Using Bluesky for Social Media Analysis

Erik Tjong Kim Sang

Netherlands eScience Center, Amsterdam, The Netherlands

June 6, 2025

TwiNL (Tjong Kim Sang and van den Bosch (2013)) is a Dutch archive of Twitter messages (tweets) widely used in research for social media analysis. However, Twitter, currently known as X, restricted access to their API in March 2023, which made collection of new tweets for the archive infeasible. In this paper, we examine the new microblogging platform Bluesky, set up by the founders of Twitter, as an alternative source for Dutch social media messages for research.

Keywords: corpus linguistics, social media analysis, Bluesky, Twitter

1 Introduction

TwiNL (Tjong Kim Sang and van den Bosch (2013)) is a Dutch archive of Twitter messages used for corpus linguistics (Schmitz et al. (2018)), opinion mining (Wang et al. (2020)), predicting election results (Sanders (2023)) and more. The archive contains messages from 2010 and later. However, in March 2023, Twitter, currently known as X, restricted access to their software API[1] which was used to collect tweets for the TwiNL archive, which meant that the archive could not be updated with new messages. In addition, tweet hydration, a process used by researchers to share tweets without sharing the text and metadata of the tweet[2], has become more difficult because of Twitter's policy changes. In this paper, we examine the Bluesky microblogging platform[3] as an alternative source of Dutch social media messages for research.

2 Bluesky

Bluesky was a project for Twitter's microblogging platform in 2019 with the goal of decentralizing the platform. The project became a company called Bluesky Social in 2021. Users were first accepted on an invitation basis in 2023 before the platform opened for all users in 2024. Bluesky reached the 30 million user mark in January

[1] An API (application programming interface) is an interface used by computer programs to communicate with each other in this case one computer requesting and retrieving data from another computer.

[2] The Twitter data access policy only allows researchers to share tweets ids. The tweet texts and meta data should then be accessed via the ids on twitter.com.

[3] Bluesky website: `https://bsky.app/`

Table 1: Ten most frequent words in Dutch Twitter (2011), Dutch newspapers and Bluesky (2025). Word frequencies on Twitter in 2011 were more similar to those in daily speech than those in written text. The Bluesky frequencies of 2025 are more similar to those of written newspaper text than those in daily speech, but also note the high rankings of the pronouns *ik* (*I*) and *je* (*you*). In 2014 Dutch Twitter switched to the written vocabulary distribution as well (Bouma (2015)).

	Twitter (2011)		Newspapers		Bluesky (2025)	
1.	ik	(I)	de	(the)	de	(the)
2.	je	(you)	van	(of)	het	(the)
3.	de	(the)	het	(the)	een	(a)
4.	een	(a)	een	(a)	en	(and)
5.	en	(and)	in	(in)	ik	(I)
6.	het	(the)	en	(and)	van	(of)
7.	niet	(not)	dat	(that)	dat	(that)
8.	rt	(RT)	te	(to)	is	(is)
9.	is	(is)	is	(is)	in	(in)
10.	dat	(that)	op	(on)	je	(you)

2025[4], while the former parent platform Twitter is estimated at 600 million users[5]. Posts on Bluesky have a maximum length of 300 characters which is 20 more than the current limit of Twitter.

Bluesky is starting to be recognized as an interesting source for research on social media analysis (Failla and Rossetti (2024)). One of the attractive sides of Bluesky for social media analysis is that it provides an open software API with good documentation[6] which makes it easy to collect Bluesky posts. The API allows searching for posts in specific language, which is useful for our purposes since we are only interested in posts written in Dutch.

For this study, we collected six months of Dutch Bluesky Social posts from 1 November 2024 to 31 May 2025. Through the API we retrieved all recent posts written in Dutch that contain one of the thirty-two most common Dutch words[7]. After removing duplicates, we had a collection of 4,207,642 Dutch Bluesky Social posts.

3 Analysis of Bluesky posts

This section presents a few examples of how Bluesky posts can be used for research[8]

3.1 Studying language usage

The TwiNL tweet corpus was originally composed for language use, in particular language usage on social media (Tjong Kim Sang (2011)), which is different than, for example, in newspapers, books and in daily speech. A Bluesky example of this can be found in Table 1. Table 2, repeats the gender study of Tjong Kim Sang (2011) with Bluesky posts. We also generated two biorhythm graphs shown in Tjong Kim Sang (2011) for the Bluesky data (Figure 1).

