Accessible User-Generated Social Media for People with Vision Impairments

Cole Gleason

CMU-HCII-20-110
December 2020

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Jeffrey P. Bigham (Co-chair)
Kris M. Kitani (Co-chair)
Patrick Carrington
Chieko Asakawa (CMU & IBM Research)
Meredith Ringel Morris (Microsoft Research)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This work was supported by a NSF Graduate Research Fellowship as well as grants from National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR) and Google.

Keywords: Social media, image descriptions, accessibility, vision impairments, screen readers, memes, GIFs, accomodations, blindness, Twitter.

For my wife, Laura.

Abstract

Social media platforms are becoming less accessible to people with vision impairments as the prevalence of user-generated images and videos increase. For example, over 25% of content on Twitter contains visual media, but I have found that only 0.1% of images contain descriptions for people with vision impairments. Through interviews with some of the few sighted social media users who currently write image descriptions, I've uncovered that poor feature design and a lack of user education is stymying efforts to increase accessible content on social media platforms.

Some unique categories of media on these platforms, such as memes and animated GIFs, are hard to describe while maintaining their humorous or emotive effects. I explored alternative methods using audio to convey this media in richer non-visual format beyond alternative text, and built a system to make these accessible by re-using templates created by online volunteers. While audio-based methods should not replace textual descriptions of visual media, they can add a new, richer method to convey a similar tone and increase understanding.

To address the seemingly insurmountable problem of making all of this user-generated content accessible, I built and deployed Twitter A11y to demonstrate and evaluate multiple methods for sourcing image descriptions including text recognition, automatic image captioning, and human crowdsourcing. Participants with vision impairments who used Twitter A11y saw a drastic increase in accessible content on their accounts, with every image having a description and majority being high-quality.

By combining rich human descriptions and automatic methods, my work seeks to make visual media on social media platforms accessible at scale. Through automatic methods we can close the accessibility gap on this platform by rehabilitating inaccessible content, while still working towards the ultimate goal of helping original content authors create accessible content from the start. This work recommends that social media platforms and researchers enact a model of shared responsibility for the deluge of inaccessible content on technology platforms, requiring all actors to work towards more inclusive online spaces for people with disabilities.

Acknowledgments

Throughout the pursuit of my degree and the completion of this thesis, I have been lucky to receive the support of many people, and I would like to acknowledge that they helped to make this work possible.

First, I would like to thank my advisors, Jeff Bigham and Kris Kitani, who supported and guided my research, while providing advice throughout my entire Ph.D. career. Their mentorship has helped me develop into a researcher and I look forward to continue working with them in the future! The other members of my committee have also contributed to and guided this work: Chieko Asakawa, Meredith Ringel Morris, and Patrick Carrington. I could not have done this without all of you.

In addition, I would like the acknowledge the many lab mates, collaborators, and colleagues that worked with me. This includes the members of both of my advisor's labs, the Cognitive Assistance Lab, Microsoft Research Ability Team, and everyone else I have collaborated with over the years. They may have contributed to work in this thesis, other research we completed together, or helped me in other ways!

I want to thank my friends and family, especially my parents who provided the much needed guidance throughout life that lead to me attending CMU in the first place. My fellow Ph.D. students and friends Alexandra To, Judeth Oden Choi, Nathan Hahn, Qian Yang, Rushil Khurana, Michael Madaio, Joseph Seering, and many others have helped me persevere through some of the harder moments during the past five years. Thank you all, and I hope we are able to see each other at many conferences in the future!

Finally, I am writing this in the height of a pandemic after months of remote research and writing from home without seeing many people in person. There was no way I would have been able to do that and accomplish this work without the support of my partner, Laura Licari, who married me even as our wedding plans were torn to shreds by COVID-19. I will be forever grateful for her support of, and patience with, my research goals. Last but not least, I would be remiss not to acknowledge our two cats, Turing and Lovelace, who turned out to be exactly the kind of animal companions you need during a pandemic when you cannot spend time with other people out of your household.

Contents

Part I

Introduction and Background

Chapter 1

Introduction: Access to the Digital Public Square

In the past two decades, social media platforms have arisen as venues for all sorts of discourse, but are quietly excluding many people with disabilities due to inaccessible content. Twitter, Facebook, Reddit, Instagram and others serve as a digital semi-public square for people to share news, join around common interests, or start political movements. Because the platforms all prominently feature rich visual media (*i.e.,* images, animations, videos), users with vision impairments are not able to fully participate. Inaccessible content is also disrupting private communication as one-to-one messaging apps additionally include emojis, stickers, animations, and more. I investigate how social media platforms can ensure that user-uploaded content can be made accessible for users with vision impairments through various methods of generating and sourcing image descriptions.

1.1 Importance of social media for people with disabilities

Social media platforms have become more important to political and professional life recently. Both heads of state and local politicians frequently announce policies and converse with each other through tweets. Social media, and Facebook in particular, was used to organize the Egyptian revolution of 2011 [98]. Twitter served as an organizing tool for disability activists during the #CripTheVote and #HandsOffMyADA political campaigns in US [8, 29]. An active social media profile is also critical for many job profiles, as it serves as a networking, marketing, and public relations platform all in one. Any lack of access to these platforms is not only preventing users with disabilities from accessing humorous memes or information about their friends' daily lives, but it is impeding their ability to fully participate as a citizen and professional in society.

While there are serious implications in excluding people with disabilities from critical professional or political infrastructure, I also do not want to minimize social media's usage for humor and recreation. Even when content like memes, jokes, or photos of innocuous daily life fails to reach the level of "political movement", that content is still important to someone's family, friends, and online followers! Recreation tends to be left out of many accessibility considerations [51], implying that fun or leisure activities are optional. People with disabilities deserve

equal access to use social media, no matter how society views the utility of the content consumed or posted. As a blind participant evaluating memes (Chapter 6) said "If there has to be a lot of useless content out there, it ought to be accessible".

1.2 Investigating social media accessibility through Twitter

Social media was not always an inherently visual medium. In fact, when Twitter launched in 2006, it was entirely textual in nature. But over the following 10 years, visual media crept up to over 25% of Twitter's English-language content [72]. Facebook has similarly seen an increase in images and videos on the site with hundreds of millions of photos uploaded per day. Recent applications like Instagram, Snapchat, and TikTok are entirely centered around images or videos.

Realizing that blind Twitter users were experiencing less and less accessible content over time, Twitter introduced an "image description" feature in 2016 so that users could add descriptions to their images. The core method of making content accessible is translating it from a visual medium (images) to an accessible alternative format (typically text or audio). Image descriptions, often called alternative text online, has served this purpose for decades on websites. They describe the visual content of an image, and may differ from a caption that avoids describing content that sighted readers can see in an image.

In my investigation two years after launch (Chapter 3), almost no one was adding image descriptions to their tweets with images, and only 0.1% of image tweets have any description added at all. Some accounts had a higher percentage of alternative text, including accounts of US congresspeople and accounts followed by blind users, but less than 5% of images were described in either case. This indicates that almost all images on Twitter are inaccessible to people with vision impairments, and as the amount of visual media on the platform increases this will only worsen. There is a wide accessibility gap between sighted and blind users on Twitter and other social media platforms.

Why do so few users add image descriptions? I conducted interviews with sighted Twitter users who add alternative text to their image posts, and the main factor limiting them from adding more alternative text is that the image description feature is hard to discover and use on Twitter and other platforms. Once the feature is discovered and users know why they should add alt text, many are not aware of best practices or what comprises good alternative text, especially for image types like memes or art. For alternative text that was written by human authors, less than 50% earned the highest rating of quality in our analysis. The greatest improvement social networking sites could make to increase the amount of high-quality descriptions is to support and educate users on why and how to add it to their own posts.

Facebook has battled the enormous scale of inaccessible images through automation. Wu et al. deployed a system to automatically describe objects and tags in the image, as long as the system is reasonably confident those objects exist in the image [112]. These automated descriptions list recognized objects and tags in a list. For example, "Image may contain: person, tree, text". In my interviews, blind users reported that these automated lists are better than no descriptions, but they fail to come close to fully describing the image. They are useful for a quick glance, however, and often clue blind users that another tool may be more useful. When "text" is listed in the image description, that indicated that an optical-character recognition application

may yield good results when extracting the text. Facebook also includes the names of recognized friends if they are tagged in the image.

Other social networks, like Reddit, remain inaccessible because the platforms do not provide the opportunity for alternative text to be added to images posted by users at all. My investigation into this accessibility issue on social media indicated the first steps platforms should take are to (1) make it possible for users to add alternative text, (2) make it easy for users to discover and understand how to add alternative text, and (3) provide incentives and reminders for users to add alternative text when posting content.

1.3 Automated tools to increase the quality and quantity of image descriptions

Ensuring that all visual content on social media contains an accessible format such as alternative text is a great first step, but this content should also be accurately and fully described to ensure complete access to the image. If we hope to increase the quantity of alternative text on the platform by encouraging unfamiliar users to add it, there also must be a way to ensure the descriptions are high-quality. Thus, social media platforms need to measure the quality of alternative text being written by their users. Image description novices are likely unaware of accessibility best practices, as the most well known are technical documents intended for web programmers [20], To address this, I developed a tool to provide an automated assessment of description quality and give real-time feedback to content creators on how they can improve (Chapter 5). Provided a social media post with an image, HelpMeDescribe automatically updates feedback as the novice user composes their image description. The inclusion of HelpMeDescribe on social media platforms, as well as other places where image descriptions are input, will increase the quality of user-provided descriptions and train users to write high-quality descriptions in the future.

Increasing the accessibility of social networks solely by educating billions of users is an arduous process that will take time, even with feedback. The automatic alt text employed by Facebook is useful, but not entirely accurate or descriptive. Could we combine automatic ways to reuse human-written alt text, and automated methods of generating alt text to make many images accessible?

I developed Twitter A11y to bring the percent of accessible images on Twitter from 0.1% to close to 100% (Chapter 4). Users install a browser extension that fetches or generates alternative text in real time as they browse the Twitter website. The extension progresses through six different methods including automatic image captioning, text recognition, reusing captions from around the web, and crowdsourcing. Methods that are fast, cheap, and produce high-quality results are emphasized first, while slower and more expensive methods like crowdsourcing are last. Before resorting to crowdsourcing, Twitter A11y can fulfill 80% of alt text requests with the first 5 methods. In user interviews, blind participants emphasized the value of automatic text recognition for text-heavy images, and admired the ability of automatic image captioning to provide at least some descriptions for most images. Twitter A11y has now been launched publicly both as a browser extension and as a automated Twitter user account that responds to any description request across the platform, no matter what device the user is accessing it from.

1.4 Some native social media content is hard to describe well

For some types of social media content, like image memes and animated GIFs, both automated and human attempts to make the visual content accessible struggle to convey the humor or emotional tone present. Figure 1.1 displays an image macro meme called "Success Kid", where different text is overlaid on top of a background image of a toddler. A simple description of the image may not convey why the toddler is relevant, and in fact may confuse the reader. This is partially because the background image encodes a template that the reader learns through repeatedly seeing the "Success Kid" meme online; in this case the template is "tiny triumphs/victories, exaggerated celebrations".

Figure 1.1: Example of a Success Kid meme.

The repeated background image of an image macro meme helps a sighted user recall this template, but I propose to harness it to make memes accessible to screen readers (Chapter 6). I use the repeated visuals to automatically match images to pre-written alt text templates explaining the meme. Recognizing that explaining the template directly may negatively affect the humor or emotion encoded in the meme's template, I also developed audio templates that use sound effects to convey a similar template.

In Chapter 6, I detail our work with blind participants to evaluate these different "translations" of image macro memes into text or audio content. We found that participants reported a preference for alternative text formats, partially because it comfortably integrated with existing screen reader software. However, the audio templates did not impair the understandability of the memes, and may even improve them. The blind participants we worked with were very excited by this work, and emphasized that access to even the silly social media content is important. When I asked about other forms of media online, surprisingly most participants said they were not as concerned with videos, as a lot of content does not need an audio description to be understandable. They reported that animated GIFs were a much larger problem on sites like Twitter.

GIFs are silent, looping animations that typically last only a few seconds. Their popularity has been rising on both social media platforms and messaging applications. Typically, GIFs are used as either a reaction to someone else's post, and often feature movie or TV characters speaking dialogue, or contain facial expressions intended to a convey an emotion. GIFs can also be used to as moving meme, similar to an image macro meme, where text is overlaid on top of a short video.

As these GIFs are silent, they are inaccessible to blind users if image descriptions are not provided. This is not supported on many platforms, although Twitter introduced the ability to add descriptions for GIFs in January 2020. I analyzed the prevalence of user-authored descriptions of these (Chapter 7) a few months after the feature launched, finding that few were described (0.3%), similar to images. Through interviews and an evaluation of different accessible GIF formats with blind Twitter users, I identified important visual information in GIFs to describe in an effort to ensure alt-text meets the needs of visually impaired users.

Figure 1.2: Four frames extracted from a popular reaction GIF. Sourced from an Apple promotional video from the early 1990s, a child is seen using a Mac computer before turning to the camera to give a thumbs up. The GIF is now commonly used online to denote approval.

Based on these discussions of important elements to describe, such as characters and action occurring in the GIF, I then designed alternative text descriptions to fulfill those requirements. However, traditional alternative text may not be expressive enough to convey the visual content of a GIF, depending on the specific content of the GIF. Therefore, I explored if GIFs excerpted from other media, such as TV shows, could be made more accessible by restoring their original sound and adding voiceover audio descriptions. This is very helpful when a GIF contains dialogue, as it is spoken in the characters voice which may be recognizable to a familiar audience. Like image memes, blind participants thought that alternative text must always be present as an accessible option, but a majority were interested in using audio descriptions for GIFs as a richer experience.

My investigation into the accessibility of memes and GIFs indicates that there are pockets of content on social media that are both poorly supported by general image description practices and ripe for re-using prepared alternatives. Most instances of memes and GIFs are not unique and creation of new ones is less common that sharing existing popular instances. Therefore, using the insights from prior work and discussions with blind Twitter users, platforms could develop libraries of accessible popular GIFs. Twitter has recently launched a similar feature for GIFs shared from their integrated library, although the descriptions are poor and do not typically convey the visual contents of the GIF.

1.5 Dual investment in human and automated approaches for accessible content

My experience with developing accessibility solutions for the images and animated GIFs on social media platforms indicates that there is no one apprach that will achieve both high-quality and scalable alternative text for all forms of media. However, certain types of visual content can be retrofitted to achieve accessibility goals, such as memes, GIFs, or other kinds of images where automated approaches succeed (e.g., text recognition for screenshots). Social media platforms, and in fact all technologies involved in the creation and sharing of digital media, must pursue both methods that increase the true accessibility investment from content creators and scalable retrofitting accommodations to address the wide accessibility gap present between sighted and visually impaired users.

To apply these dual approaches and recognize the media types they best apply to, technology platforms should note what these approaches lack: content knowledge or accessibility knowl-

edge. Content knowledge, in this case, is the understanding of the visual contents of an image as well as surrounding context such as the intent of the poster or information that is not captured in the image itself such as the photographer. Accessibility knowledge is the understanding of what elements are most important for people with vision impairments and how to structure the content non-visually to best convey this information (e.g., ordering info in alt text or choosing to use an audio description). Human authors, especially the original content posters, typically have rich content knowledge but may lack accessibility knowledge without a system like HelpMeDescribe. Automated image captioning methods can be trained on accessibility knowledge to format descriptions appropriately, but will likely always lack content knowledge compared to a human author. Platforms can seek to augment automated approaches with human effort or train humans to be better image describers, thus creating approaches with both high content and accessibility knowledge.

