

Policy Issues vs. Documentation: Using BERTopic to Gain Insight in the Political Communication in Instagram Stories and Posts during the 2021 German Federal Election Campaign

Michael Achmann, Christian Wolff

Media Informatics Group, University of Regensburg, D-93040, Regensburg, Germany

Abstract

We give first insights in the political communication of the 2021 Federal election campaign in Germany. We focused on political messages found in ephemeral stories (n=2208) and permanent posts (n=718) shared in the last fortnight of the campaign. Topic modeling with BERTopic did not yield topics as finely grained as the categories of prior content analyses, yet two main themes emerged: The majority of posts deal with policy issues, while the majority of stories does not deal with policy issues. We found a large body of stories to be documentation of the rallies and campaign trail.

Keywords

topic modeling, social media analysis, instagram stories, visual social media, political communication

1. Introduction


Instagram, initially an image-sharing platform, is now one of the most popular social networks [1]. It has added features such as the algorithmic timeline, stories, and reels, all of which focus on visual media [2]. As such, it has become an important network to be used in election campaigns. The political communication of these campaigns, has been analyzed in several studies in the past [3]. The story feature, however, has attracted little attention from scholars even though the ephemeral character of stories stands out in a world of technology where "forgetting has become the exception, and remembering the default" [4]. In order to gain an initial understanding of an election campaign on Instagram, we analyzed the differences between stories and posts, proposing to focus on text-integrated images and classify the content using topic modeling. By comparing these two forms, we want to shed light on how political communication is evolving in response to changes in technology and social media usage and see our work as part of a larger body of research that seeks to understand the ways in which social media is changing political communication. Thus we try to answer the following questions:


DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

✉ michael.achmann@informatik.uni-regensburg.de (M. Achmann); christian.wolff@informatik.uni-regensburg.de (C. Wolff)

🌐 <https://go.ur.de/michael-achmann/> (M. Achmann); <https://go.ur.de/christian-wolff/> (C. Wolff)

🆔 0000-0002-4754-7842 (M. Achmann); 0000-0001-7278-8595 (C. Wolff)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 DHNB Publications, DHNB2023 Conference Proceedings, <https://journals.uio.no/dhnbpub/issue/view/875>

1. What are the main political message types on the Instagram accounts of front-runners and political parties in the final two weeks of the 2021 election campaign?
2. How do these messages differ between ephemeral stories and permanent posts?
3. How well can we answer these questions computationally through the lens of BERTopic?

1.1. Political Communication on Instagram

Despite the young age of Instagram, the political communication on the platform has already been studied in numerous papers, with a focus on different political actors and different nations: Bast reviewed 37 studies, systematizing them according to the methodological approaches, theories, and sampled data. She found studies to address three key areas: "Who uses Instagram, how do they use it, and with what effect?" [3]. The majority of studies employed a quantitative approach, with quantitative content analysis being the most prevalent methodology utilized. The variety of approaches and theories lead to a multitude of study designs: Some studies compare communication strategies on different platforms (e.g. [5]) or between party accounts and politician accounts (e.g. [6]). Others, such as Lalancette and Raynauld's highly-cited analysis of Justin Trudeau's Instagram use, which utilizes the theoretical framework of celebrity politics, concentrate on a single individual [7].

The overarching result of these studies shows political figures are taking advantage of Instagram to present a positive and encouraging image, rather than delving into policy issues, meeting with constituents, or organizing voting efforts. Most posts feature pictures of the political actor themselves or, in the case of political party accounts, images of their leading candidate. A comparison to a similar literature analysis focusing on Twitter usage during political campaigns reveals similar usage styles: both platforms are rarely used by politicians to interact with voters, though there is considerable variance between individual players.

Since the literature review's publication several new studies about political communication using Instagram have been published, for instance a first longitudinal study using *CrowdTangle*¹ to retrospectively collect posts to shed light on changes in Instagram use of European political parties over time [8]. Further, studies regarding the visual communication of European right-wing populist politicians [9], differences in user engagement with political parties between Instagram, Facebook and Twitter in Canada [10], Instagram use of Spain's major political parties [11], a cross-country study of politicians' self-depiction [12] using computer vision, once more Justin Trudeau's use of Instagram [13] and finally the use of Instagram stories by Trump and Biden in the 2020 presidential election [14] have recently been published.

Bast concludes her review study arguing in favor of more systematic comparisons with larger and more heterogeneous samples, as well as more longitudinal studies that go beyond single election campaigns. She suggests that the lack of precise and coherent definitions of concepts and content analyzed be remedied by transferring established analytical concepts in order to build solid evidence. In addition, she argues that the relatively new Instagram functions *Instagram Video* and *Instagram Stories* present a valuable opportunity for further research, an argument backed by others [6, 14]. Stories are distinct from posts, which are the original content

¹<https://www.crowdtangle.com/>

on Instagram. Posts can include one or more images and / or videos that are permanently shared on a user's profile, often accompanied by captions (text content). In contrast, stories are a relatively new feature (see below) that are ephemeral and solely composed of an image or video. Unlike posts, they disappear after a certain period of time and do not remain on a user's profile.

1.2. Computational Analysis of Social Media Content

We see potential to increase the comparability of social media analyses through the use of computational methods to create reproducible and valid analyses. In addition, computational approaches enable us to handle a growing amount of user generated content [15, ch. 1], namely visual content in the context of Instagram. We propose to focus on text-integrated images and captions in order to apply computational text analysis methods, which are well established [16], and may serve as a bridge towards the computational analysis of visual media, which is yet a challenge [17, 18]. Overall, we want to explore the potential of computational approaches to discover and analyze visual social media content, with present work focusing on topic modeling as one possible candidate in the development of a workflow for computational visual content analysis.

