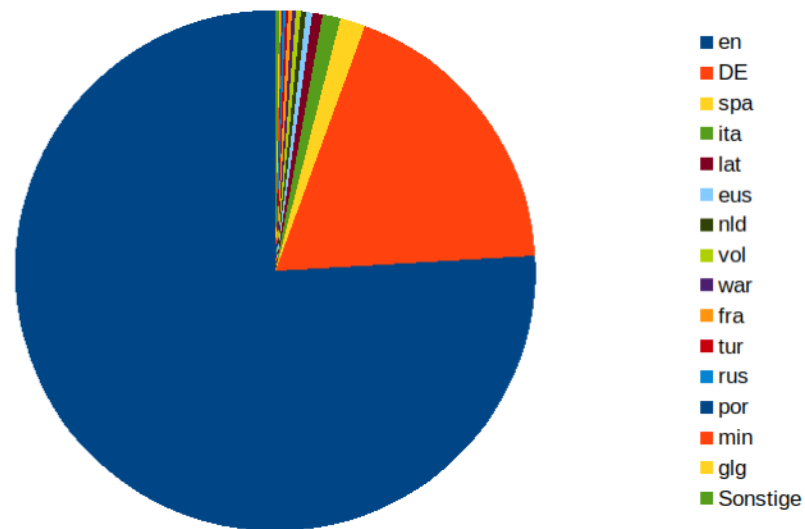
- Löschen → Bedeutungsverlust
- Beibehalten → Negativen Einfluss auf weiteren Tools
- Umschreiben → problematisch (z.B. bei Emojis unterschiedliche Semantik

🇷🇺
<https://nara.getarchive.net/amp/media/us-army-reserve-maj-tulsi-gabbard-2nd-from-left-87b181>
Asow-Stahl),
военно-биологическая
#COVID-19
2018-2020|
🇷🇺🇺🇦
✅🇺🇦 Rosemont
Sinn🤔
<https://rumble.com/vwzpu1-tucker-carlson-why-does-ukraine-have-secret-biolabs-in-ukraine.htmlnull>
«🇺🇸
<https://odysee.com/@RedPill78:e/Big-Wins-For-Truth%2C-Biolabs%2C-Covid%2C-Eelction-Fraud%2C-All-Will-b-r=68MDyxzZTCkVlDSAUWbCqihoikCqrbvC>
full-blown
<https://video.foxnews.com/v/6311272948112#sp=show-clipsnull>
♦ However,
🎵💡🇺🇸
([https://thenationalpulse-com.translate.goog/2022/03/08/obama-led-ukraine-biolab-efforts/?](https://thenationalpulse-com.translate.goog/2022/03/08/obama-led-ukraine-biolab-efforts/?x_tr_sl=auto&x_tr_tl=de&x_tr_hl=de&x_tr_pto=wapp)
x tr sl=auto& x tr tl=de& x tr hl=de& x tr pto=wapp)

Korpusanalyse

Multilingualität

- 41 unterschiedliche Sprachen nur in den Text-Nachrichten (ohne Captions)
- cf. OpenNLP Language Detection Model (<https://opennlp.apache.org/models.html>)
- Primär englische und deutsche Posts
- Ergebnisse müssen unbedingt manuell überprüft werden.



Korpusanalyse

Stopwordlists

- Stopwordlists müssen in einem Vorbereitungsschritt definiert werden
- sie sind Sprachabhängig
- Uni- Bi- und Trigramme (6575 1-2-3gramm Typen) als Hinweis?
 - wie geht man mit Abkürzungen um? (US, Q)
 - bei Trigrammen riskiert man auch bedeutungstragende Wörter zu löschen (z.B. "job")
 - aber auch sehr viele Trigramme mit Piktogrammen, Zahlen.....

ria
rid
rig|
PhD
!
Pi,
t.,
t..
aid
/Es
aim
2
c'è
0
air
Phi
"
al,
al.
Pic
Pig

Korpusanalyse

Weiterer Einsatz von NLP-Werkzeugen kann nur nach einem **Clustering nach Sprache** erfolgen, weil die NLP Tools von Sprachmodellen abhängig sind.

Durch die grosse **Variation von Textlängen** muss ein weiteres Cluster-Verfahren eine Normalisierung hervorbringen.

