M. Klein, D. Krupka, C. Winter., V. Wohlgemuth (Hrsg.): INFORMATIK 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023 821

Computergestützte Bildtypenanalyse durch Zero-Shot Klassifikation mit CLIP

Michael Achmann[1], Christian Wolff[1]

Abstract: Wir stellen unsere computergestützte Bildtypenanalyse mittels CLIP am Beispiel einer Analyse visueller *Social Media*-Inhalte vor und evaluieren diese. Dabei betrachten wir 2208 Instagram *Stories* aus dem Bundestagswahlkampf 2021, die in den letzten zwei Wochen des Wahlkampfs auf den Kanälen von acht Parteien und 14 Spitzenkandidierenden veröffentlicht wurden. Durch die Bildtypenanalyse konnten wir feststellen, dass der Großteil der *Stories* Wahlkampfveranstaltungen dokumentiert und in einem kleineren Teil Inhalte anderer Plattformen und Formate geteilt werden. Bei ca. einem Fünftel der *Stories* stehen sachpolitische Themen im Mittelpunkt. Die automatisierte Klassifikation hat insgesamt mittelmäßig funktioniert, ein Teil der Bildtypen konnte aber mit guter Performance klassifiziert werden. Deshalb sehen wir den Bedarf für systematisches *Prompt Engineering* und schlagen eine *Few-Shot / Ensemble* Klassifikation in zukünftigen Vorhaben zu testen, um die *Performance* über alle Bildtypen zu erhöhen.

Keywords: Visual Social Media; Zero-Shot; CLIP; Political Communication; Instagram Stories; Bundestagswahl 2021

1 Einleitung

Seit der Veröffentlichung von ChatGPT im November 2022 erregen große Sprachmodelle wie GPT-4, das am 14. März als Weiterentwicklung von GPT-3 erschien, öffentliches Interesse[2]. OpenAI, die Entwicklerfirma hinter ChatGPT, kündigte mit diesem Modell auch multimodale Fähigkeiten, insbesondere die Aufnahme visueller *Inputs*, an. Obwohl diese Fähigkeiten aktuell nur einem kleinen Kreis zugänglich sind, hat OpenAI bereits im Januar 2021 das neuronale Netz CLIP veröffentlicht [Ra21], das Zero-Shot-Klassifikationen von Bildern durch natürliche Sprache ermöglicht. Unsere Arbeit präsentiert Experimente zur Vorbereitung auf künftig verfügbare multimodale Modelle und erste Ansätze des *Prompt Engineerings* ([Yo22]) als Ergänzung zu bestehenden *Machine Learning*-Verfahren im Bereich *Computer Vision* für die Geistes- und Sozialwissenschaften.

Wir testen den CLIP-Ansatz zur automatisierten Klassifikation von Bildtypen für Instagram *Stories* der Parteien und Spitzenkandidierenden der deutschen Bundestagswahl 2021. Bildtypen kombinieren qualitative mit quantitativen Ansätzen [GA11] und wurden schon zur Untersuchung der politischen Kommunikation in Instagram Posts verwendet [HKK21; LB17].

[1] Universität Regensburg, Lehrstuhl f. Medieninformatik, Universitätsstr. 31, 93053 Regensburg, Deutschland
vorname.nachname@informatik.uni-regensburg.de

[2] https://openai.com/research/gpt-4, Aufgerufen am 25.04.2023.