This work represents a study in a already wide, but still worsening, gap in accessibility for people with vision impairments in online communication. Failure to make user-generated content on these platforms accessible is already excluding groups of people based on their disability, and we must develop approaches to close that gap, preferably by increasing the amount of content made accessible by the original poster. While the specific elements of my investigations are primarily focused on images on social media, the lessons of human approaches versus automation and content knowledge versus accessibility knowledge can broadly apply to other technology that involves the creation of digital media inaccessible to people with disabilities. By pursuing dual investment in approaches that increase initial accessibility and automated alternatives to scalably retrofit existing content, we can ensure that technology-mediated spaces provide equal access to everyone with a disability.

Chapter 2

Background and Research Challenges

My work on making social media accessible for people with vision impairments is related to (1) online accessibility for images, (2) automatically generating alternative text, (3) crowdsourcing alternative text, and (4) image accessibility on social media platforms.

2.1 Online accessibility for people with vision impairments

People with vision impairments primarily access the Internet through screen reader software or Braille displays. Screen readers read the content of the computer's user interface aloud, and provide interaction mechanisms to traverse the hierarchical structures of applications and web sites. An early and widely-adopted screen reader for graphical web content was the IBM home-page reader [6], but now people with vision impairments may use a variety of screen readers on their desktop or laptop computers, Talkback on Android devices, or VoiceOver on Apple phones, tablets, and even smartwatches. Screen readers on desktop or latptop computers primarily provide interaction through a keyboard, while devices with touchscreens often provide much of the navigation interactions through swipe or tap gestures. All screen readers provide customizable output, such as setting a reading speed or changing the voice used.

People with vision impairments, especially those who are deaf or hard of hearing (DHH), may use a refreshable Braille display instead. The displays are composed of one or more lines of cells, each of which can display a Braille character. The user moves their finger along the line, left to right, to read the display, which refreshes to display the next line of text. These devices are especially useful situationally, such as reading or note taking during a lecture where the user may not want to listen to a separate audio stream. Blind web developers may also prefer a Braille display to more carefully inspect punctuation and whitespace important to writing computer programs [3].

Image descriptions, often referred to as alternative text or "alt text", are captions for images online or in other software. Screen readers and Braille displays read these when the user encounters the image, and they are intended to replace the visual content that a sighted reader might need to understand the document [58]. Alternative text is most commonly encountered on webpages, as it was added to the HTML 2.0 specification in 1995 [9]. However, at the time it was intended for users who used non-graphical browsers or preferred not to render images when accessing

the web. Visually impaired users are not mentioned as an intended audience until the HTML 4.0 specification in late 1997 [21]. Accessibility of images for screen-reader users is one of the most commonly cited reasons to add alternative text to images today, but it is also recommended in case the image does not load for sighted users. Image descriptions are also expected to be added to software in other domains, including mobile phone applications on the iOS [5] and Android [40] platforms. Commercial software can also add alternative text to images in documents, such as Microsoft Word and Adobe Acrobat [2, 65].

Alternative text is not solely meant for human consumption, as various search engines consume it and use it to rank pages [53]. In fact, image descriptions have been used for a number of different applications including "semantic visual search, visual intelligence in chatting robots, photo and video sharing in social media, and aid for visually impaired people to perceive surrounding visual content" [46]. Image labels, captions, and descriptions provide a solid foundation for many of these kinds of applications.

The majority of alternative text is written manually by website developers or authors of the website content. While authors are recommended to follow Web Content Accessibility Guidelines [20], much of image content on the web does not contain image descriptions. In a historical analysis of websites from 1997-2002, Hackett *et al.* found that websites were getting increasing complex and less accessible [45]. In 2006, Bigham *et al.* found that less than 40% of significant images on the top 500 high-traffic websites contained alternative text [10]. This motivated the authors to create a tool to generate alternative text automatically from surrounding web context and optical character recognition. They found that on top-ranked sites by traffic, they automatically generated captions for around 50% of not-described images. Notably, in both studies, government websites tended to be more accessible than other groups. The increased accessibility of these sites could indicate that those organizations are more aware of the accessibility needs of their citizens, or some jurisdictions (such as the US) require it by law [75]. A more recent (2017) survey of top websites by Guinness *et al.* found 20-35% of images lacking alt text in various categories [43]. It is unfortunate that such a significant portion of image content on the web remains inaccessible, but as I demonstrate in Chapter 3, there is far more alternative text on the general web than there is on social media platforms such as Twitter.

The technology for alt text descriptions has been standard since 1995, but recent research by Morris *et al.* contends that this standard may be stale, and modern computing platforms could support richer representation of visual content, including audio [71]. The audio content could be played instead of text-to-speech content that a screen reader normally provides, or the audio could be played as an ambient background track or sound effect. I explore the use of richer alternative text for the creation of accessible image memes in Chapter 6 and animated GIFs in Chapter 7.

2.2 Automatically generating alternative text

Currently, alternative text is primarily created by the developers of the website or authors of the website content. This text is manually written, and authors are recommended to follow the Web Content Accessibility Guidelines [20]. However, as many images on the web are not labelled correctly or at all [10, 45], researchers have sought to automatically generate image descriptions.

Optical character recognition attempts to extract text characters captured in images and correct errors to make coherent words or sentences [10]. Object recognition algorithms can locate and identify entities in the image that the model has been trained to recognize, such people or animals [112]. On Facebook, image descriptions list objects in an image, such as "Image may contain: 1 person, tree, text". Instead of a list of objects, scene description methods generate a caption for the image, attempting to describe aspects of the image in a grammatically-correct sentence structure, such as "A person standing in front of a tree". This approach is available in commercial applications like Microsoft Seeing AI [64]. MacLeod *et al.* explored the impact of these captions when viewed by people with vision impairments, finding that they are not sufficiently accurate [61]. When a caption failed to accurately describe an image, blind participants often were unable to recognize that the caption may be incorrect, instead rationalizing explanations to make sense of any incongruencies with the surrounding text. The authors also evaluated ways of expressing the uncertainty in the caption model to engender skepticism when viewed as alt text, finding that a negative framing of the caption provided a statistically significant effect. This framing could be added to the front of the caption, such as "I'm not very sure, but this image might be of a person standing in front of a tree".

In Chapter 4, I explore the use of scene descriptions, specifically those provided by Microsoft Cognitive Services that are used by Microsoft Seeing AI. I also use optical character recognition to perform text recognition for images that contain text content. Both of these methods fair well at providing middle-quality alternative text and work for many images, although the scene description results are often vague and inaccurate.

2.3 Crowdsourcing alternative text

Automated approaches are popular because they are fast and cheap, allowing platforms to deploy them at scale to make large swathes of the web accessible for screen readers. However, they are often less descriptive compared to human-written alternative text on websites that prioritize accessibility. Human-in-the-loop systems can generate accurate alt text of images by soliciting descriptions from sighted crowd workers [10] or friends online [12]. Salisbury *et al.* explored employing crowd workers to correct for errors in automatic captions [89]. The authors then allowed people with vision impairments to ask clarifying questions from crowd workers, but users were unable to recover from significantly inaccurate captions.

Human-in-the-loop methods are often framed as solely using workers on crowd platforms to label images on the fly, but alt-text can also reuse human-written text from around the web. ALT-Server was an architecture proposed in 1997 that stored image descriptions written by sighted people in a central database for future use [24]. When a user accessed an image without alt text, they could check ALT-Server to see if any description existed for that URL.

Instead of looking for alt text written for the image at a specific URL, the Caption Crawler project retrieves existing alt text by searching for the same image posted elsewhere [43]. As this method utilizes reverse image search to find the image on other websites, the image must appear elsewhere and be indexed by a search engine for this to be successful.

WebInSight [10] retrieves alt text utilizing both OCR and crowdsourcing, but also looks for images with links and retrieves alt text from the linked webpages' title and headings. Twitter

A11y (Chapter 4) utilizes the Caption Crawler method of reusing alt-text from other websites, a variation of WebInSight to collect alt text from image links, and stores alt text for later use similar to ALT-Server. However, these projects all focused on images on the web in general, and the images on social media may differ enough to thwart these methods.

2.4 Image accessibility on social media platforms

With the rise of social media platforms, a significant amount of image content on the web is now generated by end-users, not website authors. This has lead to a large amount of content being inaccessible, as users did not have the option to add descriptions to their posts. Morris *et al.* found that over 25% of English tweets in June 2015 contained an image, and Twitter did not allow alternative text to be added at the time [72]. The post text itself was not a substitute for alt text, as only 11.2% of tweets would serve as good descriptions for their accompanying images. In 2016, Twitter added an opt-in feature for users to write image descriptions for their images, which I examine in Chapter 3. Twitter later extended this alt text feature to include GIF animations in January 2020, which I investigate in Chapter 7.

Facebook has addressed the issue of alt text at social media scale by deploying object recognition software to efficiently create a large quantity of images descriptions [112], but they often do not contain enough detail to fully meet the needs of people with vision impairments.

Images posted on social media are unlikely to be shared elsewhere on the web, and if so, may not be indexed quick enough for Caption Crawler to find. Specialized types of images, such as screenshots or memes, are more common on social media [72]. I explore a method to make memes accessible in Chapter 6, but another approach is to automatically recognize facial expression in memes in an attempt to convey its emotional tone [81].

2.5 Research challenges

The introduction of alternative text features from major social media platforms shows they are beginning to think about accessibility for the deluge of user-generated content. But years after Twitter and Facebook first added some sort of alternative text in 2016, I demonstrate that very few users are writing alternative text. The automatic approach deployed by Facebook is better than having no alternative text, but it seems a partial solution for a widespread problem. With these features deployed, has the accessibility of the platforms improved for users with vision impairments? If not, what could be changed?

Twitter and Facebook developed these alternative text features a decade after their first launch, meaning there were 5-10 years where content was completely inaccessible to blind users. They built on a design of alternative text that has not changed since 1995. My concern is that today's burgeoning social networks (e.g., Snapchat, TikTok) contain content such as memes, augmented reality, greenscreen effects and other media types that may not be well served by traditional alternative text. Additionally, no social network has deployed audio descriptions for standard videos on social media, let alone consider what descriptions for short viral videos should contain. What should accessible alternatives for these media formats look like, and can we develop solutions

without another 10-year gap for people with vision impairments?

This work lays out both general philosophies and specific methods to stem the apparent rise of inaccessible images and other media on social networks. To that end, this thesis investigates the following research questions:

RQ1 How inaccessible is the visual content on social networks now for people with vision impairments? How does this differ from content experienced by sighted users?

RQ2 What approaches work well to make images on social networks accessible (such as providing alternative text)?

RQ3 What unique forms of images, such as viral memes or animations, exist on social networks that are difficult to make accessible? How might they be made accessible?

I explore and answer these questions through this thesis and draw on common themes that inform how researchers and platforms should approach the inaccessibility of user-generated content in the future.

Chapter 3

The State of Accessibility on Twitter

First, to understand the state of image accessibility on social media (RQ1), I examined Twitter's image description feature. Unlike newer platforms like Snapchat or TikTok, the Twitter platform does not solely feature visual content, meaning that there is already a contingent of blind users on the platform [72] as textual content is accessible to screen readers. While descriptions can be added on Facebook, those images already include default alt text. Twitter is therefore an ideal network to examine if users are aware of image descriptions and if they write high-quality image descriptions.

Work in this chapter was also published as a conference paper. The use of "we" in this chapter refers to all of the authors who contributed to that work. The full citation for that article is:

Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's Almost like They're Trying to Hide It": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 549–559. 9781450366748 http://dx.doi.org/10.1145/3308558.3313605

3.1 How accessible is Twitter now?

Twitter and other social media platforms have become increasingly visual [72] over the past decade as media such as photos, videos, and GIFs have become more prevalent as content. As visual media makes up a larger portion of total content on a social media platform, such as Twitter, such platforms risk becoming less accessible to people with vision impairments who use screen reading software to access the site [72].

An estimated 39 million people around the world are blind, and many access online sites through screen reader software. They are typical users of social networks [12, 110], but they are not always afforded access to content on social networks like Twitter [72], which are increasingly part of public discourse. Twitter is a platform for members of the media to disseminate and discuss information, as well as users to interact with celebrities in a different format than more traditional media [111]. Many research efforts have examined participation on Twitter in the context of politics around the globe [7], especially during elections [99]. It is important that all visual content on Twitter, including those relating to these topics, be accessible to people with

vision impairments so they may have equal participation in public life on the Internet.

Social media platforms have taken different approaches to make visual content more accessible to blind users, although all approaches provide textual descriptions (alternative text) to user-posted images. For instance, Facebook and Instagram both automatically tag each image uploaded to the site using image detection and recognition algorithms [50, 112]. Users can edit and override this text after the image has been posted. Twitter, on the other hand, allows users to add their own descriptions when the image is posted, provided the user has previously enabled that feature.

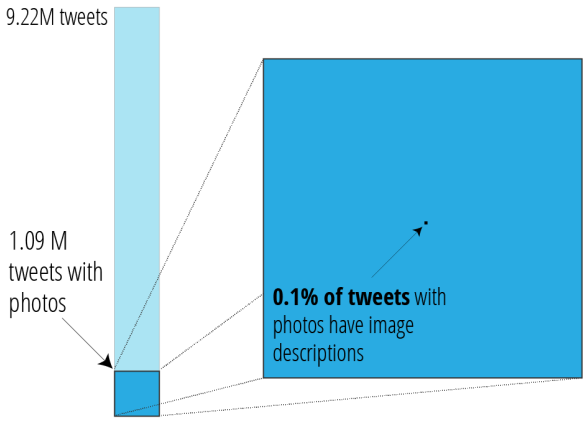


Figure 3.1: We sampled 9.22 million tweets, collecting 1.09 million with images. Only 0.1% of these tweets contained alternative text for people with vision impairments created using the opt-in image description feature on Twitter.

The automatic alternative text provided by Facebook is always available, but is not yet trustworthy to blind users compared to high-quality alternative text written by humans [112]. However, few users have enabled Twitter's image description feature for their account, and those that have do not always remember to write alternative text. We were interested in further understanding why Twitter users chose to enable and use this feature to better understand what motivates them to provide image descriptions. This understanding will help us improve similar features on social networking sites and increase the number of users providing high-quality alternative text for their social media content.

To understand the current state of alternative text provided on Twitter, we collected a sample of 1.09 million photo tweets and found that only 0.1% contained alternative text. By looking at a sample of posts with alternative text written in English, we found that 83.4% of human-written descriptions were of high quality. We then interviewed 20 Twitter users who had written image descriptions to understand their motivations for writing them.

Our findings suggest that very few users enabled the ability to post alternative text, indicating that Twitter could increase accessibility by turning the feature on by default. Those who do use the feature often do so infrequently, but generally provide alternative text of high-quality (excluding automatic posts from bots). Users who do use the feature could still benefit from training or tools that would help them write better image descriptions. We suggest that researchers or Twitter community members who wish to improve accessibility for the site develop these tools and measure their impact on the accessibility of social media content.

3.2 Quantifying image description prevalence

We sought to quantify the usage of the image description feature across Twitter to understand how the introduction of image descriptions have made Twitter more accessible. We describe the image

description feature, in general, followed by overall creation of alternative text by individual, popular, and public accounts.