Die Benutzung von **Methaphern** (besonders für NE) und **Ironie sowie anaphorische und kataphorische Referenzierungen** sollten wenn möglich markiert werden: z.B durch ein Emoji-Tagging

Netzwerkanalyse als Auswertungsverfahren

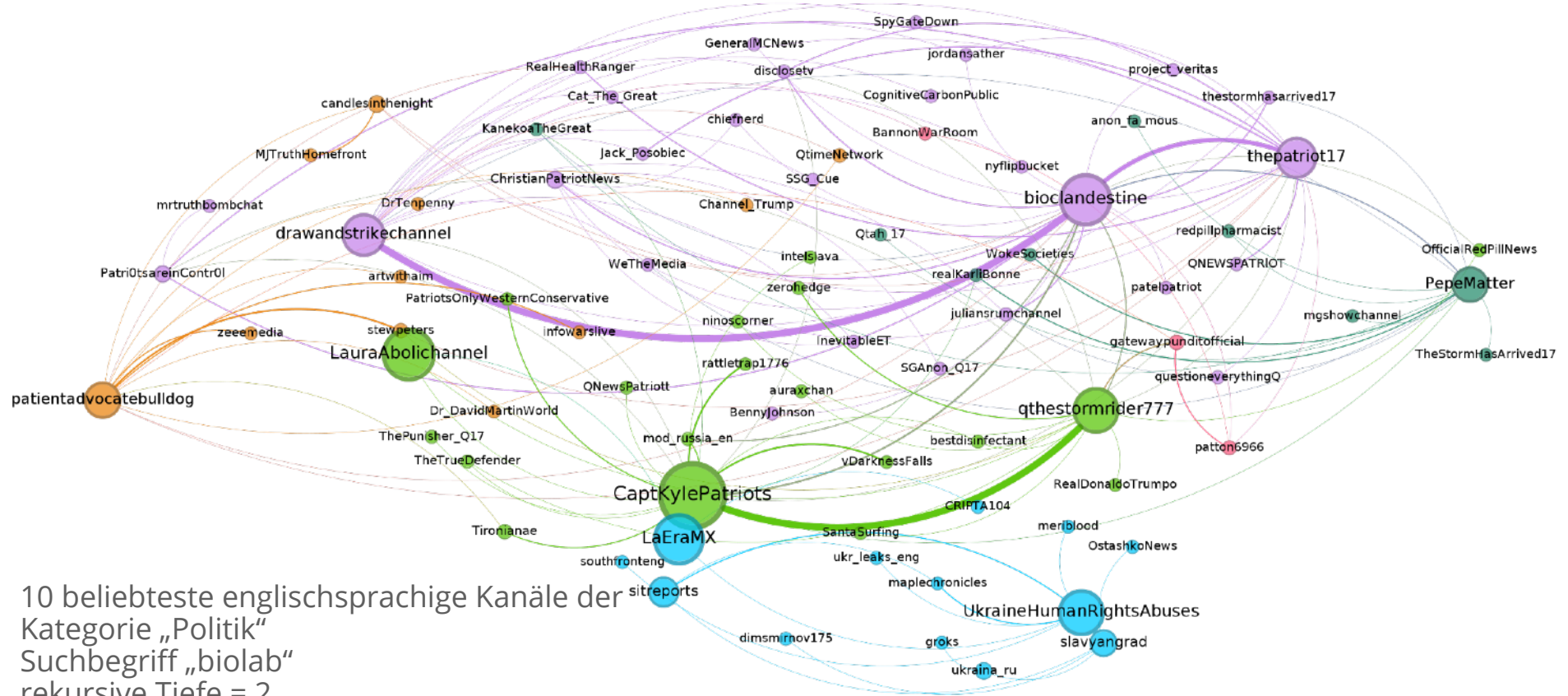
Netzwerkanalyse als ein möglicher **Übergang zwischen Distant und Close Reading**. Z.B. Identifikation zentraler Knoten.

Eigener Erkenntniswert z.B. in der Generierung von Clustern.

Eigene methodische Herausforderungen (z.B. Grad-, Betweenness-Zentralität, Pagerank für „zentrale“ Knoten?)

Bedarf der Erhebung zusätzlicher Merkmale (z.B. Sprache der Posts) um Hypothesen testen zu können. Erfordert **Anreicherung** des Materials **mit abgeleiteten Merkmalen**.

Netzwerkvisualisierung und Clusterbildung



- 10 beliebteste englischsprachige Kanäle der Kategorie „Politik“
- Suchbegriff „biolab“
- rekursive Tiefe = 2
- Knotengröße: Betweenness-Zentralität
- Visualisierung: gephi
- Farben: Communities (Louvain) & OpenOrd Algorithmus für Cluster

Nächste Schritte

- **Weiterentwicklung des Crawler-Prototypen**
 - Ermöglichung komplexerer Such-Methoden: kombinierte Schlagwörter, Regular Expressions?
 - Erstellung einer GUI
- **Alternative Methoden der Speicherung** und Nachnutzung, z.B. Graph-Datenbank
- Erstellung einer **Preprocessing Pipeline**
- Erprobung unterschiedlicher **Topic Modeling Verfahren**
- Umgang mit **multimodalen Inhalten** (im Moment nur Captions ausgewertet)

Datenethische Fragen

How public is a public channel?