1.1 Politische Kommunikation auf Instagram & Vergängliche Stories

Die politische Kommunikation auf Instagram wurde in den letzten Jahren mit diversen Ansätzen und Fragestellungen untersucht. Bast analysierte für ihre systematische Literaturübersicht 37 Studien und ordnete sie entlang des methodologischen Ansatzes. Dabei fiel ihr auf, dass die meisten Studien versuchen zu beantworten wer Instagram, wie und mit welchem Effekt nutzt [Ba21]. Die Studien zeigen einen unterschiedlichen Aufbau, einige vergleichen Kommunikationsstrategien zwischen Plattformen, andere zwischen Accounts, einzelne fokussieren sich auf einen einzigen Account. Insgesamt zeigt sich, dass Politikerinnen und Politiker Instagram nutzen, um ein überzeugendes Bild zu schaffen, indem sie die Wahlkampforganisation oder das Treffen mit Wählerinnen und Wählern dokumentieren statt sich mit politischen Themen auseinanderzusetzen. In einer jüngeren, europaweiten Studie zur Selbstdarstellung der Europaparlaments-Kandidierenden wurden computergestützte Verfahren unter anderem zur Auswertung des Aufnahmewinkels der Bilder genutzt [HJ21]. Darüber hinaus haben Towner & Muñoz 2022 eine erste Studie zur Nutzung von Instagram *Stories* im US-Wahlkampf 2020 veröffentlicht. *Stories* sind ein besonderes Format der Instagram-Plattform, das ursprünglich von Snapchat etabliert wurde, sie sind vergänglich und laufen nach 24 Stunden ab. Die Autoren sahen aus der Marketingperspektive mehrere Schwachstellen der Kampagne, wie verpasste Chancen, nutzergenerierte Inhalte zu teilen und eine inkonsistente Nutzung der Instagramkommunikationsnormen. Sie beobachteten, dass Wahlkampfveranstaltungen und Kundgebung die *Stories* dominierten [TM22].

In Anlehnung an die Arbeiten von Towner & Muñoz [TM22] und Haßler et. al. [HKK21] sehen wir den Bedarf, die vergänglichen Instagram-Inhalte aus dem Bundestagswahlkampf 2021 näher zu analysieren, um zu verstehen, welche Inhalte von den Akteuren geteilt werden und welche Kommunikationsnormen und Nutzungsmuster im deutschen *Social Media*-Wahlkampf genutzt wurden.

1.2 Computergestützte Bildanalyse & Distant Viewing

Um die wachsende Menge an visuellen *Social Media* Inhalten verarbeiten zu können, wollen wir computergestützte Methoden zur Analyse nutzen. Arnold & Tilton [AT19] haben 2019 ein grundsätzliches Framework für das *Distant Viewing* zur Arbeit mit großen visuellen Korpora vorgestellt. Ein Anwendungsfall für das *Distant Viewing* und damit auch ein Anwendungsfall für die Übertragung der bestehenden Werkzeuge und Code-Systeme ist die Analyse ikonischer Fotos. In einer systematischen Literaturübersicht zu ikonischen Bildern und computergestützten Verfahren beschreibt van Noord [No22] den *Semantic Gap* als eine der größten Herausforderungen für die *Computer Vision* (CV), besonders auch im Bereich der Geistes- und Kulturwissenschaften. Die kulturelle Lücke beschreibt dabei die fehlende Übereinstimmung zwischen den Informationen, die man aus visuellen Daten extrahieren kann, und den Interpretationen, die dieselben Daten für kulturelle Gruppen im Laufe der Zeit haben [No22]. Um die kulturelle Lücke zu schließen, bzw. zu minimieren, schlägt van Noord die Miteinbeziehung von Kontextinformationen vor, um Unsicherheiten zu überwinden. Während die einzelnen Bildtypen im vorgestellten

Vorhaben zunächst objektiv beschreibbar erscheinen und damit zur Analyse der politischen Kommunikation computergestützt klassifizierbar wirken, gibt es auch hier den Bedarf, auf Kontextinformationen zurückzugreifen, beispielsweise die im Bild eingebetteten Texte, die mit CV-Verfahren zunächst keine Beachtung finden. Wir sehen hier Parallelen zwischen den Herausforderungen der Geistes- und Kulturwissenschaften bei der computergestützten Bildanalyse und den Sozialwissenschaften, weshalb beide Disziplinen von methodischen Weiterentwicklungen profitieren können.