1.3. Topic Modeling & Instagram

While Instagram is primarily focused on visual media, text has already been used to explore themes of posts: Rodina and Dligach analyzed the themes and topics of posts by Ramzan Kadyrov, dictatorial head of the autonomous Chechen Republic. Using the Latent Dirichlet allocation (LDA) model they found two dominant themes across 6854 analyzed posts and 24 topics: A personal and a political theme, which over time started blending [19]. In health domains several social media studies relied on topic modeling: Murashka et al. tried to identify objectification elements from image captions and comments of popular #fitspiration accounts. Kim et al. [21] identified how people managed their daily lives in the face of the pandemic's fear and discomfort by applying topic modeling on captions and image descriptions. Similarly Muralidhara and Paul [22] looked into Instagram posts with health-related hashtags and identified 47 health-related topics in their corpus. They trained their model on hashtags and caption words to automatically generate image tags. In a journalistic context Al-Rawi et al. [23] employed topic modeling in a mixed-methods approach to explore the most liked news topics across several news accounts.

1.4. Ephemeral Content & Stories

While a research gap in political communication exists, ephemeral Instagram stories have been investigated in other disciplines. Stories are a special type of post as they expire after 24 hours and became the platform's main growth engine [2]. After expiry, they are archived for the authoring user but not for other users. They consist of videos or images, or collages of media and so-called stickers, platform specific affordances [24] to tag other users; hashtags; locations; or allow for interaction through e.g. questions or quizzes. Since Snapchat invented the ephemeral feature, it is worth to look at Rettberg's study of Snapchat content. She suggests that the app changed online communication and its affordances enable the discovery of more

conversational methods of communicating and telling stories, thus "Snapchat is a conversation, not an archive". Through qualitative content analysis, observation and in-depth interviews, Amancio found four narrative elements used by Snapchat and Instagram storytellers to tell their stories and construct a narrative: actions (demonstrating emotions, eating, interacting), happenings (updates), characters (people, self-portraits and animals) and setting (environment), making use of images, texts, videos, emoji, doodles, instant information and filters [26]. Bainotti et al. investigated 292 Instagram Stories by private users using an ethnographic coding approach. They claim to have identified specific grammars by matching the content and context-of-use, the two main ones are: "a grammar for documentation and a grammar for interaction" [27]. Other areas of interest for stories were ephemeral journalism [28] and Female Athletes' self-presentation [29]. Closer to political communication, a study of the candidates for the 2016 U.S. presidential primaries identified ten frames used on Snapchat [30]. Finally, Towner and Muñoz [14] published a first analysis of political communication in Instagram Stories, studying the stories published by the two U.S. presidential candidates in the 2020 campaign. They collected a sample of 304 images one week before and after the election campaign. From a marketing perspective, they saw several flaws, like missed opportunities of sharing user-generated content and inconsistently following communication norms for Instagram Stories. Further, campaign events and rallies were the most popular type of messages.

2. Methods

In order to uncover the main political messages of posts and stories in the 2021 election campaign, we used word frequency measures, word clouds and topic modeling, an unsupervised machine learning technique. Since Instagram primarily consists of visual media, we applied optical character recognition (OCR) to translate text-integrated images into machine readable text.

2.1. Data Collection

We collected a sample of 2208 stories and 718 posts shared by politicians and parties within the last fortnight of the 2021 federal election campaign. Stories were collected daily at 0:00 (CET) using Selenium, a Python package to simulate a human user browsing the stories.² Posts were collected retrospectively through *CrowdTangle* and *Instaloader*. Germany's multiparty system has witnessed a growing trend of fragmentation in recent years. As a result, we conducted a comprehensive data collection, focusing on posts and stories shared by the eight political parties participating in the election, which currently hold seats in state or federal legislatures and possess verified Instagram accounts (refer to Table 1). Additionally, we ensured the inclusion of at least one front-runner from each party (refer to Table 2).

2.2. Preprocessing

There are three different sources for text which we we have used in our analysis: 1) post captions, which is computer readable text added by users to posts, 2) text-integrated posts, and 3) text-

²Data for Sep 14 is incomplete due to technical problems. For present proof of concept work the incompleteness of the sample has been ignored.

Table 1

Selected parties and their Instagram handles at the time of data collection.

Party (Abb.) @handle	Party (Name) Translation
AfD @afd_bund	Alternative für Deutschland Right-wing Populist Party (Alternative for Germany)
CDU @cdu	Christlich Demokratische Union Deutschlands Centre-right, Christian Democrats (Christian Democratic Union of Germany)
CSU @christlichsozialeunion	Christlich Soziale Union in Bayern Bavarian Centre-right (Christian Social Union)
Die Grünen @die_gruenen	Bündnis90 /Die Grünen Green, Environmental Politics (Alliance 90/The Greens)
Die Linke @dielinke	Die Linke Democratic Socialists, Left-wing (The Left)
FDP @fdp	Freie Demokratische Partei Classical Liberals, Pro-business Free Democrats (Free Democratic Party)
FW @fw_bayern	Freie Wähler Centrist, Citizens' Groups (Free Voters)
SPD @spdde	Sozialdemokratische Partei Deutschland Social Democrats, Centre-left (Social Democratic Party of Germany)

integrated stories. The text of the latter two is embedded in either images or videos. While captions and embedded text for posts are available through *CrowdTangle*, the embedded text in stories is not. Thus we used EasyOCR for Optical Character Recognition – for consistency on both, stories and posts – to extract the embedded text from text-integrated images. A majority of stories (n=1246) turned out to be videos. As the Instagram app just allows to add one combination of text and stickers which is displayed across all frame of videos, we extracted the first video frame using OpenCV.³

2.3. OCR & Relevance Classification

Since EasyOCR turned out to be overambitious recognizing the embedded text, e.g. transcribing shop signs from the image's backdrop, we trained a CNN⁴ to classify relevant and irrelevant text-snippets (see figure 1). A human annotator corrected the OCR results and annotated the relevance of each text snippet for 50% of all captured stories. Sticker content has been labeled as irrelevant since their content is available in the metadata. Through the annotation process more than half of the OCR annotations were deemed irrelevant (4794 out of 9850). Our model reached an f1-score of .94 which we deemed sufficient (see table 3). At the same time, only

³This approach disregards embedded text in videos itself, like subtitle. We see future work taking every frame into account, controlling for repeated text across frames.

⁴A Convolutional Neural Network, a type of neural network used in machine learning to classify images. After some experiments we archived the best results using only the cropped images of text snippets as input data.