3.2.1 Adding image descriptions

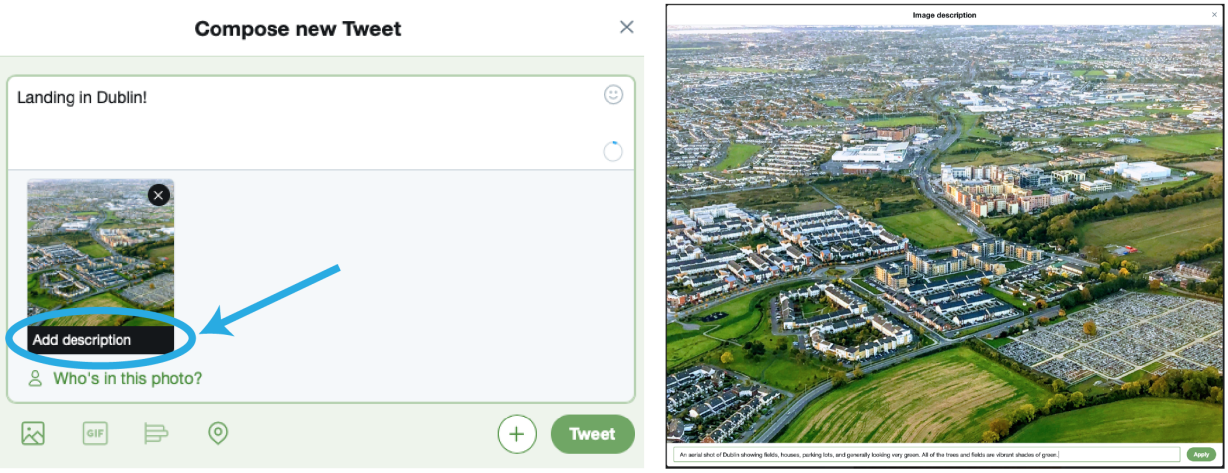


Figure 3.2: The user interface for adding image descriptions on Twitter. The left image shows the "Compose Tweet" window with the "Add Description" pop up circled in blue. The right side is the subsequent window showing the added image with a field at the bottom to add a description.

In 2016, Twitter added a feature to allow users to add image descriptions in their tweets that contain images [95]. By default this feature is disabled. The feature must be enabled by going to *Settings and privacy, Accessibility* and clicking *Compose image descriptions* [101]. A note below the checkbox says *"Adds the ability to describe images for the visually impaired."* and a link provides more information in Twitter's documentation. Once the feature is enabled, a visual cue appears when a user uploads an image on the website or mobile application (Figure 3.2). The tweet author may add a description of up to 420 characters to each image (a tweet may contain up to four). Descriptions may not be added to videos or animated GIFs, the other common media formats on the site, and descriptions may not be added or edited after the tweet is posted. Image descriptions can be added to tweets posted through the API using the "ext_alt_text" tag [102], so third party tools may enable image descriptions for their users as well.

3.2.2 Snapshot of image description usage

We first aimed to measure the amount of image descriptions on Twitter. Using Twitter's public API [103] we collected a sample of public tweets across 5 days in June 2018 for an average duration of 12 hours per day. This resulted in a collection of over 9 million tweets from all languages, including both original tweets and retweets (which may be duplicated). Approximately 1.09 million, or 11.84%, of those were tweets with at least one image. Only 1,144, or 0.1%, of those tweets (with images) contained image descriptions for at least one image.

Original photo Tweets

Upon further examination of the tweets with images, we noted that 271,330, or about 25%, were original tweets. The other 75% were retweets. Only 177, or 0.07%, of original photo tweets contained image descriptions. When we examined the image descriptions in our sample we noticed that some images had URLs as alternative text (often the source of the image), which we filtered out as they are not descriptive. After removing URLs, there were 166 tweets remaining; 0.06% of the original 271,330 photo tweets.

Retweets

If a tweet contains an image description, retweets of it will still contain the description, but new descriptions cannot be added by other users. Retweeted tweets accounted for 75% of the image tweets in our sample, although many of these were duplicated. We had a total of 820,469 retweets, but only 426,084 were unique (51.2%). For unique retweets, 0.11% contained descriptions, which after removing URLs left only 0.05% or 207 tweets.

Photos vs Tweets

Twitter allows a user to attach up to four images to a single tweet, each with its own description. When examining each photo from original tweets we observed that 336,584 photos were shared, and similar to the other categories only 0.05% contained image descriptions (only 0.03% after filtering out URLs).

The details of our sample and breakdown of our analysis is shown in Table 3.1. Overall, this suggests that less that 0.05% of the image content on Twitter is accessible to screen reader users.

3.2.3 Accessibility of popular accounts

We wondered if, despite the fact that tweets from randomly-sampled users were mostly inaccessible, perhaps more popular accounts would enable and use the feature regularly. These accounts, run by celebrities or organizations, were more likely to have professionals writing their content. Additionally, if popular accounts were accessible, they would have a greater impact on the overall accessibility of Twitter, as they would appear in more user's timelines.

We collected the image tweets from the top 50 most popular Twitter accounts [109] by number of followers. The tweets collected included every image tweet available between the time the feature was launched and October 2018. We found that only 3 of the 50 accounts had ever used the feature: the official Twitter account, Bill Gates, and BBC Breaking News. Assuming they enabled the feature just before their first tweet with an image description, these three accounts collectively added descriptions to 14 (8.8%) of 159 image tweets. This assumption does not make much sense for the Twitter account, of course, as it introduced the feature [95]. Only 5 of Twitter's 44 image tweets since the introduction of the feature have included image descriptions. The other 47 popular accounts that never added descriptions for accessibility to their images included prominent news organizations (New York Times, CNN, ESPN), politicians (Barack Obama, Donald Trump, Narendra Modi), and celebrities (Katy Perry, Justin Bieber, Taylor Swift).

Table 3.1: Number of tweets and retweets containing photos and alt-text.

	Day 1	Day 2	Day 3	Day 4	Day 5	Total
Tweets and Photo Tweets						
Total Tweets	1,860,947	1,757,012	1,366,244	2,243,235	1,997,713	9,225,151
Photo Tweets	225,679	189,445	145,778	269,098	261,789	1,091,799
Photo Tweets with alt text	195	199	205	297	248	1,144
Original Photo Tweets						
Original Tweets with photos	57,530	46,793	37,949	65,546	63,512	271,330
Original Photo Tweets with Alt	40	34	24	53	26	177
Photo Retweets						
Photo Retweets	168,149	142,652	107,839	203,552	198,277	820,469
Photo Retweets with alt	155	165	181	244	222	967
Original Photos and Retweeted Photos						
Total Original Photos	71,065	56,736	46,626	81,262	80,895	336,584
Total Retweeted Photos	287,909	241,828	183,244	355,346	349,894	1,418,221

Based on prior work that found government websites more accessible [10, 45], we wanted to see if this trend carried across to government accounts on Twitter. Using a list (created by C-SPAN) of 577 accounts associated with members of the current U.S. Congress [16], we performed the same analysis as above. In all tweets since the introduction of the feature, 42 accounts had used the feature at least once, with an average of 3.8% of their image tweets containing a description.

3.3 Accessibility for blind users

Understanding the state of accessible images on Twitter as a whole is valuable to inform how easy it is for a screen reader user to interact with any piece of content on the platform. However, this is not how most users experience Twitter content. They follow a set of users and only see the content authored, retweeted, or liked by the users they follow.

We analyzed the timelines of 94 self-reported blind users to determine if the experience of using Twitter is more accessible for people with vision impairments. To find blind users (who likely use screen readers to access Twitter), we collected a random sample of the accounts that follow the Twitter account of the National Federation of the Blind (@NFB_Voice), a large US-based organization led by blind people. From this sample, we selected an initial 100 users who self-identified as blind or visually impaired in their Twitter profile description. There is no

definitive indication that these users also use a screen reader, but we assume that many do.

We wanted to understand if the timelines of these users had the same level of accessibility as Twitter as a whole, or if the accounts they follow posted accessible content more frequently. For each user, we collected all of the accounts they follow, known as "friends" on Twitter[1]. For 6 users, we were unable to retrieve this data, as the accounts they follow were private. For each friend of the 94 remaining accounts, we collected 200 of their most recent tweets (and retweets), or as many as were available. Using these tweets sorted chronologically, we recreated each user's Home timeline for one day in October 2018.

Table 3.2: Summary of the tweets in one day of recreated Home timeline for 94 blind Twitter users.

	Min	Max	Average	Median
Total friends	6	5,001	720.7	371
Total tweets in timeline	2	29,231	3,379.7	1,554.5
Percent photos in timeline	0.0%	39.7%	18.4%	18.3%
Percent alt Text in photos	0.0%	41.4%	4.6%	2.0%

This is not a perfect recreation of a user's Home timeline, as Twitter does not show every tweet chronologically [100]. Additionally, we could not gather tweets from accounts with higher privacy settings (protected accounts), and some posted tweets may have been deleted by the time we collected them. However, we believe this to be a good approximation of the content these Twitter users would have been exposed to if they logged into Twitter that day.

Overall, we found that the recreated timelines for these users included 18.4% of tweets with photos on average, and 4.6% of photo tweets contained image descriptions. Table 3.2 contains information about the range of content we observed in these users' timelines, and a visual depiction is shown in Figure 3.3. In general, these timelines were an order of magnitude more accessible than Twitter as a whole, indicating that these users may be involved in communities with more awareness of the image description feature. An alternative explanation is that these users chose not follow some accounts that post inaccessible images. Regardless of the explanation, while these timelines are more accessible than our random sample of Twitter as a whole, they were still largely inaccessible.

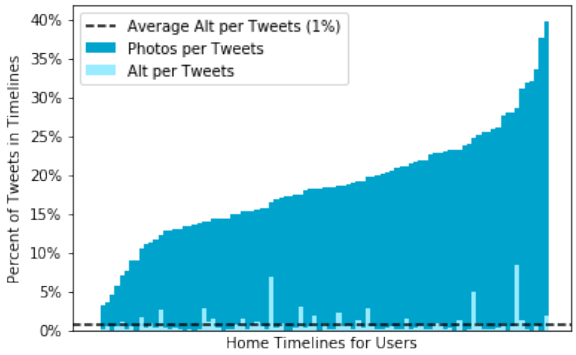


Figure 3.3: Each timeline analyzed, with the y-axis being percentage of total tweets. Dark blue: photos as a percent of tweets; lighter bars: alt text as a percent of tweets.

[1]This does not indicate a mutual relationship, as it does on many other social networks.

3.4 Quality of image descriptions

After investigating the prevalence of image descriptions on Twitter and finding low usage of the feature, we were curious about the quality of descriptions that do exist. Do authors write image descriptions in a way that is useful for people with vision impairments? Are the descriptions relevant to the photos?

To answer these questions, we developed a four-point rating scale from "Irrelevant" to "Great description" to assess the quality of image descriptions on Twitter posts (Figure 3.4). We filtered our sample of original photo tweets with descriptions to those in the English language (based on the "lang" attribute in the Tweet metadata), as we could not effectively assess non-English descriptions. This left 93 tweets from 71 users. To get a larger sample of tweets, we downloaded all tweets from these users with alt text.

3.4.1 Evaluating image descriptions

Prior work by Salisbury *et al.* [89] constructed conversations between crowd workers who were not allowed to view the image in a tweet, and crowd workers who were allowed to and expected to describe the image. From this, they built a set of structured questions to help guide composition of alternative text. We merged these questions and guidance from the Web Content Accessibility Guidelines [20] to develop a rating scale for image descriptions on Twitter. We used a rubric to rate the quality of image descriptions, and examples are shown in Table 3.4.

3.4.2 Findings

Two of the researchers used the rubric to redundantly code the quality of 500 photo tweets. We estimated the inter-rater reliability for this set by calculating a weighted Cohen's Kappa of 0.83 for these 500 ratings, which can be seen as strong agreement [63]. One of the researchers then rated an additional 500 photo tweets, resulting in a total of 1,000 rated tweets.

Using this set of 1000 rated tweets, we found that 62.6% of alt text rated as irrelevant or somewhat relevant. Only 15.8% rated as "great", with most elements of the image described. More details are available in Table 3.3.

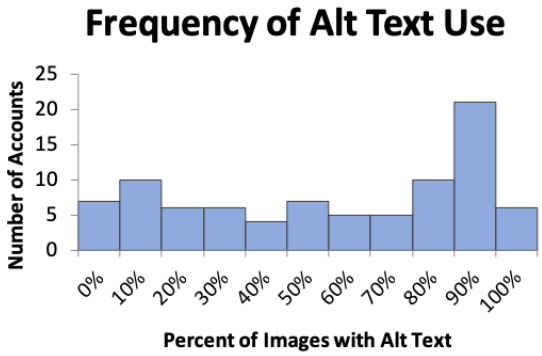


Figure 3.4: Histogram showing the frequency of image descriptions posted for the accounts in our sample. There is a spike in the 90-100% range that is comprised mostly of bots.

3.4.3 Frequency of use

Some accounts included in our sample used alternative text often, while others only had included a description in a small fraction of the im-

Table 3.3: Percentage of image descriptions rated using our rubric for different samples.

Rating	Sample 1	Sample 2 (no bots)	Combined
Irrelevant	32.6%	1.1%	16.9%
Somewhat relevant	30.0%	15.9%	23.0%
Good	21.6%	34.1%	27.9%
Great	15.8%	48.9%	32.4%
Sample Size	1000	1000	2000

ages they tweeted. We looked at all of the images each account had posted since they first used the image description feature. On average, accounts included image descriptions for 60.5% (median = 66.67%) of images tweeted. Many of the accounts who included image descriptions 90-100% of the time were automatically posting content as bots, however. After manually removing 19 of these accounts (Section 3.4.4, we found that humans tended to write image descriptions for 49.8% of their photos (median = 50%).

3.4.4 Quality of bots vs. people

From this analysis, we noticed that many of the tweets with image descriptions were coming from the same accounts, and these accounts were either explicitly bots or exhibited bot-like behavior. The accounts we knew were explicitly bots indicated they were automated in their name or profile description. Some bots generated specific kinds of memes, such as taking images from museum archives and superimposing fictional characters on them. Others posted automated information about location-based information such as earthquakes or air-quality, and included maps with their posts. Other accounts did not explicitly state they were bots but exhibited bot-like behavior such as blogs that re-posted articles from their website to drive traffic there from Twitter.

These bots exhibited different levels of accessibility. Some included information that described the entirety of a generated image. Others, such as blog-associated accounts, included human-written alt text for the images in their articles. Overall, however, we saw a general trend of descriptions associated with bot posts receiving a score of ”Irrelevant” or ”Somewhat relevant”.

As we were interested in human-written alternative text, in addition to overall accessibility on Twitter, we sampled an additional random 1000 tweets from our users that excluded the 19 accounts that were explicit or suspected bots. One member of the research team rated this additional sample, finding that just under half of the sample rated as ”Great” (48.9%) and 34.1% as ”Good”. The complete numbers of these samples are available in Table 3.3.

The difference between these two samples is stark, and indicates that the majority of poor image descriptions may be generated from automated sources. The human tweet authors in our sample described images at a moderately useful level (“good”) or higher. Bots are currently important and lively sources of content in the Twitter ecosystem, and their content should be just as accessible to people with vision impairments. These results for human authors are promising,

Table 3.4: Alt-Text Rating Examples. *Links and Usernames are removed from post text for anonymity.