1.3 Zero-Shot Klassifikation mit CLIP & Prompt Engineering

Konkret wollen wir die Bildtypen mit Hilfe des CLIP-Modells automatisiert klassifizieren. CLIP wurde im Januar 2021 vorgestellt [Ra21], seitdem wurde es in unterschiedlichen Domänen getestet und evaluiert. Yong et. al. haben sich für ihre Arbeit im Bauingenieurwesen intensiv mit dem *Prompt Engineering* und der Performanz der *Zero-*, bzw. *Few-Shot*-Klassifikation unter Nutzung des CLIP-Netztes beschäftigt. Bei der systematischen Optimierung ihrer *Prompts* für die Klassifikation von Bildern mit Gebäudedefekten erzielten sie die besten Ergebnisse durch die Nutzung ganzer Sätze, die Nutzung von Domänenwissen und die Eingabe von Beispielbildern [Yo22]. In den Digital Humanities haben Smits und Kestemont mit der binären Klassifikation von Laterna-Magica-Glasplatten aus dem späten 19. Jahrhundert begonnen. Während die *Zero-Shot*-Klassifikation für die Kategorien innen / außen vergleichsweise gut funktionierte, erzielte ein *transfer learning*-Modell bessere Ergebnisse [SK21]. In einer weiteren Arbeit klassifizieren Smits und Kestemont Bilder aus einem niederländischen Kinderbuchkorpus. Die Autoren sehen einen *multimodalen Turn* in den Digital Humanities bevorstehen, und nennen auch das *Few-Shot*-Konzept, wie im obigen Beispiel durch Eingabe der Beispielbilder in das Modell umgesetzt, als Chance zur Klassifikationsverbesserung.

Zusammenfassend wollen wir mit Hilfe der Bildtypenanalyse die Inhalte der Instagram-*Stories* aus dem Bundestagswahlkampf 2021 untersuchen. Die Analyse bietet uns einen besseren Einblick in die Nutzung vergänglicher Medien durch politische Akteure und tragen zur Untersuchung der politischen Kommunikation in den sozialen Medien bei. Um die Bilder auszuwerten, nutzen wir ein computergestütztes Verfahren, die Zero-Shot-Klassifikation mit dem CLIP-Modell. Wir wollen in dieser Arbeit folgende Fragen beantworten: (1) Welche Bildtypen dominieren die Instagram *Stories* der Parteien und Spitzenkandidierenden im Bundestagswahlkampf 2021? (2) Wie gut können die Bildtypen mit Hilfe des CLIP-Modells im *Zero-Shot*-Verfahren computergestützt klassifiziert werden?

2 Methoden

Wir haben vom 13. bis zum 26. September 2021 insgesamt 2208 Instagram *Stories* von acht Partei-Accounts und 14 Spitzenkandidierenden gesammelt. Neben allen im Bundestag sitzenden sieben Parteien wurden noch die Freien Wähler als achte Partei aufgenommen. Darüber hinaus wurden für alle Parteien zwei Spitzenkandidierende ausgewählt, die Union

wurde bei der Auswahl als eine Gruppe angesehen. Die *Stories* wurden täglich um 00:00 (MESZ) mit Hilfe von Selenium gesammelt, das die menschliche Betrachtung der *Stories* simulierte. Im folgenden stellen wir einen Ansatz zu computergestützten Auswertung visueller Social Media-Inhalte mit Hilfe des OpenAI CLIP-Modells vor. Zunächst gehen wir auf die Bildtypenanalyse als Methode der quantitativen Inhaltsanalyse ein und stellen unsere aus der Literatur gewonnenen Bildtypen vor. Danach gehen wir auf das CLIP-Modell und den Klassifikationsprozess ein.

2.1 Bildtypen

Die quantitative Bildtypenanalyse kombiniert das qualitativ ikonografisch-ikonologische Vorgehen der Kunstgeschichte mit der quantitativen Inhaltsanalyse aus der Sozialwissenschaft. Ziel ist es, Bildelemente und Symbole sowie Bildinhalte und Bildmotive zu erfassen und daraus Bildtypen abzuleiten, die zur Interpretation des Korpus genutzt werden können. Grittman und Ammann stellen hierfür einen dreistufigen Ablauf vor: Zunächst findet die ikonografische Analyse statt, darauf aufbauend werden die Bildtypen gebildet und schließlich folgt die ikonologische Interpretation der Ergebnisse [GA11]. In der Kommunikationswissenschaft wurde diese Methode bereits zur Untersuchung von Instagraminhalten genutzt. Liebhart & Bernhardt nutzten den Ansatz, um das politische *Storytelling* auf Instagram am Beispiel des heutigen Bundespräsidenten Österreichs, Alexander Van der Bellen, zu untersuchen [LB17]. Haßler et al. [HKK21] bauen für Ihre Untersuchung der Bundestagswahl 2017 auf Instagram auf den Bildtypen der österreichischen Arbeit auf.