Table 2

Selected politicians' accounts and their positions and party affiliation at the time of data collection. (The party GRÜNE is referenced as B90DieGruenen later on.)

Name	Party	Position	@handle
Alice Weidel	AfD	Front-Runner	@alice.weidel
Jörg Meuthen	AfD	Head of Party	@joerg.meuthen
Armin Laschet	CDU	Chancellor Candidate	@armin_laschet
Markus Söder	CSU	Head of Party	@markus.soeder
Annalena Baerbock	GRÜNE	Chancellor Candidate	@abaerbock
Robert Habeck	GRÜNE	Front-Runner	@robert.habeck
Ates Gürpınar	Die Linke	Deputy Head of Party	@atesgurpinar
Susanne Henning-Wellsow	Die Linke	Head of Party	@susanne_hennig_wellsow
Christian Lindner	FDP	Front-Runner	@christianlindner
Nicola Beer	FDP	Deputy Head of Party	@nicola_beer
Engin Eroglu	FW	Deputy Head of Party	@engin_eroglu
Gregor Voht	FW	Deputy Head of Party	@grey_gor
Olaf Scholz	SPD	Chancellor Candidate	@olafscholz
Saskia Esken	SPD	Head of Party	@saskiaesken



Figure 1: Application of the trained model for relevance classification. The rescaled image of a story on the left shows yellow bounding boxes for relevant and red ones for irrelevant text snippets. The right-hand image shows a step of the processing pipeline: Classification takes place for each extracted snippet. The top line, for example, shows the correct classification of a location sticker as irrelevant since we are able to extract the sticker's content from the metadata.

minor human adjustments were required for the OCR of relevant text-snippets: Using R's `adist` (approximate string distances), we calculated the generalized Levenshtein-Distance between OCR and human-corrected text: there was a mean distance of only .51 characters. Overall 104 transcriptions (across 57 pictures) were added entirely from scratch throughout the annotation process.

Table 3

Classification report of the classification model.

	precision	recall	f1-score	support
Irrelevant	0.97	0.93	0.95	5537
Relevant	0.91	0.96	0.94	4349
accuracy			0.94	9886
macro avg	0.94	0.94	0.94	9886
weighthed avg	0.94	0.94	0.94	9886

All text has been preprocessed using the `text-clean` python package to tackle encoding discrepancies and to remove emojis. For the word clouds, German characters have been converted to ASCII characters, punctuation has been removed and finally stop words were deleted using NLTK’s German stopword list.

2.4. Topic Modeling

In order to gain a computational insight of our corpora, we employ topic modeling. It is a set of algorithms that assist in identifying recurring themes in a corpus of documents [31]. Probabilistic models like the Latent Dirichlet Allocation (LDA) assume that each document exhibits multiple topics in different proportions [32]. A topic is a set of terms which frequently co-occur across the documents. These topics and corresponding terms aid in uncovering themes across the set of documents, in the case of Instagram posts and stories we employ topic modeling to help uncovering the content and identify policy issues of the 2021 election’s Instagram campaign. The result of our topic model(s) serves as the basis for the message type analysis.

Considering recent developments in language models, we sought out approaches based on modern language models. We identified *BERTopic* [33] as suitable software, since the author reports BERTopic to perform well on a corpus of tweets. Egger and Yu compared several topic modeling approaches and confirm the performance on tweets. While we are dealing with Instagram content, tweets are by definition social media content and contain short texts, thus they are comparable to our corpora.

Once the relevance model was applied to the two OCR corpora of text-integrated posts and stories, we created an overarching corpus consisting of the captions and relevant image-text. Multiple text lines of image text were concatenated to one document, thus one story image corresponds with one document in the corpus and one post corresponds to at least two documents, one for the caption, and one or more for the image text (with one post corresponding to one or more images). We trained the following three models:

Overall Model The overall model is our naive starting point. It was trained using all documents from the corpora together, thus it includes post and story OCR as well as post captions. We used the `fasttext` [35] German word vector model `cc.de.300.bin` and `paraphrase-multilingual-MiniLM-L12-v2` sentence transformer. As per BERTopic documentation, we removed stop words using `scikit-learn`’s `CountVectorizer`

method. NLTK's German stop word list was applied after the word embeddings have been created and documents were clustered.

Post Model Once we spotted first weaknesses in the overall model (see below), we decided to train models per corpus. First tests using either OCR or caption documents yielded a low amount of topics. Consulting the BERTopic documentation, we decided to split captions by sentences to generate a larger set of documents. As hashtags and mentions started dominating the topics, we added an additional preprocessing step, removing any hashtags and mentions from the caption corpus.

Story Model The story model was trained using the same parameters as the overall model. It was only trained using the story OCR documents.

2.5. Political Messages

In a second step we use the results of the topic modeling as a basis for message type classification. Due to the topic modeling results we suggest to categorize the messages into two main categories along the existence or absence of policy issues. Towner and Muñoz introduce ten main types with higher granularity: campaign events, thank you, voter endorsement, policy issues, hybrid, character, attack, behind the scenes, mobilization and other [14]. They based their categories on Liebhart and Bernhardt's study of an Austrian presidential campaign [36], which also served as a foundation for Haßler et al.'s [6] image types. The later developed nine image types: policy, campaign events, call for action, negative campaigning, media work, campaign material, supporters, everyday political work, private background story. While our results of the unsupervised topic modeling approach did not yield topics as finely detailed as to sharply distinguish between the messages types from the literature, we are optimistic that the two types will be sufficient to elucidate the differences between stories and posts.

Policy Issues We considered policy issues to be actual political content regarding a variety of domains. A post or story was considered to regard policy issues if any issue was present at all. Fringe cases (e.g. short sentences by B90DieGruenen using their neologism "Klimaregierung" (climate government), calling for more climate action) were considered as policy-bearing while slogans (e.g. "Entlasten statt Belasten", relieve rather than burden, by CSU) by itself was not considered to refer to policy issues, since the first example explicitly takes up a political issue, while the slogan in the second example merely refers to abstract relief without any specific indication of what type of relief is meant.