Rating	Image	Post Text & Image Description	Rating and Reason
Irrelevant to image		**Post:** New Music Video - Beach House (@[username]) "Black Car": [link][link] **Alt:** Beach House	We rated the alt-text for this image as **irrelevant** because it does not describe anything in the actual image shown.
Somewhat relevant to image		**Post:** Trailering training class in Henderson yesterday. **Alt:** Trailering training class in Henderson yesterday.	We rated the alt-text for this image as **somewhat relevant** because it relates to the purpose of the image but just repeats the post text without describing more of the image.
Good: some aspects of image described		**Post:** Collecting all the shirts #DNNSummit **Alt:** Three t-shirts from DNN summit.	We rated the alt-text for this image as **good** because it describes most of the things in the image. It could also describe the color and/or what is on each shirt.
Great: almost everything is described		**Post:** "Taking a moment to appreciate beauty. #DogsOfTwitter @[username] #beauty #spring" **Alt:** Styx the dog sits in front of a tree with plastic Easter eggs dangling from the branches. It's very pretty.	We rated the alt-text for this image as **great** because it describes almost everything in the image. This particular description also points out information that might not be immediately apparent visually, either.

and show that many users write good image descriptions, if they enable and use the feature.

3.5 Description author interviews

We interviewed people who had used the image descriptions feature to add alternative text to images in their tweets. Prior work has demonstrated a clear accessibility need on social media platforms from the perspective of people with disabilities [12, 72, 110]. Twitter users that create content must be able to make their images accessible, as the responsibility of making Twitter and other social media platforms accessible should not fall solely on people with disabilities. We were interested in users' motivations for using the feature, their process for composing image descriptions for tweets, and why they may not always add descriptions to their tweets.

3.5.1 Participants

We interviewed 20 Twitter users who had written at least one tweet including an image and description. Participants were recruited via direct messages or emails to users identified in our sample from Section 3.4: Quality Ratings. We also recruited participants via an advertisement on Twitter posted by one of the authors. The mean age for participants was 41.6 years (std. dev.= 11.7) with a range of time using Twitter from 4 to 12 years (very early adopters). None of the participants reported any visual impairments or use of screen reader technologies. Thus, participants in our study do not represent visually impaired users on Twitter, but rather represent the creators of content that visually impaired users might consume.

3.5.2 Interview format

We conducted interviews in person, over the phone, and through Twitter direct messages. The interview consisted of four topics: activation and use of the image description feature, interactions with blind or screen-reader users regarding alt text, process of writing and examples of their alt text, and suggested changes for the image description feature on Twitter. Demographic information is listed in Table 3.5. The interview sections are described below.

Feature use

We asked participants three questions regarding their use of the image descriptions feature:

1. why they activated the feature,
2. how they decide when to add descriptions, and
3. and how often they believe they add descriptions to the images they tweet.

Our aim was to understand what motivated these users to activate the feature, how actively they provide image descriptions, and for what purpose.

Table 3.5: Participant Demographics.

PID	Age	Gender	Years On Twitter	Occupation
P1	32	male	10	Make and sell jewelry
P2	32	male	10	Client services
P3	45	male	10	Network Administrator
P4	38	woman	9	Translator
P5	48	female	10	Public Engagement
P6	61	female	10	Dir. of IT Accessibility
P7	22	male	7	Research Fellow
P8	25	male	7	Ph.D student
P9	28	female	4	Research Associate
P10	42	female	9	Photographer
P11	44	female	9	Student
P12	32	male	4	Software Engineer
P13	57	female	9	Retired Hydro-geologist
P14	48	male	6	Construction Manager
P15	45-50	trans	9	Scientist
P16	43	male	12	Software Engineering
P17	55	female	12	Marketing
P18	59	male	8	Engineer
P19	48	male	9	Musician, Developer
P20	32	female	9	Marketing

Interactions with followers

We asked participants about their interactions with followers who were blind or used screen readers. Specifically, we asked about any direct interactions participants had and how many of their followers they believed benefited from image descriptions.

Examples of image descriptions

We discussed participants' process for composing image descriptions for their tweets. We first asked participants to describe the elements they included in their descriptions and what they think about when composing them. Next, we chose specific examples from participants' tweets to discuss; one where they wrote alt text and one where they did not. We discussed how they wrote the specific image description for the example and reasons they may not have added descriptions for other images.

Changes to image description feature

Finally, we asked participants to think of one thing they would change about the Twitter image descriptions feature.

Data and analysis

Each interview was transcribed and analyzed using a theoretical approach to thematic analysis [14] based on the sections of the interview described above. We coded each transcripts based on the topics of the interview. The first and second authors redundantly coded the first five interviews, then discussed and refined the code book before independently coding the remaining 15 transcripts.

3.5.3 Findings

Feature discovery

Participants discovered the image description feature through a variety of ways. The most mentioned (6) was by the suggestion of someone they were following or a tweet they saw mentioning the capability to add descriptions to images on Twitter.

> *"Because [a specific user I follow] suggested it, I suspect. I can't say for certain. I would never have looked into the accessibility settings if someone hadn't said that this feature was hiding there." – P3*

Some participants (2) discovered the image descriptions feature through announcements, presumably made by Twitter, when the feature was released. Three participants anticipated the release of the image description feature before it was announced. P1 had even requested that Twitter add the capability, almost two years in advance of its release.

> *"I'd been waiting for it to be an option for a while - me and friends had contacted Twitter about it but nothing happened for a while. So as soon as it became available I switched it on...Quite a few of us have disabilities so we try to push for accessibility features like captioning/transcripts or image description." – P1*

Three other participants mentioned they activated the feature because someone else had used alternative text to describe images or they were involved in accessibility related work or communities.

> *"I learned about it through [work] colleagues and wanted folks using screen readers to be able to access media, too." –P15*

Motivation

In terms of why they chose to use the feature, some participants cited that adding alternative text to images was a low effort activity, but would have a high impact on the people it was intended for.

> *"I knew I wanted to use it because it's such a simple small thing that is no effort for me, but may mean a lot to the reader." – P4*

Four participants cited personal or professional connections to someone with a disability as motivation for their use of alternative text. P18 described their professional connections:

> *"I ran [several] - all accessibility-related accounts. It would have been disingenuous to have the disability community as our majority base of followers and not include alt-text, so*

we did as soon as it was available. Same for my personal account since I am in that field and was perhaps more aware of it than others might be." – P18

Two ther participants, like P4 and P7, were influenced more personally by their relationship with blind users:

"Maybe subconsciously because a close friend of mine is visually impaired and I used to describe some pictures I find on twitter to him over lunch table conversations but now since we don't meet a lot, this is one way I can keep my friend engaged with my conversations." – P7

Overall, participants demonstrated that, while perhaps not personally connected, providing image descriptions was a matter of "inclusion" that makes things better for everyone.

Habits of use

With regard to participants' intended use of the feature it became clear that most (12/20) participants explicitly intended to add image descriptions to every image they posted. When we asked how often participants actually used alt text, we found that people experienced varying levels of success:

"Ok I'm not as good at doing alt-text [as] I thought. In my last 10 images, 5 had alt text." – P8

For others, adding the descriptions was a matter of convenience, citing that the applications they used either allowed it or made it straightforward to add the descriptions. Two participants described that when they had time, they would add image descriptions:

"Maybe its really the time which is the factor, when I feel like I am in a rush, I kind of miss adding it to alt-text and instead just describe it in the tweet text."–P7

Two other participants described being reminded by the interface (see figure 3.2) to add a description.

"The add a description prompt under the photo [which appears once users have activated the ability to write alt text in their accessibility settings] reminds me. I think I do it for every photo I add."–P11

We had two participants, P2 and P19, who created bots that generated original tweets with images that included descriptions. Both bots generated meme-like images by imposing content from online image archives with other images. The alt-text in both cases was just the description of the image that was in the library or museum archive. For these bots, including alternative text as part of the tweet metadata was supported by the Twitter API, thus making the inclusion of a description for every image very convenient.

One participant highlighted that they try to add image descriptions whenever the image is their own photo, whether it is a photo they took or a graphic they created. One participant, P20, mentioned that she switched to using the Twitter web interface or mobile app specifically because another third party tool she was using did not support adding image descriptions. This was similar to P6's professional commitment:

"If I include an image, I add alt text. If I don't feel like bothering, I don't include a picture. Professional pride" – P6

Follower knowledge and interaction

Half of participants (10/20) had no knowledge of any of their followers being blind or using screen readers. A majority of participants (11/20) had never been contacted by a blind person regarding their use of alt text.

> *"I have put stuff in my alt text to try to solicit a response and never gotten one, so I am not sure any of my twitter followers use screen readers." –P15*

The other half of participants (10/20) were aware of at least one blind individual among their followers. A few (8) estimated specific numbers between 1 and 40, or "a small percentage". Participants had limited Interactions with their networks regarding alt-text; however, one participant mentioned having discussions where alternative text was mentioned.

Authoring image descriptions

We asked participants to share their process for writing image descriptions, both in general and in relation to a specific example of one of their own tweets. Some participants had specific strategies for describing an image. Depending on the intent of the post (mentioned explicitly by 6 participants), a majority of participants (11/20) described writing a general description of the image. One participant imagines trying to explain an image to a friend on the phone.

One participant mentioned describing the colors that appeared in the image. Others described determining the importance of objects, background, and other content in the image that the reader may not be able to ascertain from the main text of the tweet. Participants mentioned transcribing the text included in images or describing the objects, actions, and facial expressions in the image. Based on the content in the image, three participants tried to highlight the focus of the image.

For bot creators and specific content-focused accounts, participants mentioned that they used the image descriptions to convey the purpose of the tweet. For instance, to describe the important visual elements in the image to represent a joke or meme, or to convey why the image makes the joke funny in the context of the post (examples in Figure 3.5). We encountered two notable examples of content creators that had difficulty writing descriptions, a photographer and a bot creator with accounts that posts memes. P10 grappled with trying to describe the important photographic elements of her images:

> *"When it's a photograph being shown as a piece of art, that's where it gets difficult - especially since many of my photographs are quite abstract and tend to defy description! I try to touch on the straightforward visual facts of it (what is it a photograph of) but also get across the sense & feel of it where I can. The latter is in some ways more important with my photography. Things like colour tones (is it cool or warm, soft colours or vivid colours), are there any textures, what does it resemble." –P10*

P2, who creates Twitter bot accounts, described nuances of conveying jokes presented through memes:

> *"[The first bot] deviates from my personal answers since its alt-text is the punchline and doesn't describe the visual content of the meme. That can be hard. For [these tweets] the verbosity is the joke, so I used [the alt-text] to convey that verbosity. For [A different bot account] the joke is both that the character looks intelligent and is making a foolish*

(a)

(b)

Figure 3.5: The alternative text for (a) was: ”Cartoon man in glasses holding book looking at a butterfly labeled ’microbes’, asking ”Is this ponies” For (b) it was: ”Monochrome photograph of a number of dandelion seeds tangled together - much of the focus is soft, the the seed heads and fluff are clearly visible in places. There are hints of blue colour tone in the background.”

mistake...the alt text template there mentions the book, the glasses, and then names the two disparate concepts [in the image].” –P2

All of these examples illustrate a very nuanced process. The approaches vary from person to person and with the intent of the post and image.

Non-use

Participants cited time constraints as the most common reason they might have missed adding alternative text, either because it was too time consuming, they were in a hurry and forgot (8), or they had to write multiple tweets in succession and missed adding alt-text in the process (5).

Other reasons were that participants simply forgot what the feature was for or that the feature was there. We encountered four participants who mentioned that they relied on the Twitter interface to remind them to add the description.

“Actually, looking at twitter on web, the cue for alt-text is ”description” which sounds pretty optional at a glance. there’s a chance i read ”description” and forgot it was for the visually impaired in my rush to tweet and accumulate all the engagement. – P8

Some participants stated that adding image descriptions from a mobile device was still not possible. However, it is currently possible to add descriptions from the Twitter mobile application. This suggests that either participants did not notice the feature or have not used the application since the capability was added.

Changes

The most common suggestion, mentioned by eight participants, to improve the image description feature on Twitter was to make the descriptions visible to sighted users, especially for their own tweets. P4 states:

"[I want] to be able to add the alt text after the tweet is posted. I don't need to edit tweets as a whole, but I would really want to be able to go through all my pics and add alt text. And I wish hovering over images would show what alt text I wrote, like they used to with other images online. Maybe more people would notice/become aware then?"–P4

The only current process for viewing these descriptions requires the user to view the source code for the web page. This suggestion is also in line with the desired ability to edit the tweet content (and image description) after it has been posted, which is typically not supported on the platform.

The second most common suggestion (by 5 participants) was to make the image description feature "active" by default instead of as a setting that you have to turn on.

"I would make it automatically enabled for people so that users don't have to wade through the settings to turn it on. They could be helping so many more people if only they used this feature up front."–P20

Other suggested changes included improving the interface reminder, addressing bugs, and providing automated support to generate image descriptions. Participant 1 also mentioned the need to increase and "normalize" the use of the feature in a similar manner to captioning. In addition to support from Twitter, one participant (P11) wanted access to individuals with vision impairments to provide guidance on writing good image descriptions. Finally, three participants mentioned increasing the character limit (currently 420 characters) for the alternative text to allow more thorough descriptions, especially for screenshots or pictures of text.

3.6 Recommendations to increase description authorship

Our analysis of image description on Twitter revealed that very few people (including popular and government accounts) use the feature, with less than 0.1% of original image tweets having any descriptions at all. The image tweets exposed to blind users contain descriptions slightly more often (4.6%), but still are very inaccessible. The Twitter users that do use the feature author descriptions for about half of the images they tweet, and the descriptions they write tend to be "good" or "great" 83% of the time (bots excluded). Our analysis of interviews with image description writers examined the reasons for use (and non-use) of the feature, and lead us to two paths for improvements: those that Twitter (or similar social platforms) could currently undertake, and those that require further research and additional tools.

3.6.1 Improvements for the image description feature

Interview participants identified many issues with the image description feature that, if fixed, would lead to higher or better usage. Participants requested that alternative text be visible to them and editable after posting. Interview participants had trouble recalling which images they added descriptions to, and what they wrote. The most common reason they did not add a description was that they forgot when posting quickly, and being able to add image descriptions after the fact would be valuable. Taken further, if users could add image descriptions to retweeted images, volunteers or friends of screen reader users could then make this content accessible.

Another common request was for Twitter to just turn on the ability to author image descriptions for everyone, rather than requiring users to find and enable this capability.

"Top wish: it should be turned on by default. It's almost like they're trying to hide it." –P6

Most participants turned on the feature as soon as they found out about it, and try to include image descriptions for every tweet with images that they post. We agree that the image description feature is hard to find and understand, and Twitter should enable it for everyone to increase accessibility on its site.

However, just enabling the feature for everyone is likely not enough. It is currently designed for people who are familiar with image descriptions for people with vision impairments, and instantly enabling it for everyone could lead to misuse. It may be abused to make tweets appear higher in Twitter or external searches or include spam URLs. We have already observed users including subtle messaging in Twitter image descriptions for followers who know to look for it. Even without intentional abuse, new users who do not understand the purpose of the feature or how to write alt text may not produce high quality image descriptions. Twitter should provide clear on-boarding instructions for users when they first use the feature, explaining its purpose and why users should provide image descriptions for their images. In order to reduce confusion when posting a new image tweet, on image upload Twitter should provide instructions on how to write alt text or a template of structured questions, which prior work has found results in higher quality alt text in other media (*e.g.,* STEM textbook diagrams) [69].

3.6.2 Additional tooling and training for users

We see two major opportunities for researchers to make social media platforms more accessible through additional tools.

Some content on Twitter is ripe for auto-generation of image descriptions. Automatic captions for generic images has been exemplified by ALT text bot [23], which provides automated alt text in response to tweets containing images. However, researchers could go further in areas where the content format is more constrained. Some photos are just photos of tweets, and could be accessible if linked to the original tweet. Screenshots or photos of text are popular (9.7% of the no-bot sample in Section 3.4), and robust optical character recognition could make these accessible.

