

Compiling Social Media Corpora in the face of ever-changing social media landscapes

Alexander König (CLARIN ERIC)

Egon W. Stemle (Eurac Research)



eurac
research

The problem

- Social-media data is - contrary to most other language data - very contemporary in its form
- There are a wide variety of social networks, not all of them widely known
- Each of them is constantly evolving to meet user and market needs (or because of the owner's inscrutable whims)
- Most corpora do not reflect this in their metadata

The problem

- Social-media data is - contrary to most other language data - very contemporary in its form
- **There are a wide variety of social networks, not all of them widely known**
- Each of them is constantly evolving to meet user and market needs (or because of the owner's inscrutable whims)
- Most corpora do not reflect this in their metadata

“Big in Japan”

- Hyves
 - Focused on the Netherlands (available in Dutch and English)
 - 2004 - 2013
 - At its peak 10.3 million accounts (NL population ~ 16 million)
- Orkut
 - Brazil
 - 2004 - 2014
- StudiVZ
 - Germany
 - 2005 - 2022
- VKontakte
 - Russia
 - 2006 -

“Remember those?”

- Myspace
 - Surprisingly still online
- Google+
 - 2011 - 2019
 - Following Google Wave and Orkut
 - Heavily pushed and then abandoned by Google
- Geocities
 - 1994 - 2009
 - Closure resulted in huge concerted archiving efforts

“In the margins”

- Diaspora
- Mastodon (before 2022)
- Tumblr

“The next big thing?”

- Bluesky
 - For a long time invite-only
- Threads
 - Started “late” in the EU
 - Heavily pushed via Instagram
- Mastodon
 - if you’re German
 - and/or are in academia

The problem

- Social-media data is - contrary to most other language data - very contemporary in its form
- There are a wide variety of social networks, not all of them widely known
- **Each of them is constantly evolving** to meet user and market needs (or because of the owner's inscrutable whims)
- Most corpora do not reflect this in their metadata

“My, how you’ve changed!”

Twitter evolution and devolution

- A tweet has a limit of 140 characters (including your username)
- A tweet has a limit of 140 characters (**ex**cluding your username)
- A tweet has a limit of 280 characters

- URLs count towards the character limit
- URLs do not count toward the character limit

- URLs in a tweet generate a preview
- URL previews obfuscate the domain of the URL

- All tweets are openly visible
- No, they are not
- Maybe they are? Sometimes

This is a chat

XChat View Server Settings Window Help

▼ freenode

Official Ubuntu Support Channel | IRC Guidelines: <http://ubottu.com/>

1 ops, 1871 total

#ubuntu

- * Now talking on #ubuntu
- * Topic for #ubuntu is: Official Ubuntu Support Channel | IRC Guidelines: <http://ubottu.com/y/gl> | This channel supports: Ubuntu and its official flavors, versions 12.04, 14.04, 15.10 and 16.04 | Unofficial derivatives: use your distro's support channel, not here. | IRC info: <http://ubottu.com/y/irc> | Pastes to <http://paste.ubuntu.com/> | Download: <http://ubottu.com/y/dl>
- * Topic for #ubuntu set by Pici!bnrubin@ubuntu/member/pdpc.profession al.pici at Thu Apr 21 21:54:49 2016
- ChanServ- [#ubuntu] Welcome to #ubuntu! Please read the channel topic. This channel is logged. Use of this channel implies acceptance of terms at <https://wiki.ubuntu.com/IRC/TermsOfService>
- * #ubuntu :<http://www.ubuntu.com>
- * **hatter** (~smuxi@c-2ec2339a-74736162.cust.telenor.se) has joined #ubuntu

tecmin

_m3
_moogs
_morphis
_ruben
S4MUR4I
_Trullo
`ryban
`z
Aaaaand
aavrug
Aayush
abhishek
AbigailBuccane
abort_
abra0

This is a chat

The screenshot displays a Discord chat window for a server named "Ballance". The channel is "# new", which has a description: "Server news; New speedrun videos, entertainment, custom maps etc.". The chat history shows several messages:

- A date separator for "2019年8月27日".
- A message from user "2jly" (2019/08/27) stating: "Super Monkey Ball will be on Steam soon" with a link to a Steam store page: https://store.steampowered.com/app/1061730/Super_Monkey_Ball_Banana_Blitz_HD/. This message has 4 reactions.
- A message from user "Sjeto" (2019/08/27) with a single smiley face reaction.
- A date separator for "2019年9月11日".
- A message from user "Cyberscore User" (昨天23:08) asking: "Is Banana Blitz HD coming out on PC?!" and adding: "Now I'm actually interested in that game."
- A date separator for "2019年9月12日".
- A message from user "2jly" (今天00:44) replying: "Yes, it is."