No Policy Issues This category is the opposite of the policy issue category and merely describes the absence of political topics in the textual content. Examples include text-integrated posts bearing "Damit es möglich wird: Am 26.9. DIE LINKE wählen" (to make it possible: vote for DIE LINKE on 26.9.) or "Danke für eure Unterstützung" (thank you for your support) posts. It includes the types campaign events, thank you, voter endorsement, character, attack, behind the scenes, mobilization and other by Towner and Muñoz.

3. Results

Throughout the period of investigation, a total of 2208 stories were collected, most of them (n=1246) videos. In the same period 713 posts were published by the parties and politicians. While B90DieGruenen created the most stories (n=578), the CDU published the most posts (n=144). The most stories were created on Sept. 24 (n=311), and the most posts on Sept. 23 and 24 (n=75 on both days), with an average of 157.71 stories and 50.93 posts published per day. The application of the preprocessing steps described above resulted in three text corpora derived from the Instagram content: 1. captions, 2. post OCR, and 3. story OCR.

Captions are genuine digital text added by the user when creating a permanent image or video post. 706 posts were accompanied by captions (99%). Unsurprisingly, captions were the longest texts we observed with an average of 75.80 (median=60.00) words per post. Each caption incorporated an average of 2.72 hashtags (median=2.00). Among the ten most frequent words across all posts were terms like “deutschland” (Germany), “heute” (today), “wählen” (to elect), “stimmen” (to vote), “bundestagswahl” (federal election). Before filtering the hashtags “btw21”, “csu” and “cdu” ranked among the most top words, hinting at a consequent use of these hashtags by the CSU and CDU parties in combination with them being the most active posters.

Post OCR consists of the OCR results classified as relevant. Each post may contain several images or videos⁵. We collected a total of 1299 image files belonging to 713 posts. easyocr identified text in 1092 images, out of which 650 images contained text classified as relevant.⁶ Overall, the 650 images belong to 532 different posts, thus 74.61 % of posts contain text-integrated images. On average, text-integrated post images contain 15.87 words (median=12.00). The most frequent words look similar to the captions with “deutschland” (Germany), “wählen” (to elect), “stimmen” (to vote) taking the top places.

Story OCR comprises the relevant text snippets found in stories, with each story containing one image that may include multiple text snippets. Further, stories offer the possibility to add so-called stickers, like hashtags, mentions or locations. Our relevance classification model was trained to reject text from the stickers since we can retrieve sticker information from the metadata. Out of a total of 2208 stories 1939 (87.82%) contain relevant text-snippets. On average each story contains 15.66 (median=12.00) words. The most frequent term is “heute” (today), followed by “uhr” (o’clock) and “danke” (thank you). While “deutschland” (Germany) and “wählen” (to elect) also rank among the top ten words for stories, their relative frequency is lower in comparison to captions and post OCR.

Overall, we see text to be part of the majority of shared content. Almost every post contained a caption and the majority of post images and stories were text-integrated. Top word frequencies show first similarities and differences across the corpora. In order to gain a better insight of the

⁵We used the cover image of videos for our analysis.

⁶Since we trained our model using annotated story images we qualitatively explored the rejected text snippets and are confident not to have missed important content. In addition, we took a deeper look at images without any text (after relevance classification) and did not find a single text-integrated image rejected erroneously.

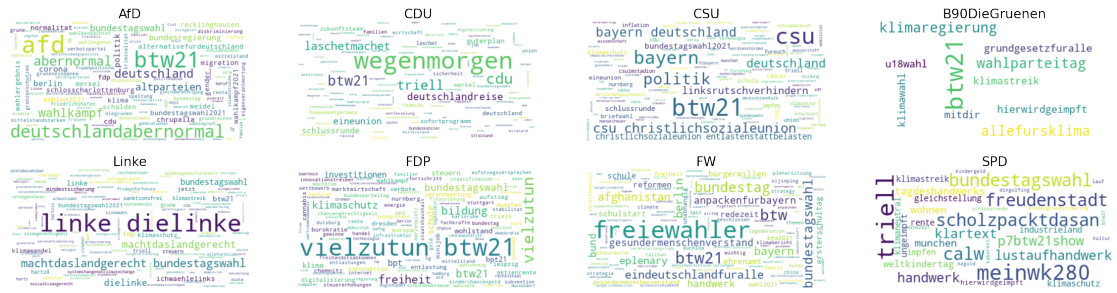


Figure 2: Word cloud of hashtags used in captions grouped by parties.

content shared by politicians and parties we first take a look at the word clouds and thereafter present an overview of topics as generated by BERTopic.

3.1. Looking at the Word Clouds

One platform-specific affordance [24] of Instagram is the use of hashtags. While hashtags may be used in both, stories and posts, our data shows that hashtags are rarely used in stories. However, since they have been used in the past [37] for automatic image annotation, we will first take a look at caption hashtags through word clouds and then proceed to our three corpora. The word clouds have been generated using the wordCloud package. We grouped the text documents by party, this split allows, on the one hand, to see differences between the parties, and, on the other hand, to control the dominance of parties with a higher posting frequency.

Hashtags Most parties referenced policy issues in their hashtags (e.g., “#klimaschutz” (climate protection), “#digitalisierung” (digitization), “#impfen” (vaccination). FDP used the most varied hashtags, followed by SPD and B90DieGruenen. Across all parties the #btw21 was the most used hashtag. Non-policy related hashtags, including party names and slogans (e.g., “#wegenmorgen” (abouttomorrow), “#vielzutun” (lots todo), “#deutschlandabernormal” (germanybutnormal)), were dominant for all parties. The CSU appears to be the only party to have established a negative-campaigning hashtag “#linksrutschverhindern” (see figure 2).

Captions Similar to the hashtags, caption word clouds rarely offer insight into policy issues important for the campaign. Mentions of election, Germany, party names and front-runners were frequent. CSU was the only party with more occurrences of “bayern” (Bavaria). B90DieGruenen most consistently referenced their posts with “klimaschutz” (climate protection). Other policy issues mentioned were “hartz IV” (social welfare benefit program, Linke), “familie” (family, CSU) and “schulen” (schools, FDP).