Researchers should also develop tools to help users write better image descriptions. Many users do not know what elements to include in an image, and would benefit from specific instructions, such as the structured questions developed in Salisbury *et al* [89]. Automated tooling could rate how descriptive alt-text is, provide specific instructions based on recognized objects in an image, or even pre-fill the image description with an auto-generated scene description [89, 116]. This may help prompt the user to change or refine the image description before publishing.

3.6.3 Supporting authorship through automated feedback

To provide a starting point for the goal of creating tools for image description authors, we used the sample rated in Section 3.4 to create an automatic rater for image descriptions. We merged ratings of "Irrelevant" and "Somewhat relevant" alt text as "low-quality" and ratings of "Good"

or "Great" as "high-quality". An Extra Trees classifier [32] was trained on features extracted from a subset of the sample (1,320 tweets). The specific features used were: counts the of parts of speech in the alt text and post text, shared words between the alt text and post text, as well as the length of the alt text. This classifier achieved an overall accuracy on the remaining 680 tweets of 85.3% (AUC = 0.84, precision = 0.83, recall = 0.94), demonstrating it is able to distinguish between much of the alt text quality. The five most important features for this classifier were: number of prepositions in alt text, number of words shared between post text and alt text, length of alt text, number of present verbs in alt text, and number of plural nouns in alt text.

At a very simple level, a classifier like this demonstrates that automatic feedback could be given to users, when they compose their descriptions, on whether it appears to be low or high quality. Specific feedback could focus on how similar the alt text is to the post text, objects in the image that are not mentioned, or a lack of actions (verbs) and objects (nouns) in the written description. I expand on this work in the development of HelpMeDescribe (Chapter 5).

3.6.4 Implications for other platforms

While we only examined Twitter in this work, we can infer that accessibility features that are similarly hard to find or use will also see low adoption on other large social networks. Social networks that do not give users ways to make their content accessible, or do not make it easy to enable will see little accessible user-generated content. This low adoption indicates that platforms may seek to employ automated methods to make images accessible, as we have seen Facebook launch object recognition algorithms on their platform. However, other automated methods exist, including text recognition or automated image captions, which produce full sentence descriptions of an image. In the next chapter I examine these methods, among others, to determine if platforms can and should employ them to make content on their networks accessible. This includes an analysis of how many images each method can be made described by a given method, and if so, how well it can be described. I examine these approaches again in the context of Twitter, but the exact methods that are useful on a social network may depend on the type of images present on that network.

The other important implication of this chapter is that celebrity and government accounts, who may provide some of the most seen or most important content on social networks, are also extremely inaccessible even though they have more resources to manage a social media presence compared to the average user. Due to network effects on social media that make this content more visible to users, platforms should consider targeting user education efforts on content that is seen by many people. Additionally, platforms might take an active role in ensuring that accounts produce accessible content they deem to be important, such as election information or emergency updates from a local governments.

3.6.5 Limitations

The primary limitation of this work is that we have not comprehensively studied the accessibility of Twitter as a whole, only photo tweets. Other forms of media, such as animated GIFs, videos, polls, and URL previews exist. URL previews in particular can contain alternative text that is

pulled from the linked page, but this was not the subject of our analysis. I address some of this in future chapters, such as linked pages (Chapter 4), memes (Chapter 6), and GIFs (Chapter 7).

Our rating scale was developed from prior research on the experience of blind users interacting with alt text, but it was still designed and executed by sighted researchers rating image tweets, Therefore, it does not fully reflect what screen reader users seek when browsing images on Twitter. I validate this scale in a survey with blind social media users in Chapter 5.

3.6.6 Conclusion

Our findings show that image descriptions are very rarely provided on Twitter. This is in large part due to very few users having the feature turned on, but even the users who enabled the ability to provide alternative text descriptions did not always write them. Twitter would likely dramatically improve its overall accessibility if it encouraged all users to provide descriptions and enabled them to do so across all types of media.

Access to social media is increasingly important for participation in many aspects of society, including social connection, entertainment, civic participation, and news consumption. By improving descriptions of visual content on social media networks, many users with vision impairments [110] will again have equal access to these vital platforms. As P15 states:

> *"Only built-in accessibility from the start provides more equitable access. Only doing it because/when someone asks for it or because we know a specific individual needs it, puts the burden on the person with the need, and that's not how accessibility should work. And in terms of accessibility online, alt text for static imagery is a low hanging fruit, easy and inexpensive to implement. Failure to do it is just inconsiderate laziness."–P15*

Part II

Automated Tools to Improve Social Media Accessibility

Chapter 4

Making Images Accessible on Twitter

It is clear that social networks like Twitter are not seeing widespread adoption of image description authoring. As that platform has had alternative text for the longest and likely more blind users compared to platforms solely devoted to visual content, few images on social media are likely to be accessible. While I recommend that social media platforms work to increase support and education for users to write their own image descriptions, it is not clear that will quickly solve the issue of inaccessible social media content. However, researchers have proposed alternatives to make the web accessible that include: text recognition, automatic image captioning, crowdsourcing, and more. In this chapter, I detail my work to build Twitter A11y, a tool that utilizes and evaluates multiple methods to make images accessible on Twitter.

Work in this chapter was also published as a conference paper. The use of "we" in this chapter refers to all of the authors who contributed to that work. The full citation for that article is:

Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M Kitani, and Jeffrey P Bigham. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, USA, 12. 978-1-4503-6708-0 `http://dx.doi.org/10.1145/3313831.3376728`

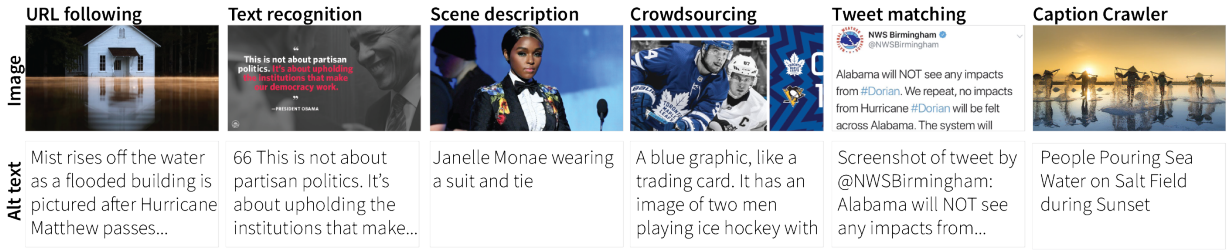


Figure 4.1: Twitter A11y describes images posted to Twitter depending on the image type including user-posted images, external link previews, and text-based screenshots. For each method, we include a sample tweet image (top) and sample alt text produced by the method (bottom).

4.1 Automatically generating and reusing image descriptions

Social media platforms provide a medium for online discussion and information dissemination, but accessibility barriers on these sites can prevent users from accessing them with a screen reader. People who are blind or low-vision use screen reader software to read the text on a webpage or application aloud, but social networks lack the necessary descriptions for visual content, like images or videos. For example, Twitter was originally a very popular social network for people with vision impairments, as the text-based posts were accessible via screen readers [89]. However, the steady increase in the number of images posted by users has lead to these platforms becoming less accessible because they do not include image descriptions (alternative text). As demonstrated in the last chapter, around 12% of content on a random sample of Twitter consists of images, and only 0.1% of the images on Twitter include descriptions.

Access to social media platforms is critical for people with vision impairments to both communicate with friends and colleagues, and to participate in public discourse. People with vision impairments interact with social media features to the same extent as sighted users, but prior work notes a decrease in interaction with visual content and features on Facebook [110]. In addition, people with disabilities often use social media to share information about their disabilities or organize around disability activism. For example, the #HandsOffMyADA and #CripTheVote campaigns on Twitter were organized by disability activists around pending legislation in the US [8, 29]. Auxier et al. found that even in the #HandsOffMyADA campaign, only 7% of images contained alternative text, leaving most of the images inaccessible to people with vision impairments.

To provide high-quality descriptions for images on social media platforms, we designed an end-to-end system, Twitter A11y, to generate or retrieve alt text for images on Twitter (Figure 4.1). Prior research has developed several automatic and human-in-the-loop methods to generate image descriptions, and these methods are now robust enough to deploy at scale to address the growing lack of image accessibility on social media. Twitter A11y includes three methods that automatically add alt text for user-posted images: text recognition (optical character recognition), scene description, and the Caption Crawler method [43] (reverse image search). Two additional methods seek to address Twitter-specific images categories: screenshots of tweets and preview images for external links. Finally, if none of the prior methods produce a satisfactory alt text description for the image, Twitter A11y asks a crowd worker to describe the image on Amazon Mechanical Turk using a set of provided guidelines. Twitter A11y's browser extension dynamically requests alt text in the background for images as a user uses the Twitter website, and adds it to the image as if it had been there originally.

To evaluate the coverage and quality of alt text from the six methods, we performed a static analysis of images from 50 blind Twitter users' timelines. We randomly sampled tweets they may have read over the course of a day, creating a sample of 1,198 images. Through a combination of automatic methods, Twitter A11y increased the alt text coverage from 7.6% to 78.5%. We then rated a subset of these images to compare the quality of the descriptions returned from each method on a four-point scale. We consider the alt text that achieves either the highest rating ("Great") or second-highest rating ("Good") to be high-quality alt text. The highest percentage of quality alt text was from the text recognition (32.7% "Good", 44.9% "Great") and scene description (53.1% "Good", 14.3% "Great") methods. We also evaluated crowdsourcing

as an additional method, finding that 62.5% of the resulting descriptions were rated "Great" (and 18.8% "Good").

We recruited 10 participants who access Twitter via a screen reader, to evaluate the perception of Twitter A11y and the six methods. Twitter A11y was able to add automatic descriptions to 82.4% of the content they accessed, crowdsourcing the remaining 17.6%. On average, participants' perceptions were that 12.1% of images were accessible in their timelines before the study, and 72.3% of images were accessible when using Twitter A11y.

In this work, we make social media content accessible by asking people with vision impairments to install a browser extension, but the technological and financial costs of making social media accessible should not be borne solely by people with disabilities in the long term. Rather, the platforms should bear the responsibility to ensure their hosted content is accessible through more accessibility features, user education, and employing the methods used by Twitter A11y. Therefore, the Twitter A11y approach and user evaluation results should be informative for application developers, not used as a justification for user-installed solutions.

This work represents a combination and comparison of known methods that are now robust enough to address the accessibility issues plaguing social media platforms. We show the potential for dramatic improvement in accessibility, the differences between coverage and quality of different methods, and the impact of this tool on the social media experiences of our participants. This work opens future directions for researchers to improve and combine the methods used by Twitter A11y and provides guidance for social media designers on integrating methods to make their platforms accessible at scale.

4.2 Twitter A11y: a system to make images accessible

Twitter A11y combines a browser extension (Figure 4.2, square corners) with a backend server (rounded corners) to make image tweets accessible through one of six methods.

4.2.1 Requesting alternative text

When the user loads the Twitter.com web interface, the browser extension observes new images loaded on the page. When an image associated with a tweet is loaded on the user's timeline, the extension extracts the image URL, any existing alt text, and context of the tweet (*e.g.*, tweet ID, user ID, tweet text). If the tweet contains a preview image and link to an external website, the extension also records the linked URL. The extension automatically requests alt text from the server using this data, without user input.

4.2.2 Obtaining alternative text

The Twitter A11y server receives requests for alt text that include the image URL and tweet context and attempts to use up to six methods as applicable to fetch or generate the image descriptions. The methods are ordered to prioritize a quick response, with methods that take less time returning early if they result in alt text for the image. The costs and total time taken to return a response if a method is successful is present in Table 4.1. If an image has already been

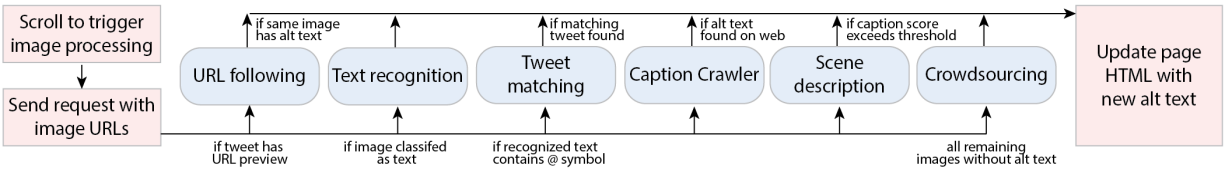


Figure 4.2: Flowchart of Twitter A11y process. Square rectangles depict steps that happen in the browser extension, and rectangles with rounded corners depict steps on the server. When the user scolls on the Twitter webpage, the extension will send all tweets with images to the server. The server progressively attempts up to six different methods to generate alt text for the image, returning a result early if one is successful.

requested by another user and alt text exists in the database, that alt text is returned to the user immediately. Because of this, requests for tweets from popular Twitter accounts will have a very quick response time.

4.2.3 Obtaining alternative text

The Twitter A11y server receives requests for alt text that include the image URL and tweet context and attempts to use up to six methods as applicable to fetch or generate the image descriptions. The methods are ordered to prioritize a quick response, with methods that take less time returning early if they result in alt text for the image. The monetary costs (*i.e.*, API request or crowd payment) and total time taken to return a response if a method is successful is present in Table 4.1. These are the costs incurred per Twitter A11y response in our study, but we expect social media platforms or third-party application developers could implement these methods cheaper and faster. If an image has already been requested by another user and alt text exists in the database, that alt text is returned to the user immediately. Because of this, requests for tweets from popular Twitter accounts will have a very quick response time. If a tweet image contains existing alternative text written by the poster, it will not be replaced by Twitter A11y, as the original alt text is likely the most suited for the image.

If no alt text is present for the image, then Twitter A11y tries the following methods in order: URL Following, Text Recognition, Tweet Matching, Scene Description, Caption Crawler, Crowdsourcing (Figure 4.2). When a method returns alt text and the alt text satisfies the threshold conditions, the progression breaks early. Otherwise, a crowd worker writes alt text to ensure all images receive a description.

External link previews

Websites such as news organizations share external articles on Twitter that contain described images on the external website but do not include alt text for the link previews that appear in such tweets (Figure 4.1, URL Following). If this is included in the tweet context data, the system crawls the page at the given external URL to extract all images. We compare the color histogram of the image to be described with the color histograms for all images in the external page (~3-10 images total for news articles). If a matching image is found and contains alt text, this is returned

Method	Cost	Agg. Cost	Agg. Time
URL Following	0.00¢	0.00¢	3.0s
Text Recognition	0.15¢	0.15¢	2.5s
Tweet Matching	0.00¢	0.15¢	2.3s
Scene Description	0.25¢	0.40¢	2.4s
Caption Crawler	0.40¢	0.80¢	14.8s
Crowdsourcing	36.00¢	36.80¢	126.7s

Table 4.1: The time and per-request costs incurred with Twitter A11y's methods. Because it tries methods in a specific order, the aggregate costs and time are presented to respond with a result from that method.

to the requester.

Text-based images

Many tweets feature images of text (*e.g.*, text that exceeds the Twitter character limit, text messages, screenshots of tweets). We determine if an image tweet depicts primarily text (Figure 4.1, Text Recognition) using whole-image labeling via the Google Cloud Vision API [39]. If the confidence of the "text" label exceeds 0.8, we record the resulting Optical Character Recognition result (also via Google Cloud Vision API) as the alt text. We selected this threshold empirically to achieve high-precision (lower recall) such that users are unlikely to receive inaccurate alt text.