The interface includes a left sidebar with channel categories: ANNOUNCEMENT (welcome, speedrun, bulletin), TEXT CHANNEL (general, new, skill, dev, missions, mapping, suggestions, off-topic, application-evaluation, spam_zh), and BOT COMMANDS. The right sidebar shows a list of users, including ADMIN-3 (BallanceBug, Cyberscore User, Sjeto), DEV-1 (Diablo), and YOUTUBER/STREAMER-1 (NEVERBRICK-SUBSCR...). The bottom of the chat shows a message input field with a "new" tag and a "发消息" button.

This is a chat

The screenshot displays a Microsoft Teams chat window for the organization "Acme Inc.". The interface is divided into three main sections: a left-hand navigation pane, a central chat area, and a right-hand details pane.

Left-hand navigation pane: Shows the "Acme Inc." header with a user profile for Sharon Robinson. Below it, there are navigation options: "All unread", "Threads", "Mentions & reactions", and "Show more". A list of channels is visible, including "Marketing team" (with sub-channels # design-crit, # design-team, # Hubspot, and # social-media), "Announcements" (with # announcements, # media-and-pr, # released, and # team-sf), and "Pals" (with Arcadio Buendía, Leland Ygle... and Florence Garret). The "# social-media" channel is currently selected.

Central chat area: The channel name is "#social-media" with 21 members. The chat history includes:

- A message from Zoe Maxwell at 11:56 AM: "No! It was my pleasure! Great to see the enthusiasm out there."
- An announcement from Acme Team at 12:45 PM: "Event starting in 15 minutes: Team Status Meeting Today from 1:00 PM to 1:30 PM".
- A message from Harry Boone at 12:58 PM: "Quick note: today @Liza will join our team sync to provide updates on the launch. if you have questions, bring 'em. See you all later... er, in 2 minutes 😊".
- A message from Lee Hao at 12:58 PM: "Meeting notes from our sync with @Liza".

A "Meeting Notes" card is visible, indicating "1/9 Meeting Notes" and "Last edited just now". A Zenith Marketing logo is also present in the chat area.

Right-hand details pane: Titled "Details #social-media", it provides information about the channel:

- About:** Topic: "Track and coordinate social media"; Description: "Home of the social media team"; Created on: "October 18th, 2019".
- Members:** 21 members.
- Organizations:** 2 organizations (including Zenith Marketing).
- Pinned:** 1 pinned item.
- Shortcuts:** 1 shortcut.

Conclusion

- Social media environments have a large variance
- The environment is important to understand social interactions and the language production situation
- Knowing that a certain communication item is “a tweet” or “a chat message” is not enough detail

Suggestion

- Add a section to the corpus metadata/description detailing the specific environment
- Basically, instead of “this corpus contains Twitter data” -> “This corpus contains data from Twitter and when it was collected Twitter looked like this”
- Which info should be collected?
 - To be discussed in the community. The following slides contain some ideas...

Which info should be collected

- Follow model, social graph
- What is the “entry point” for a user
 - “for you” page?
 - Can you only see messages from people you selected yourself?
 - Is there some additional information in a side column? (related topics)
 - Are there global, local, personal trends that users might refer to?

Which info should be collected

- Mode of interaction
 - Retweets vs Quote Tweets
 - Subtweeting
 - Drukos vs. Drükos
 - Is threading possible? common? encouraged?
 - Is there moderation? Does it follow clear and documented standards?

Which info should be collected

- Type of items
 - Text, photo, video, audio
- Mode of text entry
 - Which device is commonly used for interaction?
 - Modern autocomplete or T9 style or none at all?
 - Are there emojis? emoticons? gifs? stickers?
- Limitations
 - Character limit
 - Language support (“full” unicode, RTL languages)
 - Usable with screen readers

Conclusion

- Social media corpora can cover a huge variety of services
- Especially with Twitter no longer being the easy to use go-to service, variety of social media corpora is expected to grow
- Environments can differ a lot from service to service
 - and even within the same service over time
- Services die and will be forgotten

- Therefore: detailing the service / environment should be an essential part of the documentation

Questions & Discussion