Post OCR The text-integrated posts focus on party names and the election, with the AfD associated with “nein” (no) possibly indicating negative campaigning. FDP refers to “bildung” (education), Linke to “pflege” (nursing), “rente” (pensions) and “klimaschutz” (climate protection). CSU’s slogan is “entlasten statt belasten” (better relieve than burden) focusing on Bavaria alongside FW.

Story OCR The term “heute” (today) dominates stories across parties, as do front-runner names. The AfD used a question sticker, resulting in that question dominating their stories. The FDP held a political convention which was referenced in many stories.

The macroscopic look at the wordclouds allowed us to gain a first insight into the Instagram election campaign. Overall, we see some policy issues emerging. Mostly, however, they are obfuscated by the parties mentioning their names, referencing their candidates and using slogans or tailored hashtags again and again, thereby dominating the word frequencies which are the basis of word clouds. Thus, in order to computationally gain a better insight of the campaign we used BERTopic for topic modeling.

3.2. Looking through the lens of BERTopic

Once the first model has been trained, several policy-issues appeared among the topic representations. Over the course of several iterations we refined the model(s) and decided to split posts and stories in order to gain better results. Once we trained the overall model we inspected the topics through a dendrogram and an inter-topic distance map and decided to reduce the initial topic count of 62 to 25. Topic -1 refers to outliers, all items sorted into this topic were disregarded. For comparability we kept the topic count constant across the three models.

First, we offer a qualitative look into the results for each topic model. In a second step during qualitative inspection we assigned a new variable to each topic, in order to distinguish topics dealing with policy issues from topics with other, non-policy related issues. While a majority of topics clearly leaned towards the absence or presence of policy issues, some could not be subsumed in either of the classes. These were classified as mixed, in order to prevent misinterpretations. Since each post consists of one caption and at least one image, we considered a post to contain policy issues if at least one image or the caption contained policy issues. Thereafter we are going to use this variable for a quantitative interpretation of our corpora.

Overall Model The overall model uncovered several policy issues, like climate change, renewable energies, education & digitization, labour and social issues, and the economy. On a closer look, however, several topics showed inconsistencies between the different corpora. One topic, for example, consists of text-integrated posts with short texts about policy issues, as well as a majority of policy-focused captions. The stories in this topic, however, are mostly documentation. Further, several policy issues were mixed together in plenty of topics, thus a differentiation of policy issues by the overall topic model may not be very accurate. All in all on visual inspection of the post images and stories we saw first patterns emerge, namely the difference between policy-issues focused content and documentary content. Further, the topic representations appeared in several cases to only cluster either posts (OCR / caption) or stories into meaningful categories, we call these *mixed* topics. Hence we decided to train separate models to gain better insight, hoping to improve topic validity.

Stories Model The story model uncovered several clusters of stories documenting the election campaign, namely some "thank you!" and "hello" topics as well as a city names / geographical locations topic and a "selfie & beer garden" topic. One topic identified mostly

Table 4

Overview of the 25 topics in the posts model, the amounts of post captions per topic and the message type variable assigned for each topic.

Topic	Post Captions		Title	Message Type
	Party	Person		
0	47	28	Climate, CO2 and more	Pol. Issues
1	32	10	Focus on Germany, mostly Slogans	Mixed
2	2	2	Announcement and Angela Merkel	No Pol. Issues
3	6	2	Election-Centred	No Pol. Issues
4	1	2	Need to Fight	Mixed
5	33	5	CDU / CSU Slogans, Mixed with Policy Issues	Mixed
6	13	1	Election Day	No Pol. Issues
7	0	0	NA	No Pol. Issues
8	5	2	Thanks, Thanks, Thanks	No Pol. Issues
9	18	5	CSU Slogan, Financial Relief	Pol. Issues
10	0	0	NA	No Pol. Issues
11	13	12	Democracy, Middle Class, and more	Pol. Issues
12	0	3	Announcements	No Pol. Issues
13	11	4	Focus on the Union (CDU+CSU Party)	No Pol. Issues
14	4	0	Crisis and Scandals, mixed with negative-campaigning	Mixed
15	2	0	FDP-Slogan	Mixed
16	34	16	Digitization and Education	Pol. Issues
17	90	60	Financial Relief, Taxes, Debt: Share Pics with short Policy Snippets, mixed with slogans	Mixed
18	5	1	Specific Amount of Money	Pol. Issues
19	3	1	Children and Families	Pol. Issues
20	19	11	Schools and Education	Mixed
21	4	11	COVID and Reliability	Pol. Issues
22	96	72	A variety of political issues, centered around the change for the future	Pol. Issues
23	6	7	Change and Future Direction	Pol. Issues
24	5	9	The gap between Cities and Countryside, Germany-Centred	Pol. Issues
Total	449	264	713	

negative campaigning and another one stories focused on the chancellor candidates. While a majority of stories appeared in these documentation topics, we also discovered several hybrid topics, which consist of stories documenting campaign rallies or similar events supplemented by quotes or short snippets referring to policy issues. Namely taxes, economy, progress, social, education and children appeared as policy issues among these topics. They are different to the *mixed* topics, as they are not a mix of stories with and without policy issue, rather they combine the typical elements of documentation style images with policy issue through e.g. through short quotes, thus we assessed these stories as policy issue message types. Finally, a few policy-focused topics emerged: We observed climate change, democracy, young people, family and children as policy issues through the lens of this model.

Posts Model The initial, unreduced, posts-model based on caption-sentences offers the most detailed look into policy issues: climate change, digitization, education and family emerge, moreover we see issues like employment, affordable housing, COVID-19, police and safety, debt and economical directions, democracy, creative and cultural industry, deregulation and the reduction of bureaucracy, and agriculture and the countryside. Once the model was reduced to the target of 25 topics, some themes such as Climate change, democracy, education, schools and digitization, and children and families remained clearly visible .In

Table 5

Overview of the 25 topics in the story model, the amounts of stores per topic and the message type variable assigned for each topic.