Screenshots of tweets

Twitter provides a way for users to share other tweets by using the retweet feature, but users often share screenshots of tweets instead. This could be to preserve a tweet in case the author later deletes it or prevent harassment by sharing a tweet without notifying the original author. If the text recognition results from the previous step include a Twitter username beginning with the "@" symbol, we use the Twitter API to search for tweets by that user containing the first 10 words in the text recognition results. If a matching tweet is found, we describe that it is a screenshot of a tweet and return the tweet text directly, which can be more accurate than text recognition results.

Reverse image search

Some images shared on social media are copied from elsewhere on the web, where alt text might be present. We re-implemented the Caption Crawler [43] project to source alt text from other websites with the same image. This method utilizes the Bing Image Search API [66] to perform a reverse image search. Once we have a list of locations where the same image appears around the web, we crawl up to 25 webpages to find the image and any alt text it contains. If multiple webpages have alt text, we return the longest one, as evaluation of the Caption Crawler project found length to be a good heuristic for image caption quality.

Automatic image captioning

Other social media platforms, such as Facebook, have experimented with providing image captions automatically generated by object recognition or scene description algorithms [112]. These are useful because they can be easily scaled to many images, but can lead blind users astray if the generated caption does not accurately or fully describe the image [61]. Twitter A11y will attempt to use this method if the previous ones were not applicable or did not result in alt text. It uses the Microsoft Cognitive Services Vision API to generate an image caption, and chooses the resulting caption with the highest score. If no caption exceeds a threshold of 0.7, determined empirically, then it is ignored.

Crowdsource all remaining images

For remaining images not handled by the prior, rather quick methods, we post a crowd task on Amazon Mechanical Turk. This ensures that all requested images will receive some alt text from Twitter A11y. The task asks workers to generate image descriptions using the guidelines informed by Salisbury *et al.* [89]. The task time was originally estimated by the authors to take ~60 seconds and crowd workers were paid $0.17 per image ($10 per hour). After evaluation with some crowd workers, we found the median task to take 108 seconds (mean = 122s), so the task reward was increased to $0.30 per image. As the worker may take some time to write the description (~2-5 minutes), the browser extension displays "Waiting for crowd worker" until the written description is ready. Crowdsourcing carries the benefit that humans may be able to best describe characteristics such as humor that automatic methods miss.

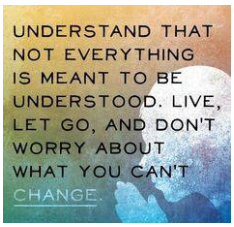
4.2.4 Displaying alternative text

From the server, the extension receives either the existing alt text in the database, newly generated alt text from one of the above methods, or the status of an uncompleted crowdsourced description. The browser extension then dynamically inserts it into the alt text tag for the image. The user's preferred screen reader can then read the newly generated alt text for an image when it focuses on the tweet, just as it would if the alt text had been there by default. While automatically-generated alt text appears soon after viewing (~2-10 seconds), crowd-generated alt text takes longer (on the scale of minutes) such that the user could view the text upon re-visiting the tweet (*e.g.*, by scrolling back in their timeline, or revisiting a user's page where the tweet appeared).

4.3 Static analysis of blind users' timelines

To gather a large number of users who may use a screen reader to access Twitter, we examined the Twitter accounts following the National Federation for the Blind (@NFB_Voice) and the American Federation for the Blind (@AFB1921). We selected 50 users from this list who self-described themselves as blind or visually impaired in the profile description. It's possible that these users do not use a screen reader to access Twitter, and therefore they may not notice the presence or absence of alternative text. However, we are making the assumption that a large

Table 4.2: A high-quality and low-quality alt text example for different images made accessible by each method utilized in Twitter A11y.

Strategy	High-Quality Alt Text Rating		Low-Quality Alt Text Rating	
	Image	Alt Text	Image	Alt Text
Original		Overhead map of the UK made up of people standing and the words During NEHW another 1,400 people across the UK will be diagnosed with advanced age-related macular degeneration		Palestinian protester
URL Following		A guide dog with a harness sits on the ground.		A point & click adventure game about the fun, alienation, stupidity and agony of being a teen.
Text Recognition		UNDERSTAND THAT NOT EVERYTHNG IS MEANT TO BE UNDERSTOOD. LIVE, LET GO, AND DON'T WORRY ABOUT WHAT YOU CAN'T CHANGE		Cittle oullorly Blesseng — (lowerserarch c Croaon
Caption Crawler		Google home device pictured next to packaging box for size perspective		How to Make Special Video Effects
Scene Description		a group of people posing for a photo		a teddy bear sitting on top of a grass covered field
Crowd workers		A row of large, white Cannon professional video camera lenses are sitting on a perforated surface like a vent panel or an appliance.		Two faces hidden in the beautiful painting

majority of these accounts that self-identify as blind or visually impaired do care about the presence of alt text on Tweets. The "Home Timeline" is the feed of tweets that a user reads when they log in to Twitter. We simulated a version of this timeline by collecting all of the tweets posted or retweeted by accounts each user followed over a 24 hour period. We then placed them in chronological order. All 50 users had at least 1,000 tweets in this simulated timeline.

These simulated timelines have some differences from those that would be experienced by users. First, Twitter does not always include all tweets in chronological order, choosing to place popular tweets first in the timeline. The Twitter algorithm may also hide replies to tweets that are not relevant, or include tweets liked by accounts the user follows. The actual timeline may also include ads. Finally, we were unable to collect tweets that had been deleted or were posted by protected (non-public) accounts, which the user may be able to see.

4.3.1 Accessibility of timelines

Accounts followed by blind users tend to include more alternative text in their tweets than Twitter as a whole. For the 50 accounts, we randomly sampled 50 tweets from each user that either contained images or links to external website. From this 2,500 tweet sample, after examining the links, 1,041 tweets remained with either an image or valid link preview for a total of 1,198 images (a tweet can contain up to 4 images).

Of these 1,198 images, 62 contained alternative text from the tweet poster, and 29 link previews had alternative text from the linked website, meaning 7.6% was already accessible. We then evaluated each method except crowdsourcing on all of the images, and calculated the ability of each to add alt text to the images (Table 4.3). The automatic methods increased the presence of alt text from 7.6% to 78.5% before applying crowdsourcing to the remaining 21.5%.

Alt Text Method	**Covered**		**Unique**		**Selected**	
	N	%	N	%	N	%
Original Alt	91	N/A	12	N/A	91	N/A
Scene Description	746	67.3	465	42.0	547	49.4
Text Recognition	216	19.5	60	5.4	199	18.0
Caption Crawler	213	19.2	70	6.3	70	6.3
URL Following	57	5.2	8	0.7	33	3.0
Tweet Matching	0	0.0	0	0.0	0	0.0
Crowdsourcing	1107	100	258	23.3	258	23.3

Table 4.3: Evaluation of alt text methods in a sample of 1,198 images from blind users' timelines. Covered indicates how many images in the sample the strategy could provide alt text for, and Selected indicates if that method was chosen according to Twitter A11y's method priority order. Because different methods can provide alt text for the same image, the Covered column does not sum to 100%.

When evaluating methods we considered three metrics in addition to quality:

- **Image Coverage**: How many images did this method produce alt text for in the sample?
- **Method Uniqueness**: For how many images in the sample was this the only method that produced alt text?
- **Selected**: Using the Twitter A11y method priority as defined in Figure 4.2, how often was this method's alt text chosen?

Using these metrics, we can change Twitter A11y's method priorities to optimize for different aspects of the user experience. We can see that in all metrics, Scene Description is providing a bulk of the alt text, meaning that if optimizing for speed then it should be first in the method priority. Twitter A11y only used 70 of the images provided by Caption Crawler (because it was last in the order before crowdsourcing), while it was able to source alt text for up to 216. As automatic descriptions may have lower quality than human-written descriptions from Caption Crawler, perhaps it should be the "last resort" automatic method instead of Caption Crawler. We examine the quality of captions returned by each method next.

4.3.2 Description quality

Two members of the research team independently rated a random subset of the data collected, consisting of 50 images and alt text for each method (except URL Following and tweet matching as they did not have enough examples). We include sample descriptions for each method in Table 4.2. As in Chapter 3, we utilized a four-point rating scale based on prior work [72, 89]. The scale ranged from "Irrelevant to image (0)" to "Great: almost everything described (3)" (available in Supplemental Material). To estimate agreement, we computed Cohen's Kappa $= 0.61$, indicating substantial agreement [63]. The two raters then met and discussed each instance where their ratings differed until they reached agreement.

Method (N)	Irrelevant	Relevant	Good	Great
Original Alt (48)	4.2%	20.8%	25.0%	50.0%
Scene Description (49)	8.2%	24.5%	53.1%	14.3%
Text Recognition (49)	6.1%	16.3%	32.7%	44.9%
Caption Crawler (38)	26.3%	42.1%	23.7%	7.9%
URL Following (26)	26.9%	46.2%	7.7%	19.2%
Crowdsourcing (48)	10.4%	8.3%	18.8%	62.5%

Table 4.4: Two members of the research team rated a subset of the entire sample to estimate the quality of captions returned by each method.

In Table 4.4 and Figure 4.3 these ratings are broken down by method. Crowdsourcing provided the highest "Great" quality descriptions as they were human-written with specific instructions for writing high-quality alt text. Text recognition has the highest "Great" percentage for an automatic method, comparable to alt text provided by the original poster and the only "Good" quality automatic captions from scene description. Surprisingly, Caption Crawler and URL following provided mostly "Irrelevant" or "Somewhat Relevant" alt text, even though they should

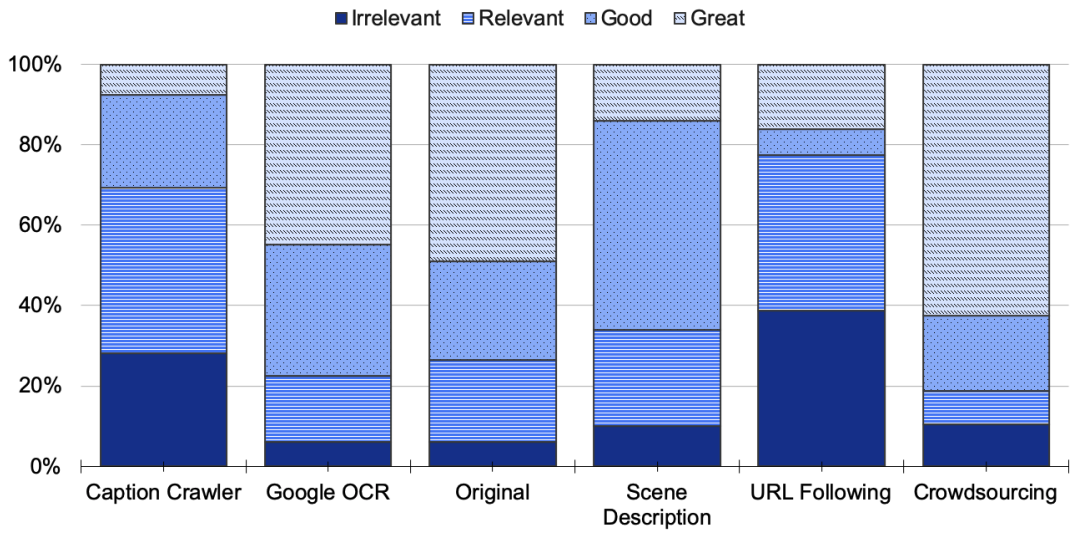


Figure 4.3: A breakdown of the distribution for level of quality for alt text descriptions generated by each strategy from Great at the top to Irrelevant at the bottom. The columns are normalized by the number of example images in our sample.

source human written descriptions. Upon examination of the low quality descriptions, we found that this was typically because website authors were using an image filename or article title as the alt text for the image, instead of properly describing the visuals in the image itself (Table 4.2).

This measure of quality by method allows us to measure the value for each method used by Twitter A11y that has external costs (*e.g.,* API fees or crowd payments). We define value as the number of "Good" or "Great" image descriptions a method can produce per $1 spent. This results in a value ranking of text recognition (571), scene description (269), Caption Crawler (79), and crowdsourcing (3). While money may not be a limiting factor for social media platforms, if third-party client developers wished to pass these external method costs along to blind users, crowdsourcing may not offer enough value to warrant using on every remaining image as in Twitter A11y. However, we do not advocate for passing these costs along to blind social media users as a long-term solution, and instead believe that social media platforms need to invest in accessibility solutions rather than relying on third-party application developers to create additional tooling.

4.4 Evaluation with blind Twitter users

The examination of Twitter A11y's performance on this static sample of images from blind users' timelines helped us get a sense of the ability of different methods to add alt text to images and at what quality level. Next, we wished to evaluate the usability of Twitter A11y and the perception of its ability to make Twitter content more accessible. We recruited 13 participants who use a screen reader to access Twitter to install and use the browser extension. The participants each completed a 30-minute semi-structured interview about their experiences with Twitter accessibility in the past. Then they installed and used the browser extension for one week. We followed

up after that week for another 30-minute interview about their experience using Twitter A11y. Three participants (P3, P9, P13) did not complete the data collection or final interview due to technical difficulties or lack of access to a computer. Their responses are included in the first interview, but not in the post interview results.

4.4.1 Participant demographics

We recruited people with vision impairments who had participated in previous studies, as well as sending out a call for recruitment on Twitter. Participant ages ranged from 19 to 54, with an average age of 33.5. Six participants were female and seven were male. All participants accessed Twitter using a screen reader, although many said they used additional applications to access Twitter either on their computers or on their smartphones. These are detailed in Table 4.5.

Participants were compensated $10 for each interview, and $6 per 10 minutes of using the system, up to a maximum amount of $62 for the entire study.

4.4.2 Pre-study interview

Before using Twitter A11y, we asked participants about their impressions of accessibility on Twitter, and what problems they saw with the social media platform and content on that platform. The specific questions can be found in the Appendix.

Regarding Twitter's accessibility as a whole, ten participants said they found the platform itself mostly accessible. A few participants (3) complained about using a screen reader on the Twitter website, but stated the mobile applications were satisfactory. However, most participants (10) still choose to use third-party clients (e.g., Twitterific, TWBlue) either because they supported screen readers more effectively or because they did not change layouts often. TWBlue was especially popular, as it is an open source application designed with screen reader accessibility in mind. Participants lauded its support of global keyboard commands, meaning they did not have to switch applications to use it. They also liked that it included a function to request the text in an image using OCR.

The most common accessibility issue mentioned by every participant was media accessibility. They noted that most images did not include alt text, even though they likely followed accounts that added alt text more often compared to the sighted population. Twelve participants said they could sometimes guess at the content of some images depending on the text content of the tweet, but they commonly said this worked for less than half of images.

To understand how participants perceived the scope of the lack of alt text on image, we asked them to estimate what percent of their feed contained images and what percent of those images were accessible. On average, they estimated 50.5% of their feed contained images (max = 70%, min = 10%) and believed 12.1% of the images they encountered contained alt text (max = 30%, min = 1%). Eight of the participants mentioned that the percentage of images that people post affect their decision to follow an account, with some stating they would not follow an inaccessible account, some stating they would unfollow someone who did not add alt text after they requested it, and one stating they blocked those accounts to keep inaccessible images out of their feed.