Diese bilden die Grundlage für unseren Ansatz. In einem ersten Schritt haben wir zur Vorbereitung die gesammelten Instagram *Stories* exploriert und 1104 Bilder auf Grundlage der Annotationsanleitung von Haßler et. al. [HKK21] kodiert. Während die meisten Typen übertragbar sind, haben wir keine Bilder aus der Kategorie Umfrageergebnisse gesichtet und diesen Bildtyp deshalb nicht übernommen. Einzelne Typen, z. B. *Negative Campaigning* oder *Call for Action*, konnten wir nicht übernehmen, da die korrekte Klassifikation von (Text-)inhalten abhängig ist. Mit der Beschreibung der Bildinhalte für unsere Vergleichsphrasen konnten wir keine zufriedenstellende Unterscheidung vom Bildtyp *Positioning* erreichen. Außerdem haben wir die Bildtypen um *Social Media Posts* erweitert, da *Stories* auch genutzt werden, um Beiträge und Inhalte aus anderen Plattformen (z. B. Twitter) oder Medientypen (Posts) zu übernehmen. Die ausgewählten Bildtypen und zugehörigen Vergleichsphrasen sind in Tabelle 1 dargestellt.

Bei Haßler et al. [HKK21] konnte ein Bild mit mehreren Bildtypen kodiert werden. Bei der menschlichen Kodierung konnten wir diese Entscheidung gut nachvollziehen. Um die Komplexität unseres Ansatzes zu reduzieren, gehen wir allerdings im Folgenden davon aus, dass jedem Bild genau ein Bildtyp zugeordnet werden kann. Dieser Schritt scheint durchaus kompatibel mit Grittmann und Ammanns Erwartung, „dass der einzelne Bildtyp intern homogen und extern heterogen ist.“ [GA11].

2.2 CLIP & Evaluation

Ziel unseres Projekts ist die computergestützte Klassifikation von Bildtypen. Zur maschinellen Klassifikation von visuellen Medien stehen verschiedene Verfahren zur Verfügung, vor allem aus dem *Deep Learning* Bereich. In der Regel setzen diese Ansätze die Existenz eines genügend großen Trainingskorpus voraus, also einer genügend großen Zahl von annotierten Daten. Daraus ergeben sich mehrere Probleme, beispielsweise der hohe Zeit- und Kostenaufwand um visuelle Medien in ausreichender Qualität zu annotieren, sowie eine begrenzte Übertragbarkeit von Trainingsdaten für spezifische Domänen.

Wir haben den CLIP-Ansatzes als möglichen Kandidaten für die Bildklassifikation für unseren Anwendungsfall identifiziert, da das Modell sogenannte *Zero-Shot* Klassifikation erlaubt. Das CLIP-Netz wurde auf einer Vielzahl von Bild-Text-Paaren trainiert, damit erlaubt es den Vergleich zwischen einem Eingabetext und einem Eingabebild [Ra21]. Das Vergleichsergebnis mündet in der *confidence*, einer metrischen Variable, die den Grad der Übereinstimmung des Bildes mit dem Text ausdrückt. Um nun eine Klassifikation mit Hilfe des CLIP-Modells vorzunehmen, wird eine Sammlung von Vergleichsphrasen angelegt. In einer Schleife werden für jedes Bild alle Phrasen mit dem Bild verglichen, zur Klassifikation wählen wir die Phrase aus, die die höchste *confidence* für alle Bild-Phrasen-Paare aufweist. Konkret wird eine *softmax*-Funktion angewandt und die Klassifikation mit der höchsten Wahrscheinlichkeit ausgewählt. In Anlehnung an [LTG22] wurden im Anschluss an die erste Kodierung insgesamt 21 Phrasen erstellt, worin auch die Erfahrungen aus der manuellen Kodierung einflossen. Die Phrasen wurden dann zur Klassifikation genutzt.

Zur Evaluation wurden die Klassifikationsergebnisse einem geschulten Annotator in LabelStudio zur Kontrolle vorgelegt. Er hat alle vorgeschlagenen Bildtypenlassifikationen korrigiert. Wie oben erwähnt, haben vergangene Arbeiten die Kodierung mit einem oder mehreren Bildtypen vorgenommen, in unserem Fall wurde die *softmax*-Funktion genutzt, um die Phrase mit der größten Ähnlichkeit zum Bild zu identifizieren, und davon abgeleitet genau ein Bildtyp klassifiziert. Entsprechend wurde bei der Korrektur der Entscheidung des CLIP-Ansatzes Vorrang gewährt.