Topic	Stories	Title	Message Type
0	107	Thanks, Thanks, Thanks	No Pol. Issues
1	135	Climate Change	Pol. Issues
2	23	Documentation	No Pol. Issues
3	72	Election & Candidates	No Pol. Issues
4	73	Documentation	No Pol. Issues
5	202	Place Name & Documentation	No Pol. Issues
6	32	Democracy, Young People, Family, Hatred	Pol. Issues
7	108	Mostly Documentation	Mixed
8	222	Short Policy Issues (Taxes, Economy) and Documentation	Pol. Issues
9	37	Pol. Issues of AfD and Die Linke; others: Documentation	Mixed
10	27	Interviews	No Pol. Issues
11	9	Party Names, Negative Campaigning	No Pol. Issues
12	38	Selfies & Beer Garden: Documentation	No Pol. Issues
13	16	Children & Family	Pol. Issues
14	47	Focus on Chancellor Candidate	No Pol. Issues
15	12	Documentation	No Pol. Issues
16	1	NA	No Pol. Issues
17	98	Documentation & Short Policy Snippets	Mixed
18	21	Progress & Persuasion, Documentary Hybrid	Pol. Issues
19	18	Interviews	No Pol. Issues
20	17	Hello, Hello, Hello!	No Pol. Issues
21	115	Press Conferences & Announcements, paired with Content	Mixed
22	42	Thanks, Thanks, Thanks	No Pol. Issues
23	192	Documentation	No Pol. Issues
24	216	Documentation, Minority Issues: Social, Education, Children	No Pol. Issues
Total	1880		

the end, the majority of posts are dealing with policy issues. However, seven topics had to be classified as **mixed** since they neither showed a clear majority of policy issues nor documentation-style images and captions.

Through the qualitative inspection of the three models and their topics we were able to gain an insight into several policy issues occurring through the posts and stories. The topics, however, did not always differentiate the posts / stories precisely into a coherent theme. Captions classified by the posts model, for example, are tough to group into one topic as they often contain one or more policy issues. At the same time the stories may be grouped rather well into a coherent topic due to the rather short text-length and focus on one issues, yet the majority of stories turned out not to deal with any policy issue at all.

All in all, we found a majority of posts to deal with policy issues (see figure 3, 64.24%). Almost

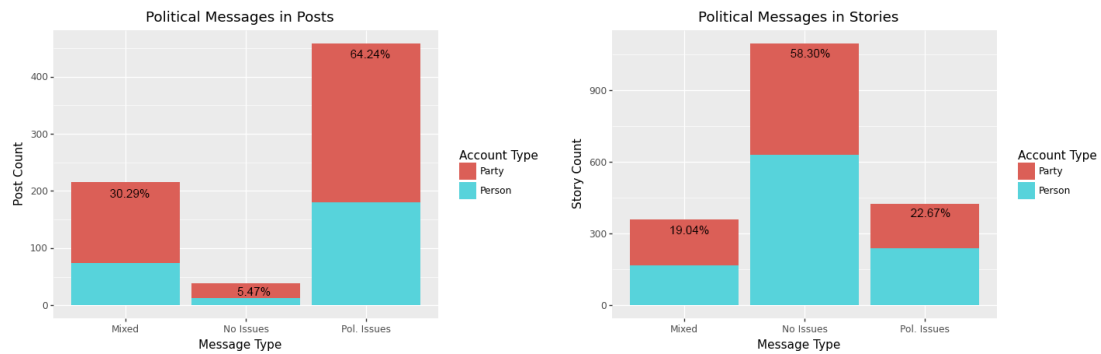


Figure 3: Policy issues in posts (left, n=713) against policy issues in stories (right, n=1880) by account type.

a third of posts (30.29%), however, fell into mixed topics, which could not be clearly classified as content bearing or not. Nevertheless, a clear minority of posts was identified to not deal with policy issues at all (5.47%). The stories, on the other hand, show a majority of items to not deal with (58.30%) policy issues. A smaller share clearly were regarded as policy-bearing content (22.67%) and a minority as mixed (19.04%). We have not observed any significant differences between user types (party accounts vs. political leaders), nor between different parties.

This story model unearthed what could be described as a subtype of both main message types: a documentation class. Exploring the stories of these documentary topics we find parallels to the findings by Bainotti et al. [27] about personal Instagram stories. They discovered a “grammar for documentation” that is used to portray both exceptional event photographs and regular, everyday situations. In the documentation topics, stories about political rallies showed campaign events, different stops along the campaign trail, front-runners on stage, and the crowd or both without any policy issues in the textual content. As such it consists of the types campaign events, and thank you in combination with campaign events by Towner and Muñoz. Nevertheless, there are some documentation topics in which a documentary style stories mix with short quotes or text snippets bearing references to policy issues. While we discovered the documentary themes almost exclusively among text-integrated stories, we took a step back and investigated the post images without text, which had eluded the attention of our text-based approach. The vast majority of those images appear to be documentary-style images without any policy issues.

4. Discussion

We presented a pipeline to explore Instagram posts and stories using tools for textual analysis. By analyzing word frequencies, constructing word clouds, and employing topic models we were able to gain an initial understanding of the 2021 German federal election campaign. We found that posts tended to focus on policy issues while stories typically did not. Further, we found a large share of stories to document the campaign trail of candidates and campaign events, consistent with a previous study of the U.S. presidential campaign. The proposed pipeline provides a valuable tool for exploring the rapidly increasing amounts of visual social

media content posted during campaigns, despite topic modeling not yielding precise topics to differentiate between policy issues or messages types as finely grained as in previous work: For example, Haßler et al. collected 581 Instagram posts from the last month of the 2017 German federal election, while we have already gathered 713 posts comprising 1299 images and 2208 stories over the last fortnight of the 2021 campaign. Thus, while our results still call for a proper validation, e.g. through (computational) content analysis, we are able to give valuable first insights into stories used in a German election campaign. Consistent with Towner and Muñoz’s [14] findings about the 2020 U.S. presidential campaign, we found stories to mostly consist of the no policy issues types, several topics discovered consist of stories documenting the campaign, similar to campaign events and rallies being the most popular message type in the U.S. elections.