Participants also complained that there was no mechanism to make GIFs (short animations), which are common on Twitter, accessible by adding alt text. When asked about video acces-

ID	Age	Gender	Years on Twitter	Level of vision	Screen reader	Twitter Applications	Other Social Media
P1	19	M	6	Blind since birth	NVDA	TWBlue	Facebook
P2	39	F	11	Blind since birth	Voiceover	Twitterific	Facebook
P3	25	F	5	Blind since birth	NVDA, JAWS	TWBlue, Twitter (mobile site), Twitter for iOS	Facebook
P4	19	M	7	Blind since birth	NVDA, Voiceover, Talkback	TWBlue, Twitterific, Twitter for iOS, Twitter for Android	LinkedIn
P5	32	M	12	Blind since birth	NVDA, JAWS	TWBlue	None
P6	41	M	12	Blind since age 21	VoiceOver, JAWS, NVDA, Narrator	twitter.com, Twitterific, Tween	LinkedIn, Facebook, Instagram
P7	47	M	12	Blind since birth	JAWS, NVDA, VoiceOver	TWBlue, Twitterific, Twitter for iOS	LinkedIn, Facebook
P8	44	F	8	Blind since age 28	JAWS, NVDA, VoiceOver, Talkback	TWBlue, twitter.com, Twitter for iOS	Facebook, Instagram
P9	41	F	8	Blind since birth	VoiceOver	Twitter for iOS	None
P10	29	F	7	Blind since age 1	VoiceOver, JAWS	twitter.com, Twiter for iOS	Facebook
P11	22	F	8	Blind since birth	NVDA, VoiceOver	TWBlue, Twitter for iOS	Facebook
P12	23	M	2	Peripheral vision, no central vision since age 13	VoiceOver	Twitterific	Reddit, Youtube
P13	54	M	11	Totally blind since age 1.5	VoiceOver, NVDA	Twitterific	Facebook

Table 4.5: Demographics of participants who participated in the online study including age, gender, years on Twitter, level of vision, screen reader, methods of accessing Twitter, and other social networks used. Note that P3, P9, and P13 did not complete the data collection and post-study interview.

sibility, responses were mixed, and many said they found videos consisting of mostly dialogue already reasonably accessible. Three participants suggested that videos on Twitter should support the addition of audio descriptions as a secondary audio track or as timestamped text content.

Some users circumvented the issues with image accessibility by utilizing other applications. Participants stated they used Microsoft Seeing AI on their iPhones to receive a textual description of an image (similar to Twitter A11y's scene descriptions) or to read text in an image, if it was present. P12 noted that he could tell if an image contained text using his peripheral vision, but people who were totally blind would not be sure if an image contained text.

Participants who used other social networks stated that the lack of image accessibility was common to Facebook, LinkedIn, Reddit, and Instagram. Most participants who used Facebook mentioned their automatic image alt text, stating that they wished they were more descriptive, but it was better than no alt text at all. P10 stated that she liked Facebook's automatic captions simply because it stated "Image may contain: text", so she knew to send the image to an OCR application.

Finally, we asked participants what they would do to make Twitter more accessible. Eight participants wished Twitter would make the alt text feature more prominent, ensuring all users are aware of the feature and how to use it. Three even suggested the feature should be mandatory, while 4 others wanted Twitter to fill in empty alt text with automatic descriptions similar to Facebook. Users who utilized third party applications stated Twitter should better support them, especially by ensuring all features (especially muting, blocking, and other reporting tools) were available via the API.

4.4.3 Twitter A11y usage

After the interview, participants were directed to install Twitter A11y on their computers, and asked to use it for about 10 minutes a day over 7 days. Whenever they accessed Twitter, the browser extension logged every tweet containing images in their feed, and logged the alt text response added.

Session length and content

Participants used Twitter A11y for a total of 2,198 minutes over 145 sessions (mean = 15.2 minutes per session). During this time, they saw a total of 3,615 unique images (mean = 301.3). Of these, 86 already contained alternative text, and Twitter A11y provided descriptions for an additional 3,505 images. The breakdown of this by method and participant can be seen in Figure 4.4. The only time Twitter A11y did not provide any alternative text for an image was due to a technical error interrupting the request.

4.4.4 Post-study interview

After 7-9 days had concluded, we asked the participants about their experiences with Twitter A11y. Overall, almost every participant stated that they enjoyed using the system and found that many images were more accessible with the alt text provided. P7 was the exception, as he stated he did not see much additional alt text on the images he viewed.

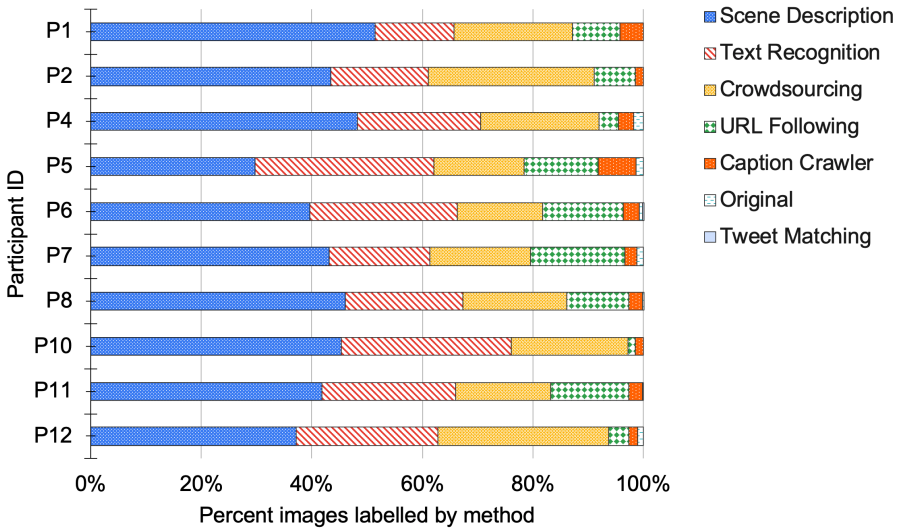


Figure 4.4: A breakdown of which method provided alt text for the images requested by a participant in our user evaluation. The bars are normalized to the number of requests for each user.

We asked each participant to rate the six methods used by Twitter A11y to generate alt text on a Likert item scale from Not useful at all (1) to Extremely useful (5) (see Table 4.6). When using Twitter A11y, the participants could tell which method generated a result, as each alt text was preceded by "From [method name]:". Some participants stated they did not see any examples of a particular method, in which case they were unable to give an answer. We also asked participants to order the methods they preferred from best to worst. Because the majority of participants only felt comfortable ranking scene description, text recognition, and crowdsourcing, we only show the mean ranks for those. We do not report statistical testing for these responses, as not all participants were able to rate or rank every method, resulting in a small sample size.

Note that this self-reported metric of "usefulness" does not distinguish between the quality of the alt-text received and the participants' perceived value of the method itself. Future work could attempt to separate these two factors. For example, when assessing the usefulness of a

Method	Mean Usefulness	Mean Rank
Caption Crawler	4.8	N/A
Tweet Matching	4.5	N/A
URL Following	4.3	N/A
Text Recognition	4.3	1.3
Automatic	3.7	3.1
Crowdsourcing	3.6	3.4

Table 4.6: Participant ratings of methods on a scale from Not at all useful (1) to Extremely useful (5). Participant mean rankings for the three most common methods (with 1 being their first choice).

description method for a particular tweet image, some participants reported the value depended on both the content of a description along with the context of the surrounding tweet. As P1 describes:

> *[Scene description] is useful when the tweet itself gives some context so if a tweet talks about a protest, and [scene description] says "a person holding a sign", then that makes sense because I understand that they're protesting but it doesn't give a ton of detail.– P1*

When assessing the usefulness of a description method as a whole, participants also considered the usefulness of the method for their particular timeline. As P4 reports:

> *[Text recognition] is one of the most useful because a lot of my timeline is tech and geek stuff which often has screenshots. Text recognition is key to understanding what's going on. – P4*

Inaccurate descriptions could lead people with vision impairments to believe a tweet image contained something that was not actually present. We asked participants if they ever felt like they did not trust a caption because it seemed inaccurate. Several participants stated they read 1-2 captions from scene description method that did not make sense with the context of the rest of the tweet, but otherwise everyone said they assumed the tweets were accurate. This implies that future versions of Twitter A11y should integrate cautionary text explored by MacLeod *et al.* [61] to encourage distrust in automatic captions when uncertainty is high.

In general, participants felt Twitter A11y could be most improved by speeding up the responses from crowdsourcing, as they did not want to wait minutes for a worker to describe the image. Participants also wanted flexibility to choose the method by which an alt text is generated, possibly through keyboard commands or a menu. Additionally, they wanted to be able to use multiple methods together, specifically scene description and text recognition, to get a sense of the image and recognize any specific text. Finally, participants were eager to see Twitter A11y integrated into their preferred clients, as they find them more usable than the Twitter web interface.

4.5 Implications for social media platforms

The participants in our user study expressed that Twitter A11y offered an impressive level of accessibility compared to what they typically find on social media platforms. This mirrors our findings from the analysis of static sample of tweets in blind users' timelines, which increased the presence of alt text from 7.6% to 78.5%, with an estimated 57.5% of descriptions being rated "Good" or "Great". We integrate our findings from these analysis to discuss the alt text generation methods holistically.

Participants preferred the text recognition and automatic captioning methods because they were quick (~2.5s) and often descriptive (~67-77% "Good" to "Great"). While most participants were familiar with these methods from other applications, they expressed that having the alt text automatically attached to images made their experience much more accessible. In our static analysis, we see the crowdsourcing provides the highest percentage of "Great" alt text (62.5%), but offers the lowest value (only 2-3 "Good" or "Great" captions per dollar) due to the expense of

paying human annotators. Participants also perceived crowdsourced descriptions as accurate, but too slow (~2 minutes) to wait for when browsing social media sites. The tweet matching, URL Following, and Caption Crawler methods were highly rated by the participants who encountered them, but results with the methods were too rare for all participants to form an opinion. However, we only used Caption Crawler when no other automatic methods produced a result (6.3% of sample), and the coverage results in the static analysis indicate it had results for many more images (19.2% of sample) that Twitter A11y did not use, suggesting that the priority of methods could be re-examined.

For social media platform designers and application developers seeking to add automatically generated alt text, we would recommend using methods that produce cheap, high-quality results first. This would include text recognition, followed by scene description methods. Caution should be used when integrating the latter, as prior research has shown that inaccuracies are not easily noticed by people with vision impairments [61, 89] . Other methods do produce additional alt text, including URL following and Caption Crawler, but the low-quality results indicate they should be a low priority. Crowdsourcing is clearly the solution that produces the highest-quality alt text, but asking crowd workers to label images is likely prohibitively expensive at scale. Instead, designers and developers should explore if they can design features to support friends and other volunteer in adding alt text [11, 12].

Several participants indicated that other social media platforms include a higher frequency of inaccessible images, including Facebook and Instagram. We designed Twitter A11y specifically for evaluation on the Twitter website, but there is strong indication that this tool would be useful on other websites. Participants were unanimous in their belief that Twitter A11y would work equally well if deployed on other social media platforms they used, and the methods that provided high coverage of images and high-quality captions are readily applicable to other platforms. The only method that could not be easily re-engineered for other platforms is tweet matching, which was not used in the static evaluation as screenshots of tweets are a rare image category.

Two participants in our user study raised the importance of distinguishing accessibility and accommodation. They viewed Twitter A11y's efforts as important to provide reasonable accommodations for images that were not made accessible from the start. However, they were not willing to use this tool unless it also made an effort to increase alt text provided by end-users who uploaded photos. The image posters have important contextual knowledge, and even the best crowd worker will not fully understand their intent when posting the image or all important details (*i.e.,* names). The participants suggested that Twitter A11y automatically notify the image poster that a blind user found their post inaccessible, and provide instructions on how to add image descriptions on Twitter. We agree with the participants that social media platforms should consider additional accessibility features and user education that could improve accessibility, not just rely on accommodations such as scene description methods. Some recommendations specific to Twitter are to increase enable image descriptions by default, train users on what comprises good alternative text, and give users feedback on the alternative text they write.

As members of the research team (all sighted) tested Twitter A11y, we were surprised at how useful alt text could be even in conjunction with seeing the image, indicating the tool could provide value for sighted users. The image captions served as quick summaries of a scene, and provided additional context. Specifically URL following, scene description, and Caption Crawler often added the names of people and places or described events that were not easily discernible

from the images (see Figure 4.1). Additionally, when an image contained alt text written by the original image poster, it served as an indication that they valued accessibility for people with vision impairments. In contrast, the constant lack of original alt text served as a reminder that the majority of images are inaccessible and why image descriptions are valuable.

4.5.1 Limitations & future work

The major limitation with our evaluation of Twitter A11y is the rather short week-long evaluation with small number of participants (10) completing the study. A longitudinal study with more participants would be necessary to understand any behavior change of Twitter users due to increased accessibility of images. Additionally, we asked participants to use the Twitter web interface, which was typically not their preferred client, so an evaluation of Twitter A11y that was more tightly integrated with their Twitter client of choice would likely yield results more representative of typical use. Finally, a major area for future work is validating that our rubric of alt text quality aligns with the expectations of blind users. While the rubric was constructed based on prior work with blind social media users, it has not been validated to ensure that the 4-point rating scale accurately captures different levels of "usefulness" that people with vision impairments might desire.

Our interviews with participants and evaluation of a tweet sample indicate other avenues for future work. First, participants raised the desire for an integration of multiple methods, such as scene description and text recognition. Additionally, Twitter A11y currently tries each method in a sequential order until an alt text result is found, but our static evaluation revealed overlap between some methods. If there was a clear approach to score the quality of image descriptions from multiple methods, Twitter A11y could ensure the best alt text is always returned. In Chapter 3, I briefly suggested generating automatic feedback for users while they write image descriptions based on the post text and the image description. The inclusion of these language features and features from the image itself could be adapted to develop a ranking algorithm (proposed in Chapter 5).

4.5.2 Automated and human methods for social media accessibility

This evaluation of Twitter A11y shows that social media platforms could deploy automatic methods to make certain types of content accessible at a quality level that would be acceptable to many users. This is clearest in the case of text recognition, which often makes photos of text or screenshots of text accessible with low error rates. However, we must remember that human-authored descriptions will always contain more *content knowledge*, information about the image contents and surrounding context, than automated approaches. Even in the case of human authors in the form of crowd workers, the original image poster knows more about the intent of the social media post. Therefore, platforms must not ignore truly accessible image descriptions by providing an automatic solution to re-mediate inaccessible content.

The prior evaluation of user-provided descriptions on Twitter shows that just enabling the capability to add descriptions is not enough. To increase the prevalence of image descriptions, platforms should explore tools that encourage and help people make the content they create accessible. The crowdworkers employed by Twitter A11y produced descriptions of slightly higher

quality than those of original posters, perhaps because they were given specific instructions, or *access knowledge* on how to write great image descriptions. In the next chapter, I introduce and evaluate an interactive tool that seeks to provide access knowledge to novice users. Platforms can utilize similar methods to ensure user provided images, or other forms of media (*e.g.,* videos, augmented reality), is accessible before resorting to automated methods.

We only address image accessibility in Twitter A11y, but other forms of visual media, such as GIFs and videos, were reported by participants to be inaccessible on social media platforms. As GIFs are short looping animations, they straddle the line between a static image and a longer video. An exploration of the best way for Twitter A11y to make these accessible might explore the use of an alt text description versus an audio description that describes the action in the GIF (see Chapter 7).

It is unlikely that the methods we tested will be directly applicable to creating audio descriptions, so new avenues will need to be explored to address video inaccessibility. There is not yet a robust method for automatic description of actions in videos [76], but there is a dataset of audio descriptions for movies to encourage future research on generating video descriptions [86]. Additionally, the YouDescribe project [80] has demonstrated how dedicated volunteers can describe videos and share audio descriptions through browser extensions, meaning Twitter A11y could explore a crowd workflow for generating audio descriptions on social media networks.