3 Ergebnisse & Evaluation

In unserer Studie klassifizierten wir die visuellen Inhalte von 2208 Instagram *Stories* im Bundestagswahlkampf 2021. Die Zahl der Übereinstimmungen mit den Vergleichsphrasen der CLIP-Klassifikation und der daraus abgeleitete Bildtyp sind in Tabelle 1 angegeben und den Frequenzen der manuellen Korrektur gegenübergestellt. Für die inhaltliche Auswertung wird Bezug auf die korrigierten Ergebnisse genommen.

Zur Evaluation der computergestützten Bildklassifikation wurden die menschlichen Korrekturen erfasst und *Precision, Recall* und *F1-Score* als Performance-Metriken berechnet. Insgesamt betrachtet erreicht das CLIP-Verfahren über alle Vergleichsphrasen lediglich

Bildtyp	Vergleichsphrase	n-Phrase C	n-Typ C	n-Typ M
Campaign events	The main subject of the photo is on stage giving a speech	328	765	888
	A photo of a crowd gathered at a political rally	41		
	A photo of an event location without a focus on people	300		
	A photo of one or more people watching TV.	21		
	A photo documenting a political campaign event	55		
	A photo showing one or more people on stage during a discussion	9		
	A travel photo	9		
	A roadtrip photo	2		
Individual Voter Contact	A photo of a politician talking face to face with constituents	225	227	106
	A Selfie	2		
Media work	The main subject of the photo is being interviewed	80	124	146
	A photo of a TV studio	44		
Positioning	A person speaking to the camera	111	519	507
	A digital collage for an election campaign	408		
Call for action	A poster or image calling to vote or an invitation to an event	0		(116)
Campaign material	A photo of flyers or other campaign material for an election campaign	0	32	49
	A photo of posters or flyers	32		
Everyday political work	A photo documenting the parliamentary work of a politician	65	90	71
	An image documenting a party convention	25		
	A screenshot of a social media post	451	451	395
NA				49
Private background story	A photo showing the private life of a politician	0	0	(3)

Tab. 1: Übersicht der genutzt Bildtypen nach Haßler et. al. mit den dazugehörigen Vergleichsphrasen. In den hinteren Spalten sind die Vorkommen der Phrasen bei Klassifikation und die Zahl der daraus abgeleiteten Bildtypen für das CLIP-Modell (C), bzw. die menschliche Kodierung (M) angegeben. Der Bildtyp *Call for Action* wurde durch CLIP nur bei der manuelle Korrektur erfasst und entsprechend dem Bildtypen *Positioning* zugeordnet. Die wenigen *Private background stories* wurden unter NA aussortiert, vgl. Diskussion.

Bildtyp	**Precision**	**Recall**	**F1**
Social Media Posts	0.88	0.79	0.84
Campaign events	0.77	0.90	0.83
Positioning	0.73	0.73	0.73
Media work	0.57	0.69	0.62
Individual Voter Contact	0.89	0.42	0.57
Campaign Material	0.37	0.60	0.46
Everyday Political Work	0.39	0.33	0.36
Overall	0.66	0.64	0.63

Tab. 2: Performance-Metriken für die Bildtypenklassifikation mittels CLIP (Makro-F1) im Vergleich zum menschlich korrigierten Datensatz.

einen *Makro-F1-Score* von 0.60. Ein genauerer Blick in die Ergebnisse zeigt allerdings auch, dass sich der Klassifikationserfolg zwischen den Vergleichsphrasen stark unterscheidet: Die Phrasen *A photo of an event location without a focus on people* und *The main subject of the photo is on stage giving a speech* schnitten mit einem F1-Wert *Accuracy* 0.92 und 0.91 sehr gut ab, die Phrasen *A photo documenting the parliamentary work of a politician* (F1=0.37) und *An image documenting a party convention* (F1=0.29) hingegen sehr schlecht ab. Da die Performance für die Bildtypenklassifikation direkt von der Klassifikation anhand der Vergleichsphrasen abhängig ist, zeichnet sich ein ähnliches Ergebnis ab, siehe Tabelle 2.