The majority of past analyses of the political communication on Instagram relied on content analysis taking into account both modalities, images and text, in some cases even video and audio of the Instagram content. Our approach concentrated on text, and overall BERTopic turned out to be a capable aid in exploring Instagram content, since the share of text-integrated images was overall high. Thus, we were able to use this text-based approach for the majority of the content available. Our approach may be applied in different domains as long as there is a large amount of text-integrated images. Through our experiments, however, one shortcoming became apparent: Parties and politicians used several catch-phrases and slogans over and over again. These eventually started to appear as their own topics, thus some policy issues, especially the ones of parties (like the CDU) extensively using the same or similar phrasing over and over, did not develop as their own topics.

4.1. Limitations & Future Work

One limitation of our research is that we have only focused on images, when in fact more than half of the stories were videos. This could potentially lead us to miss important content that may be part of the audio channel or text embedded in subsequent video frames. In order to address this issue, future studies should consider using automated transcriptions in order to access another layer of textual content which could then be integrated into the proposed pipeline. Similarly, we focus not on the image itself, but on the text as part of it. This means our analysis overlooks important visual cues and content present in the picture. We could image the use of automated image descriptions (e.g., produced by CLIP [38]) in conjunction with captions and image text to bridge the gap between our topic modeling results and the granularity of quantitative content analysis seen in prior work. A similar approach has already been demonstrated to be effective by Muralidhara and Paul using automated image tags. Further, we see potential in using BERTopic’s guided topic modeling capabilities in order to fine-tune the topics and allow for a more distinct separation between topics in order to capture policy issues and message types with higher validity. Such an approach has already been successfully used to study political topics in tweets [39]. Such a guided topic model could prove to be an invaluable asset for longitudinal studies of political communication on Instagram, a research desideratum that has been formulated in the literature.

References

- [1] Social Networks nach Nutzern 2022, ??? URL: <https://de.statista.com/statistik/daten/studie/181086/umfrage/die-weltweit-groessten-social-networks-nach-anzahl-der-user/>.
- [2] T. Leaver, T. Highfield, C. Abidin, *Instagram: Visual Social Media Cultures*, John Wiley & Sons, 2020.
- [3] J. Bast, Politicians, Parties, and Government Representatives on Instagram: A Review of Research Approaches, Usage Patterns, and Effects, *Review of Communication Research* 9 (2021). URL: <https://www.rcommunicationr.org/index.php/rcr/article/view/108>.
- [4] V. Mayer-Schönberger, *Delete: The Virtue of Forgetting in the Digital Age*, Princeton University Press, 2011.
- [5] X. Farkas, M. Bene, Images, Politicians, and Social Media: Patterns and Effects of Politicians' Image-Based Political Communication Strategies on Social Media, *The International Journal of Press/Politics* 26 (2021) 119–142. URL: <https://doi.org/10.1177/1940161220959553>. doi:10.1177/1940161220959553.
- [6] J. Haßler, A. S. Kumpel, J. Keller, Instagram and political campaigning in the 2017 German federal election. A quantitative content analysis of German top politicians' and parliamentary parties' posts, *Information, Communication and Society* (2021) 1–21. URL: <https://doi.org/10.1080/1369118X.2021.1954974>. doi:10.1080/1369118X.2021.1954974.
- [7] M. Lalancette, V. Raynauld, The Power of Political Image: Justin Trudeau, Instagram, and Celebrity Politics, *The American behavioral scientist* 63 (2017) 888–924. URL: <https://doi.org/10.1177/0002764217744838>. doi:10.1177/0002764217744838.
- [8] A. Olof Larsson, The rise of Instagram as a tool for political communication: A longitudinal study of European political parties and their followers, *New Media & Society* (2021) 14614448211034158. URL: <https://doi.org/10.1177/14614448211034158>. doi:10.1177/14614448211034158.
- [9] J. Bast, Managing the Image. The Visual Communication Strategy of European Right-Wing Populist Politicians on Instagram, *Journal of Political Marketing* (2021) 1–30. URL: <https://doi.org/10.1080/15377857.2021.1892901>. doi:10.1080/15377857.2021.1892901.
- [10] S. Boulianne, A. O. Larsson, Engagement with candidate posts on Twitter, Instagram, and Facebook during the 2019 election, *New Media & Society* (2021) 14614448211009504. URL: <https://doi.org/10.1177/14614448211009504>. doi:10.1177/14614448211009504.
- [11] A. Pineda, E. Bellido-Pérez, A. I. Barragán-Romero, “Backstage moments during the campaign”: The interactive use of Instagram by Spanish political leaders, *New Media & Society* 24 (2022) 1133–1160. URL: <https://doi.org/10.1177/1461444820972390>. doi:10.1177/1461444820972390.
- [12] M. Haim, M. Jungblut, Politicians' Self-depiction and Their News Portrayal: Evidence from 28 Countries Using Visual Computational Analysis, *Political Communication* 38 (2021) 55–74. URL: <https://doi.org/10.1080/10584609.2020.1753869>. doi:10.1080/10584609.2020.1753869.
- [13] V. Raynauld, M. Lalancette, Pictures, filters, and politics: Instagram's role in political image making and storytelling in Canada, *Visual communication quarterly* 28 (2021) 212–226. URL: <https://www.tandfonline.com/doi/full/10.1080/15551393.2021.1986827>. doi:10.1080/15551393.2021.1986827.