---

4 Number of Bluesky users: `https://bsky.app/profile/bsky.app/post/3lgu4lg6j2k2v`

5 Number of Twitter users: `https://x.com/elonmusk/status/1793779530282443086`

6 Bluesky API documentation can be found on `https://docs.bsky.app/`

7 Searching for any Dutch post is not possible on the platform; a query word is required.

8 The software used for the analyses presented in this section is available at `https://github.com/twinl/bsky`

Table 2: Comparison of most popular words in Dutch Bluesky posts of 50 users with female names compared with the t-test with posts from 50 users with male names. Left: The first group relatively used more positive self-reflecting words while the second group relatively used more political words. Right: The two groups are not linearly separable with a principal component analysis of their vocabulary: the best separating line assigns the correct category to 80 of the 100 users.

	users with female names	users with male names
1.	ik (I)	de (the)
2.	fijne (fine)	van (of)
3.	dag (day)	Wilders (politician)
4.	fijn (fine)	Trump (politician)
5.	mijn (my)	is (is)
6.	heerlijk (wonderful)	VVD (political party)
7.	lekker (nice)	PVV (political party)
8.	ook (also)	Nederland (Netherlands)
9.	en (and)	over (over)
10.	mooi (beautiful)	Rusland (Russia)

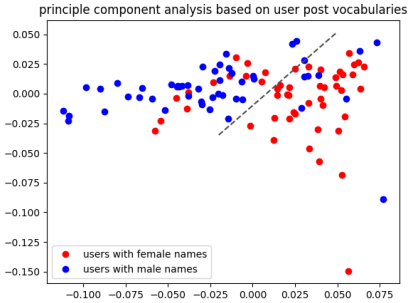
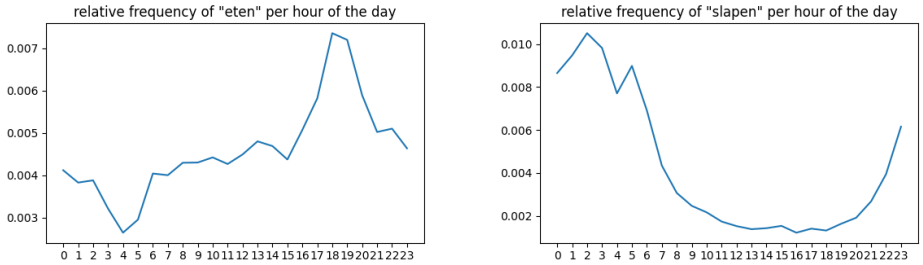


Figure 1: Biorhythm graphs generated from Bluesky posts. Left: relative frequency of "*eten*" (*to eat/food*), revealing the preferred Dutch dinner time at 18:00, but without signals for breakfast and lunch. Right: relative frequency of "*slapen*" (*to sleep*) which is higher between 0:00 AM and 8:00 AM.



3.2 Studying news events

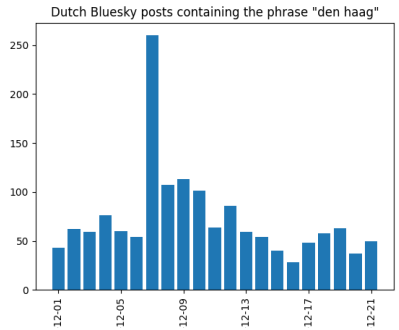
Twitter (currently X) has been widely used for studying responses to news events. The Twitter posts were excellent for this purpose, given the large volumes of posts and the wide range of interests of the users. Figure 2 shows the reaction of Bluesky to a Dutch news topic in the investigation period, the 7 December bomb attack on a bridal shop in The Hague. Table 3 shows the top 25 hashtags[9], and 25 popular political words in the six months of November 2024 - May 2025.

4 Concluding remarks

In this paper, we presented some examples demonstrating that, in the current challenging landscape of working with tweets, Bluesky posts can be used as a viable alternative for social media analysis, at least for Dutch. In the near future, we hope to collect more data from the platform and provide search access to them through a web platform.