Im Instagram *Story*-Wahlkampf ist somit der größte Teil der veröffentlichten Inhalte eine Dokumentation von Wahlkampfveranstaltungen. Diese Bilddokumentation zeigt sehr häufig Kandidierende auf der Bühne bei Reden und Blicke in das Publikum und Überblicke der Veranstaltungsorte. *Social Media Posts* war der zweithäufigste Bildtyp, dabei handelt es sich um *Stories*, die der Interaktion mit dem Publikum dienen, durch Teilen von Beiträgen anderer Accounts, oder Teilen von Ergebnissen über Instagram-Sticker. Darüber hinaus wurden in dieser Kategorie viele Screenshots von eigenen Beiträgen auf anderen Plattformen wie Twitter beobachtet, zum Teil überschnitten sich diese Inhalte mit dem Bildtyp *Positioning*. Dieser war der dritthäufigste, hier wurden *Stories* kategorisiert, die sachpolitische Inhalte aufgriffen und über Share-Pics, Text-Integrierte *Stories* oder Videos vermittelt wurden. Bei der *Media Work*-Kategorie konnten viele Interviews und Ausschnitte aus dem Triell als besonderer Kommunikationssituation zwischen drei Kanzlerkandidat:innen 2021 beobachtet werden, *Individual Voter Contact* fand primär bei Wahlkampfveranstaltungen statt. Aufgrund der Beobachtung in den letzten Tagen des Wahlkampfs wurden vergleichsweise wenige *Stories* des Typs *Everyday Political Work* gesichtet, hier stach die FDP mit ihrem Bundesparteitag heraus. Am seltensten wurden *Campaign Materials* geteilt, oft wurden hier Bilder von Wahlkampfständen kleinerer Ortsgruppen gezeigt.

Zusammenfassend hat die automatisierte Klassifikation der Bildtypen für die Typen am besten funktioniert, für die die detailliertesten[3] Vergleichsphrasen entwickelt wurden. Durch die Klassifikation der *Stories* zeigt sich ein starker Fokus der Parteien und Kandidierenden auf der Dokumentation des Wahlkampfs und dem Teilen von Inhalten anderer Plattformen oder Kanäle, der durch die manuelle Korrektur bestätigt werden konnte. Darüber hinaus gab es (nur) bei rund einem Fünftel der *Stories* einen sachpolitischen Fokus.

4 Diskussion

Unsere Auswertung zeigt vergleichbare Ergebnisse zu den Instagram-*Stories* im U.S. Präsidentschaftswahlkampf, auch dort dominierten die Wahlkampfveranstaltungen und Kundgebungen. Towner & Muñoz [TM22] haben darüber hinaus aus der Marketing-Perspektive verpasste Chancen durch Einbindung von nutzergenerierten Inhalten gesehen,

[3] Detailliert ist hier einerseits im Sinne einer möglichst präzisen Beschreibung der Bilder zu verstehen, andererseits als die Abdeckung eines Bildtyps durch mehrere Phrasen.

hier ist die große Zahl der *Social Media Posts* in unserem Korpus näher zu betrachten: Dieser Bildtyp wurde abweichend von allen anderen Bildtypen nicht aus der Literatur herausgearbeitet. Bei der Evaluation hat sich gezeigt, dass CLIP hier sehr präzise geteilte *Stories*, *Posts* und *Screenshots* von eigenen oder anderen Profilen sowie Instagram und anderen Plattformen klassifiziert hat. Bilder dieses Typs sollten in zukünftigen Arbeiten tiefgehender analysiert werden, um herauszuarbeiten, in welchem Umfang es sich um nutzergenerierten Inhalt handelt. Dies könnte teils automatisiert durch den Blick in die Metadaten erfolgen. Die *Stories* dieses Bildtyps zeigen eine weitere Limitierung des Verfahrens auf: Die Bildtypen der Literatur waren darauf ausgelegt, dass ein oder mehr Bildtypen einem Post zugeordnet werden können, ein Teil der geteilten *Social Media*-Inhalte hätte inhaltlich auch einem weiteren Bildtyp zugeordnet werden können. Ein anderer Teil der Bilder hätte allerdings auch durch den Menschen nicht spezifischer eingeordnet werden können, beispielsweise Mitmachspiele auf dem CDU-Account.