- Messenger suggerieren ihren Nutzenden möglicherweise einen eher **privaten Kommunikationsraum** (Chat-Interface, ähnlich privatem Austausch über Messenger)
- Im Unterschied zu twitter etc. verwenden User von Messengern (nicht zwingend die Channel-Betreiber) ihre auch **privat genutzten Konten** (Bindung an oftmals private Mobilfunknummer); Trennung privater/ öffentlicher Raum verschwimmt
- Zugriffs- und Auswertungsmöglichkeiten durch Dritte auf öffentliche Channel (aber auch Supergroups!) über **API mutmaßlich Mehrheit der Nutzenden nicht bewusst**
- **Crawls** finden **ohne Zustimmung** (und i. d. R. ohne **Kenntnis**) der Nutzenden (Channel-Betreiber wie auch Subscriber/User) statt

Datenethische Fragen

Implikationen

- Vielfältig **kontroverse politische Themen** mit starker politischer Agenda verhandelt; **Rechtsverstöße** (Volksverhetzung) nicht unwahrscheinlich. Welche Rolle spielt der **transnationale Kontext**?
- Daten aus **Konflikten und Krisengebieten** könnten von *malicious actors* für data mining genutzt werden.
- **Pseudo- und Anonymisierung**? Aufwendig, da Usernames der Channelbetreiber in der Regel auch in den Messages auftreten.
- Identifikation zentraler Akteure (von **zeitgeschichtlichem Interesse**) erlaubt evtl. Rückverfolgung anderer Akteure mit Recht auf Privatheit.
- Wie mit **Forwards und Mentions** umgehen, die **aus privaten Kanälen** kommen? Wie mit geposteten und verlinkten Medien (Urheberrecht)?
- Global: Braucht es eine **CARE ähnliche Governance** (wer wird dann repräsentiert?) oder decken Beiräte/Ethikräte dies ab? Rolle der NFDI?

Datenethische Fragen

DDOXing von Bioclandestine, einem zentralen Akteur des Biolabs-Narrativs

[About](#) [Research Centers](#) [What We Do](#) [Resources](#) [Take Action](#) [Donate](#) 

BLOG

Unmasking “Clandestine,” the Figure Behind the Viral “Ukrainian Biolab” Conspiracy Theory

↓

RELATED CONTENT

BLOG

Antisemitic Conspiracy Theories Abound...

March 09, 2022





US BIOLABS IN UKRAINE



Exclusive US biolabs in Ukraine, and they are financed at the expense of the US Department of Defense.

The laboratories are located in Odessa, Mariupol, Kherson, and near Crimea and Luhansk. 2 other possible locations: Kyiv and Mykolaiv.

PH 4 AM - PM 30, 2022, Twitter for iPhone

Twitter, 24.2. 2022

Kuratorische Fragen

- Projektbezogene Insellösung vs. **Standardisierung von Crawl-Formaten?**
- Verbesserung der Formalisierung **datenethischer Auszeichnung?**
- Framework für **Plattform-übergreifendes Crawling** (EU erzwingt Öffnung zentraler Plattformen, Austausch nimmt daher perspektivisch zu).
- Welche Gedächtniseinrichtungen wollen was, wann und in welchem Umfang archivieren? (Welches **Mandat** haben, welches Mandat brauchen wir?)
- Ist anlasslose, breite Archivierung (**Heuhaufen**) sinnvoll, skalierbar – und aus datenethischen Überlegungen heraus wünschenswert?
- Oder macht eine (dezentrale,) **thematisch fokussierte Archivierung** (mglw. auch fortlaufendes Monitoring einschlägiger Channel) mehr Sinn?

- **Forschungsgetriebene Eigenentwicklung eines Crawlers** für Telegram
- Vorarbeiten zu einer **Vorverarbeitungspipeline** für weitere Analyseschritte wie Topic Modeling
- **Korpuslinguistische und netzwerkanalytische Verfahren** sowohl als eigene **Auswertungsebene** wie als Übergänge vom Distant zum Close Reading
- Identifikation **datenethischer und kuratorischer Fragen**, die über den konkreten Projektzusammenhang hinausgehen.



HERDER-INSTITUT
für historische Ostmitteleuropaforschung
INSTITUT DER LEIBNIZ-GEMEINSCHAFT

Mitglied der
Leibniz
Leibniz
Gemeinschaft

Gefördert von:



Die Beauftragte der Bundesregierung
für Kultur und Medien

HESSEN



Hessisches
Ministerium für
Wissenschaft
und Kunst