[9] A hashtag is a topic-marking phrase preceded by a hash sign and added by users to their posts.

Figure 2: Summary of the Bluesky response to the 7 December 2024 bomb attack on a bridal shop in The Hague (Den Haag). There was a spike in the post counts for the city (left) and several of the posts contain terms related to the event (right): explosie (explosion), slachtoffer (victim) and Tarwekamp (the suburb of the attack).



References

Gosse Bouma. N-gram Frequencies for Dutch Twitter Data. *Computational Linguistics in the Netherlands Journal*, 5:25–36, 2015.

Andrea Failla and Giulio Rossetti. I'm in the Bluesky Tonight: Insights from a year worth of social data. *PLOS*, 2024. doi: https://doi.org/10.1371/journal.pone.0310330.

Eric Sanders. *Vox populi. On Forecasting Elections with Twitter*. MagellanBook, Oosterhout, 2023. doi: https://hdl.handle.net/2066/288385. PhD thesis Radboud University Nijmegen.

Tijn Schmitz, Robert Chamalaun, and Mirjam Ernestus. The Dutch verb-spelling paradox in social media: A corpus study. *Linguistics in the Netherlands*, 35(1):111–124, 2018.

Erik Tjong Kim Sang. Het gebruik van Twitter voor Taalkundig Onderzoek. *TABU: Bulletin voor Taalwetenschap*, 39(1/2):62–72, 2011. (in Dutch).

Erik Tjong Kim Sang and Antal van den Bosch. Dealing with Big Data: the Case of Twitter. *Computational Linguistics in the Netherlands Journal*, 3:121–134, 2013. ISSN: 2211-4009.

Shihan Wang, Marijn Schraagen, Erik Tjong Kim Sang, and Mehdi Dastani. *Dutch General Public Reaction on Governmental COVID-19 Measures and Announcements in Twitter Data*. arXiv.org, 2020. doi: https://doi.org/10.48550/arXiv.2006.07283.

Table 3: Frequent Dutch hashtags in three weeks of Bluesky social media posts (November 2024 - May 2025, left) and frequency of popular political terms selected by the authors of this paper (right). An asterisk (*) behind a word indicates that matches inside longer words have also been counted (like *oorlogen*, *wereldoorlog*, *naoorlogse*, ...). These display interests in news, sports, politics, festivities and hobbies. However, a study of less popular topic like *metoo* on Bluesky would require including also posts of other languages, in particular English.

	freq	hashtag	English	freq	term	English
1	50272	#nieuws	news	83975	Trump*	politician
2	41327	#vacature	vacancy	36183	Wilders*	politician
3	19273	#rtlnieuws	broadcaster	34810	PVV*	political party
4	15729	#algemeendagblad	newspaper	34168	oorlog*	war
5	7760	#rotterdam	city	26906	extreem*	extremism
6	5774	#nos	broadcaster	25991	Israël*	Israel
7	5671	#trump	politician	22014	Oekraïne*	Ukraine
8	5205	#natuur	nature	19638	Gaza*	Gaza
9	4931	#efteling	amusement park	18700	klimaat*	climate
10	4714	#photography	photography	14801	democratie*	democracy
11	4627	#sportnieuws	sports news	13501	genocide*	genocide
12	4590	#update	update	11771	vrijheid*	freedom
13	4522	#randstad	conurbination	11485	rechten*	rights
14	4350	#tda25	?	10034	protest*	protest
15	4116	#amsterdam	city	8517	woke*	woke
16	3883	#groningen	city	8375	fascisme*	fascism
17	3858	#nu.nl	broadcaster	7218	paus*	pope
18	3837	#almelo	city	7169	milieu*	environment
19	3208	#pvv	political party	5817	racisme*	racism
20	2970	#utrecht	city	2766	discriminatie*	discrimination
21	2931	#ajax	soccer club	845	slavernij*	slavery
22	2871	#fotografie	photography	812	uitbuiting*	exploitation
23	2636	#olympia	?	201	Nethanyahu*	politician
24	2421	#bbb	political party	145	metoo	metoo
25	2376	#wilders	politician	88	BLM	black lives matter