Bei der Korrektur der Ergebnisse zeichneten sich auch Muster für besonders herausfordernde Klassifikationen ab: Die Differenzierung zwischen *Call for Action* und *Positionierung* ist mit unseren Vergleichsphrasen nicht möglich gewesen. Die meisten Bilder, die eine Aufforderung zur Wahl oder Veranstaltungseinladung darstellten, wurden naheliegenderweise als „A digital collage for an election campaign“ klassifiziert. Gespräche mit Wählerinnen und Wählern wurden darüber hinaus zu oft erkannt. Außerdem konnten wir bei der Annotation der *Stories* noch Bilder finden, die nicht optimal unter den bestehenden Bildtypen eingeordnet werden können: Es gibt mehrere *Behind the Scenes*-Aufnahmen, bei denen Unterstützer:innen nach einem erfolgreichen Wahlkampftag noch etwas gemeinsam trinken. Weiterhin wurden mehrere Events von einzelnen Parteimitgliedern moderiert. Da die Personen dabei in die Kamera gesprochen haben, wurden diese Szenen als Positionierung klassifiziert, allerdings ist der sachpolitische Kontext hier nicht immer gegeben. Liebhart & Bernhardt [LB17] hatten auch noch den Bildtyp *Site Visits* genutzt, auch in unserem Korpus gab es mehrere Dokumentation von Besuchen bei Unternehmen und Interessenvertretungen, die durch den Bildtyp *Campaign Events* nicht exakt repräsentiert werden können – wobei die Klassifikation der Besuche und Moderation mit CLIP eine weitere Herausforderung darstellt. Im Vergleich zu früheren Arbeiten konnten wir keine Einblicke in das Privatleben der Politiker:innen beobachten, Robert Habecks *Stories* zum Besuch seines ehemaligen WG-Hauses bei der Wahlkampftour wären die einzigen Kandidaten gewesen, aufgrund der Uneindeutigkeit haben wir die Bilder nicht in die Analyse einbezogen.

Einige Bildtypen konnten nicht zufriedenstellend klassifiziert werden und gerade die Unterscheidung zwischen Bildern mit sachpolitischen Inhalten und *Call for Action* sowie die korrekte Einordnung des indivduellen Wählerkontakts hat schlecht funktioniert. Darüber hinaus wurde in der vorliegenden Studie primär das Bild betrachtet, ein Großteil der gesichteten *Stories* beinhaltet Bildtext, dessen Inhalte weitere wichtige Anhaltspunkte für eine korrekte Bewertung der *Stories* geben – interessanterweise haben die Ergebnisse zum Teil den Eindruck erweckt, als ob bestimmte Stichwörter (z. B. "Plenar-") die CLIP-Klassifikationen beeinflusst haben. Wir haben lediglich einen Zwei-Wochen-Zeitraum

betrachtet, unsere Aussagen lassen sich deshalb nicht auf den gesamten Wahlkampf übertragen. Darüber hinaus haben wir nur *Stories* betrachtet und auch hier nur Bilder oder den ersten *frame* aus Videos, wobei Videos rund 56% unsere *Story*-Korpus ausmachen. Durch die Bevorzugung der CLIP-Ergebnisse bei mehrdeutigen Bildern hat unsere Evaluation einen Bias. Um die Mehrdeutigkeiten zu reduzieren könnten künftige Arbeiten Bildtexte und Bildinhalte klarer getrennt betrachtet, z. B. durch Schwärzung der Texte für CLIP und menschlichen Annotator.

Aufbauend auf unsere Ergebnisse sehen wir den Bedarf, das *Prompt Engineering* weiter zu systematisieren und für die Bildtypenklassifikation *Ensemble*- und *Few-Shot*-Ansätze zu testen und nutzen. Dabei werden neben den Suchphrasen noch einzelne Vergleichsbilder als *Input* in das Model gegeben, in der Literatur wurde die Klassifikationsperformance dadurch stark erhöht. Für eine gesamtheitliche Betrachtung der multimodalen *Social Media* Inhalte sollten die Bildtexte und Metadaten sowie Audioinhalte bei Videos für eine valide Klassifikation zusammen mit den Bildern (bzw. *frames*) genutzt werden. Abschließend sollten zukünftige Arbeiten auch die *Posts* der Akteure beinhalten.