- [14] T. L. Towner, C. L. Muñoz, A Long Story Short: An Analysis of Instagram Stories during the 2020 Campaigns, *Journal of Political Marketing* (2022) 1–14. URL: <https://doi.org/10.1080/15377857.2022.2099579>. doi:10.1080/15377857.2022.2099579.
- [15] L. Manovich, *Cultural Analytics*, MIT Press, 2020.
- [16] C. Baden, C. Pipal, M. Schoonvelde, M. A. C. G. van der Velden, Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda, *Communication methods and measures* 16 (2022) 1–18. URL: <https://doi.org/10.1080/19312458.2021.2015574>. doi:10.1080/19312458.2021.2015574.
- [17] D. V. Shah, J. N. Cappella, W. R. Neuman, Big Data, Digital Media, and Computational Social Science: Possibilities and Perils, *The Annals of the American Academy of Political and Social Science* 659 (2015) 6–13. URL: <https://doi.org/10.1177/0002716215572084>. doi:10.1177/0002716215572084.
- [18] T. Araujo, I. Lock, B. van de Velde, Automated Visual Content Analysis (AVCA) in Communication Research: A Protocol for Large Scale Image Classification with Pre-Trained Computer Vision Models, *Communication methods and measures* 14 (2020) 239–265. URL: <https://doi.org/10.1080/19312458.2020.1810648>. doi:10.1080/19312458.2020.1810648.
- [19] E. Rodina, D. Dligach, Dictator’s Instagram: personal and political narratives in a Chechen leader’s social network, *Caucasus survey* 7 (2019) 95–109. URL: https://www.schoeningh.de/downloadpdf/journals/casu/7/2/article-p95_1.pdf. doi:10.1080/23761199.2019.1567145.
- [20] V. Murashka, J. Liu, Y. Peng, Fitspiration on Instagram: Identifying Topic Clusters in User Comments to Posts with Objectification Features, *Health communication* 36 (2021) 1537–1548. URL: <http://dx.doi.org/10.1080/10410236.2020.1773702>. doi:10.1080/10410236.2020.1773702.
- [21] S. Kim, H.-W. Lim, S.-Y. Chung, How South Korean Internet users experienced the impacts of the COVID-19 pandemic: discourse on Instagram, *Humanities and Social Sciences Communications* 9 (2022) 1–12. URL: <https://www.nature.com/articles/s41599-022-01087-7>. doi:10.1057/s41599-022-01087-7.
- [22] S. Muralidhara, M. J. Paul, #Healthy Selfies: Exploration of Health Topics on Instagram, *JMIR public health and surveillance* 4 (2018) e10150. URL: <http://dx.doi.org/10.2196/10150>. doi:10.2196/10150.
- [23] A. Al-Rawi, A. Al-Musalli, A. Fakida, News Values on Instagram: A Comparative Study of International News, *Journalism and Media* 2 (2021) 305–320. URL: <https://www.mdpi.com/2673-5172/2/2/18>. doi:10.3390/journalmedia2020018.
- [24] M. Bossetta, The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election, *Journalism & mass communication quarterly* 95 (2018) 471–496. URL: <https://doi.org/10.1177/1077699018763307>. doi:10.1177/1077699018763307.
- [25] J. W. Rettberg, Snapchat: Phatic Communication and Ephemeral Social Media, in: J. W. Morris, S. Murray (Eds.), *Appified: Culture in the Age of Apps*, University of Michigan Press, 2018, pp. 188–195.
- [26] M. Amancio, “Put it in your Story”: Digital Storytelling in Instagram and Snapchat Stories, Master’s thesis, Uppsala University, Disciplinary Domain of Humanities and Social

- Sciences, Faculty of Social Sciences, Department of Informatics and Media, 2017. URL: <https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1111663&dsid=-5700>.
- [27] L. Bainotti, A. Caliandro, A. Gandini, From archive cultures to ephemeral content, and back: Studying Instagram Stories with digital methods, *New Media & Society* (2020) 1461444820960071. URL: <https://doi.org/10.1177/1461444820960071>. doi:10.1177/1461444820960071.
- [28] J. Vázquez-Herrero, S. Direito-Rebollal, X. López-García, Ephemeral Journalism: News Distribution Through Instagram Stories, *Social Media + Society* 5 (2019) 2056305119888657. URL: <https://doi.org/10.1177/2056305119888657>. doi:10.1177/2056305119888657.
- [29] B. Li, O. K. M. Scott, M. L. Naraine, B. J. Rühley, Tell Me a Story: Exploring Elite Female Athletes' Self-Presentation via an Analysis of Instagram Stories, *Journal of Interactive Advertising* 21 (2021) 108–120. URL: <https://doi.org/10.1080/15252019.2020.1837038>. doi:10.1080/15252019.2020.1837038.
- [30] E. A. Nashmi, D. L. Painter, Oh Snap: Chat Style in the 2016 US Presidential Primaries, *Journal of Creative Communications* 13 (2018) 17–33. URL: <https://doi.org/10.1177/0973258617743619>. doi:10.1177/0973258617743619.
- [31] D. M. Blei, Topic Modeling and Digital Humanities, *Journal of Digital Humanities* 2 (2012). URL: <http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/>.
- [32] D. M. Blei, Probabilistic Topic Models, *Communications of the ACM* 55 (2012) 77–84. URL: <https://cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-models/fulltext>.
- [33] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure (2022). URL: <http://arxiv.org/abs/2203.05794>. arXiv:2203.05794.
- [34] R. Egger, J. Yu, A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts, *Frontiers in sociology* 7 (2022) 886498. URL: <http://dx.doi.org/10.3389/fsoc.2022.886498>. doi:10.3389/fsoc.2022.886498.
- [35] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning Word Vectors for 157 Languages, in: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018, pp. 3483–3887.
- [36] K. Liebhart, P. Bernhardt, Political storytelling on Instagram: Key aspects of Alexander Van der Bellen's successful 2016 presidential election campaign, *Media and communication* 5 (2017) 15–25. URL: <https://www.cogitatiopress.com/mediaandcommunication/article/view/1062>. doi:10.17645/mac.v5i4.1062.
- [37] A. Argyrou, S. Giannoulakis, N. Tsapatsoulis, Topic modelling on Instagram hashtags: An alternative way to Automatic Image Annotation?, in: *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, 2018, pp. 61–67. URL: <http://dx.doi.org/10.1109/SMAP.2018.8501887>. doi:10.1109/SMAP.2018.8501887.
- [38] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision (2021). URL: <http://arxiv.org/abs/2103.00020>. arXiv:2103.00020.
- [39] L. Hemphill, A. M. Schöpke-Gonzalez, Two Computational Models for Analyzing Political Attention in Social Media, *Proceedings of the International AAAI Conference on Web and Social Media* 14 (2020) 260–271.