Zusammenfassend dominiert der Bildtyp *Campaign Events* die *Stories* aus dem Instagramwahlkampf deutlich. Außerdem gibt es viele geteilte Inhalte aus anderen Medien und Formaten und ca. ein fünftel der *Stories* mit sachpolitischen Bezügen (*Positioning*). Die automatisierte Bildklassifikation mittels CLIP hat insgesamt mittelmäßig funktioniert, bei detaillierter ausgearbeiteten Bildtypen aber durchaus eine gute *Performance* erzielt. Um Bildtypen zukünftig zuverlässig mit CLIP zu klassifizieren, ist eine systematische Erarbeitung der Vergleichsphrasen notwendig, außerdem sollten *Few-Shot & Ensemble* Lösungen näher betrachtet werden.

Literatur

[AT19] Arnold, T.; Tilton, L.: Distant viewing: analyzing large visual corpora. en, Digital scholarship in the humanities 34/Supplement_1, S. i3–i16, 2019, ISSN: 2055-7671, 2055-768X, URL: https://academic.oup.com/dsh/article/34/Supplement_1/i3/5694340.

[Ba21] Bast, J.: Politicians, Parties, and Government Representatives on Instagram: A Review of Research Approaches, Usage Patterns, and Effects. en, Review of Communication Research 9/, Juli 2021, ISSN: 2255-4165, 2255-4165, URL: https://www.rcommunicationr.org/index.php/rcr/article/view/108.

[GA11] Grittmann, E.; Ammann, I.: Quantitative Bildtypenanlyse. In (Petersen, T.; Schwender, C., Hrsg.): Die Entschlüsselung der Bilder: Methoden zur Erforschung visueller Kommunikation : ein Handbuch. von Halem, S. 163–178, 2011, ISBN: 9783869620435.

[HJ21] Haim, M.; Jungblut, M.: Politicians' Self-depiction and Their News Portrayal: Evidence from 28 Countries Using Visual Computational Analysis. Political Communication 38/1-2, S. 55–74, 2021, ISSN: 1058-4609, URL: https://doi.org/10.1080/10584609.2020.1753869.

[HKK21] Haßler, J.; Kümpel, A. S.; Keller, J.: 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/, S. 1–21, Juli 2021, ISSN: 1369-118X, URL: https://doi.org/10.1080/1369118X.2021.1954974.

[LB17] Liebhart, K.; Bernhardt, P.: Political storytelling on Instagram: Key aspects of Alexander Van der Bellen's successful 2016 presidential election campaign. Media and communication 5/4, S. 15–25, 2017, ISSN: 2183-2439, 2183-2439, URL: https://www.cogitatiopress.com/mediaandcommunication/article/view/1062.

[LTG22] Lucas, L.; Tomás, D.; Garcia-Rodriguez, J.: Exploiting the Relationship Between Visual and Textual Features in Social Networks for Image Classification with Zero-Shot Deep Learning. In: 16th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2021). Springer International Publishing, S. 369–378, 2022, URL: http://dx.doi.org/10.1007/978-3-030-87869-6_35.

[No22] van Noord, N.: A survey of computational methods for iconic image analysis. Digital Scholarship in the Humanities 37/4, S. 1316–1338, 2022, ISSN: 2055-7671, URL: https://academic.oup.com/dsh/article-pdf/37/4/1316/46607811/fqac003.pdf.

[Ra21] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision./, 2021, arXiv: 2103.00020 [cs.CV], URL: http://arxiv.org/abs/2103.00020.

[SK21] Smits, T.; Kestemont, M.: Towards multimodal computational humanities. Using CLIP to analyze late-nineteenth century magic lantern slides. In: CEUR Workshop Proceedings. Amsterdam, Netherlands, 2021, URL: http://ceur-ws.org/Vol-2989/short_paper23.pdf.

[TM22] Towner, T. L.; Muñoz, C. L.: A Long Story Short: An Analysis of Instagram Stories during the 2020 Campaigns. Journal of Political Marketing/, S. 1–14, 2022, ISSN: 1537-7857, URL: https://doi.org/10.1080/15377857.2022.2099579.

[Yo22] Yong, G.; Jeon, K.; Gil, D.; Lee, G.: Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. en, Computer-aided civil and infrastructure engineering/, 2022, ISSN: 1093-9687, 1467-8667, URL: https://onlinelibrary.wiley.com/doi/10.1111/mice.12954.