4.5.3 Conclusion

The lack of accessible content on social media platforms is a major barrier for participation by people with disabilities. Our participants echoed this in their interviews, stating that it was their primary concern and they often had to find workarounds for images without alt text. Making the deluge of user-generated content accessible, at scale, seems challenging, but platforms such as Facebook are attempting this.

Twitter A11y represents an attempt to merge promising methods for finding or generating new alternative text into one tool that users can use on Twitter. We demonstrated Twitter A11y's ability to take the content followed by a blind user from 7.6% to 78.5% with accessible images. Of these images, 57.5% of descriptions receive a "Good" or "Great" quality rating.

This tool represents a large leap in making content on these major platforms accessible, and we believe it could be easily modified, refined, and deployed on other social media platforms that include images with limited alternative text (Instagram, Reddit). We also encourage social media platforms to take note of the success of some of these methods, especially text recognition and automatic captioning, and integrate them into their platforms to improve accessibility for people with vision impairments.

Chapter 5

Automated Quality Assessment of Alt Text

The interfaces provided to add alternative text descriptions on social media sites typically lack much instruction for novice image description authors. For example, the help text on Twitter specifies that "good descriptions are concise, but present what's in your photos accurately enough to understand their context." Novice describers, therefore, may struggle to understand what information to include or how to structure the text to highlight the most important information first, especially in the short moments when they are preparing a photo for upload. To address this, I applied the image description quality scale described in Chapter 3 to help promote describing the focus of an image, actions occurring, and other image elements such as text present. I then designed HelpMeDescribe, an automated system to rate description quality and provide real-time feedback for description authors. In an evaluation of HelpMeDescribe with online crowd workers, we see an increase from 64% to 76% in the two highest rating of description quality. The adoption of HelpMeDescribe by online platforms, authoring software, and consumer devices could train users to write better image descriptions, thus building a more accessible web for all.

5.1 Guiding novice image describers to improve quality

To increase alternative text prevalence on social media platforms, we must pursue approaches that improve both the quantity and quality of alternative text on social media composed by the original poster. Alt text written by the original image poster is likely to be better than computer-generated descriptions which often lack detail or are inaccurate [61, 91]. Compared to crowdsourced alternative text, descriptions written by the original author could contain additional context, such as actions that may have occurred before/after the photo was taken, or the names of proper nouns (e.g., people, places) in the image. They can also describe context that is not visible in the image frame, such as the person holding the camera or other important background details.

People with disabilities and accessibility researchers have called on social media platforms to promote accessibility features and potentially require users to add image descriptions before images are posted. In my earlier analysis of Twitter's image description feature I echoed these calls for further engagement of sighted users, and I included several suggestions to increase the number of people using it. These include 1) enabling the feature for everyone, and 2) making alternative text content visible to sighted users. However, as there is a concern that novice users

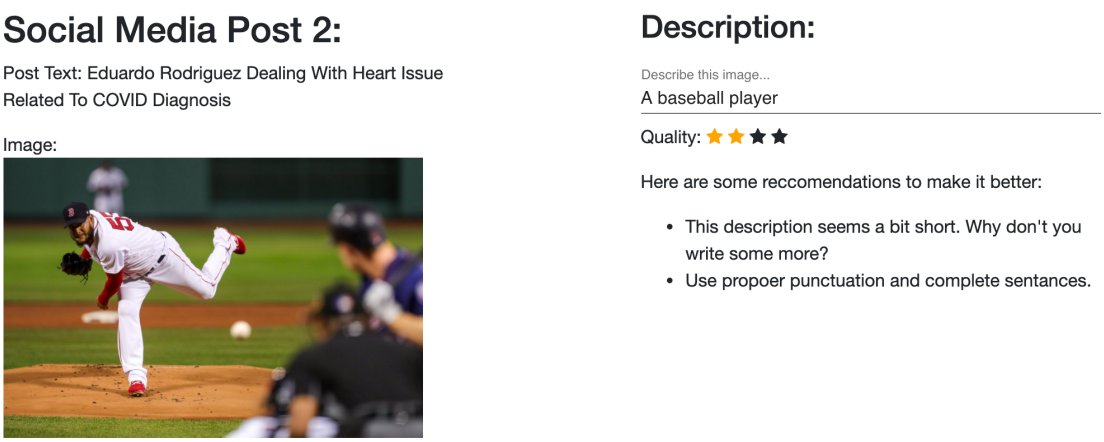


Figure 5.1: The HelpMeDescribe interface next to an example social media post. Given a social media post with text and an image, as well as a draft of an image description, HelpMeDescribe displays a quality rating (indicated by stars) and feedback to direct the author to improve their image description.

may not understand what good alternative text is comprised of, an increase in users may lead to a decrease in overall alternative text quality. Detailed accessibility guidelines exist, but they are primarily technical documents written for an audience of web developers and are not easy for users to grasp in just a few seconds. I propose a feedback mechanism that automatically displays the quality of typed alternative text in real-time, along with recommendations to improve the alternative text quality.

I introduce this approach through HelpMeDescribe, an automated approach to provide real-time quality rating and feedback to novice image description authors. Using prior work, the quality scale introduced in Chapter 3, and the results of a formative survey with 19 blind users, I identified elements of image descriptions that are important to include. HelpMeDescribe uses easily-interperable machine learning to output a rating for alt text quality on a four point quality scale, along with feedback for the author to improve their description by including this important information. I evaluated HelpMeDescribe with online crowd workers and found that workers with HelpMeDescribe increased the amount of descriptions in the "Good" or "Great" rating categories by 12 percentage points. This indicates that real-time feedback and quality assessment can change the behavior of online image describers, and this effect could be replicated if deployed by social media platforms. While I focus my efforts for online images, the addition of automatic quality assessment and feedback could also improve the quality of descriptions across all digital media, including document authoring tools (i.e., Microsoft Word) or digital photo albums. Any platform or service that allows end-users without accessibility experience to share images should encourage users to add high-quality descriptions with HelpMeDescribe.

5.2 Authoring tools for image descriptions

The typical interface for adding alternative text descriptions across most digital media includes a simple text box asking for a description. This is true on social media platforms, where that feature

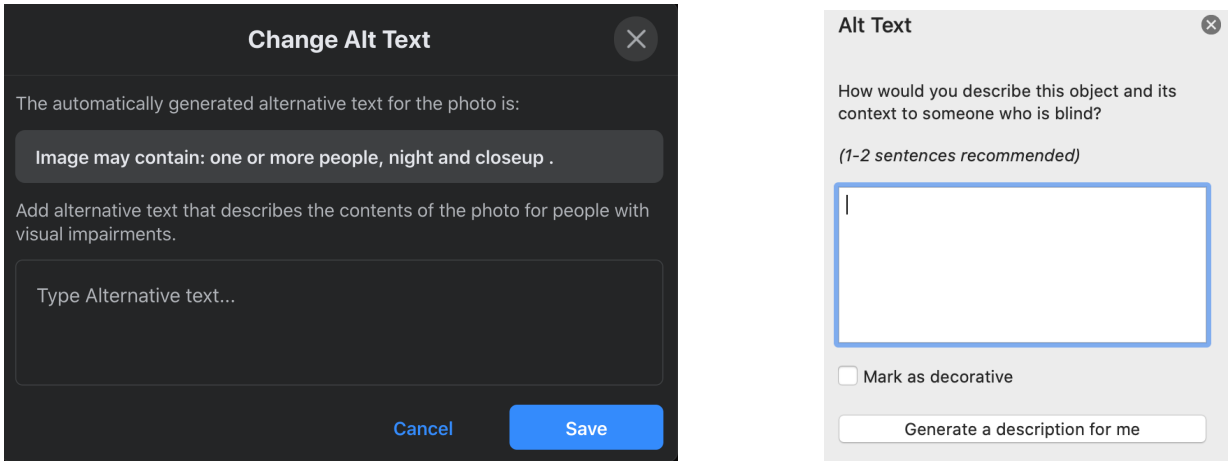


Figure 5.2: Examples of alt text entry interfaces on Facebook and Microsoft Word.

exists, as well as document creation tools like Microsoft Office [65] or Adobe Acrobat [2]. It is assumed that users will know what to describe in that text field, perhaps because the feature is typically hard to find, leaving only experienced users to add their own alternative text. The design of these interfaces differs only slightly. Facebook and Microsoft Office, for example, provide pre-filled alternative text from object recognition [112] or image captioning models [97]. Otherwise, the instructions are fairly similar:

1. Facebook: "Add alternative text that describes the contents of the photo for people with visual impairments."
2. Twitter (after clicking "What is alt text?"): "You can add a description, sometimes called alt-text, to your photos so they're accessible to even more people, including those who are blind or visually impaired. Good descriptions are concise, but present what's in your photos accurately enough to understand their context."
3. Microsoft Office: "How would you describe this object and its content to someone who is blind?"
4. Apple Keynote (on mouseover): "Type a description that VoiceOver reads aloud."

In these examples, Facebook and Microsoft office are the only interfaces that explain the purpose of image descriptions when initially opening the interface to write them. Twitter, after clicking on the help link, provides a bit more detail, instructing describers to be concise but accurate. Apple's instructions in Keynote require the user to both mouse over the text box to read tooltip text and to know what VoiceOver is and why it is used.

To learn to write better descriptions, content authors instead must turn to other sources to find descriptions. Perhaps the most well known are the Web Content Accessibility Guidelines, which states as part of their "short text alternative" technique [20]:

> The text alternative should be able to substitute for the non-text content. If the non-text content were removed from the page and substituted with the text, the page would still provide the same function and information. The text alternative would be brief but as informative as possible.

This document provides examples, although many of them are targeted at web developers,

explaining how to create text alternative for buttons or logos they may be placing on their site. Additionally, the Web Accessibility Initiative [49] provides content authors with a decision tree, informing them how to describe the image depending on it's purpose. They offer

> The text alternative should convey the meaning or content that is displayed visually, which typically isn't a literal description of the image. In some situations a detailed literal description may be needed, but only when the content of the image is all or part of the conveyed information.

These guidelines have been created for use by developers creating websites where there is typically more surrounding text context than on social media. Morris et al. found that only about 11% of images on social media were sufficiently described by the post text, meaning 89% of social media post will not be understood fully without image descriptions [72].

To expand these descriptions to audiences sharing specific media, some institutions or other practitioners write guides for other content. For example, the Diagram Center provides these guidelines about describing comic strips [17]:

> Describe the picture first to give a set-up, then write out the text. The text may be edited if it would not make sense unless there was a long explanation.

For social media specifically, disability advocate and blogger Veronica Lewis offers advice posts regarding content like memes or TikTok videos [59]. In addition to describing the visual contents of the meme including any text, she specifies that users should not be afraid to explicitly give away the punchline:

> For these types of memes, don't be afraid to share what the joke is, and don't expect your viewer to guess for themselves [59].

Researchers have additionally created tools to help novice authors write image descriptions for very challenging content, such as science, technology, engineering, and mathematics (STEM) diagrams [69]. The tool in this example asked authors questions about the image, providing a structured method for novice authors to describe individual parts of STEM diagrams such as chart titles. Image describers who used this tool preferred this method of description by query and it resulted in higher-quality descriptions.

Stangl et al. have also recently investigated the description preferences of blind people across various contexts, including social media, e-commerce, and more [91]. They find that, across contexts, elements of the photo like people present, actions occurring, and text present in the image were almost always important. Elements like people's clothing, facial expressions, or color of objects instead depended greatly on context. Perhaps some of these elements might be more important for online dating or shopping. In the social media context, they found that almost all aspects of images were somewhat relevant, presumably because the image content tends to be more varied than a shopping site or other narrower context.

HelpMeDescribe builds on prior work by adapting the knowledge inherent in static guidelines to dynamic feedback offered to the content author based on their image description draft. Novice authors who may only invest seconds into learning about accessibility can therefore get actionable feedback on improving the quality of their image descriptions, and build up their image description skills over repeated interactions.

5.3 Formative survey with blind respondents

The 4-point scale of quality proposed in Chapter 3 was designed to take into account the image description guidelines targeted at practitioners, as well as relevant insights from accessibility research. This scale places alternative text (relative to the described image) on a quality scale from 0 to 3:

Irrelevant to Image (0) : The alternative text is not descriptive of the image at all.

Somewhat Relevant to Image (1) : The alternative text broadly describes the image, but offers little detail.

Good (2) : The alternative text describes the image, and typically includes one of the person/object of focus, the action, and the setting.

Great (3) : The alternative text fully describes the image, including the person/object of focus, the action, and the setting.

A rubric detailing how I evaluate this scale on various types of image categories (e.g., screenshot or drawing) is available in the Supplemental Material.

To ensure that this quality metric is consistent with the perceptions of people with vision impairments, I organized a formative online survey with 19 blind participants to verify the survey scale. Participants were recruited by circulating an IRB-approved recruitment message on Twitter and the /r/Blind subreddit community on Reddit. I also recruited participants from a pool of prior study participants with vision impairments at Carnegie Mellon University. All participants were compensated with a $10 Amazon or PayPal gift card for completing the survey.

The survey that participants took was an online, screen-reader accessible survey designed to take approximately 45 minutes to complete. The survey consisted of three sections.

First, the participants were asked to share information about their demographics. The following section focused on image description completeness. Completeness refers to how much information was included in the description, and this measure is reflected in the quality measurement rubric introduced earlier. For this section, participants encountered 20 social media posts from users on Twitter that were anonymized to remove usernames and links. Participants were given two possible alternative text descriptions, one written by the original post author and another written by online crowd workers. The participants were asked to choose the description they preferred most. Participants were told the descriptions were accurate, meaning no information was incorrect about the visual contents of the image, but they may be incomplete descriptions of the image. I rated each description, assigning a quality score, and balanced the 20 questions to reflect a choice between different quality levels. Ten questions contained a pair of descriptions that differed by one quality level (Relevant to Image and Good, or Good and Great) and ten contained a pair that differed by two quality levels (Relevant to Image and Great). No "Irrelevant to Image" descriptions were included, as these are typically inaccurate by definition. To avoid ordering bias, all questions in this section were presented in a randomized order and all answers were also randomized.

The final section contained wrap-up questions asking participants for free-form text responses about what they thought could be improved about the image descriptions they encountered overall. All of the survey questions are available in Appendix B.

5.3.1 Findings

Five of the participants were women and 14 were men. Ten participants were blind from birth and 9 acquired blindness in childhood or later. All participants reported they were fluent in English, and they were heavy screen reader users, with 18 of 19 using a screen reader all of the time to access the internet. One respondent (P1) said they used a screen reader about half the time. Participants used a variety of screen reader across their devices, including JAWS (10), NVDA (17), VoiceOver on iOS (13) and macOS (1), Android Talkback (5), and Windows Narrator (9).

In the section designed to assess the completeness of image descriptions, the majority of participants always choose the image description with more information and a higher quality rating, with 16 of 19 participants choosing that answer on average. One question received a unanimous 19 choices of the higher-rated answer, while another split the participants with only 10 of 19 choosing the answer I rated high quality.

In the free-form text responses participants generally echoed the themes above, stating that information like the subject of the photo (especially their name if known) is the most important, followed by elements like text. Participants were divided on whether or not the image style should be specified, as some said it should be the first part of the description, while others said it could be inferred from the rest of the description and was not necessary to call out specifically. Overall, participants generally agreed that what is most important and contribute to the quality of a description is context specific: what is the purpose of this image and how does it fit in with surrounding context such as the social media post? Participants recounted examples of descriptions, human or AI generated, that were long and detailed but missed the point of the photo entirely, such as text on a food label. Quality must therefore be holistic, understanding the intent of the image poster and that contributes to the overall experience a screen reader user has with a piece of digital content, not just the relationship between the image and the textual description.

5.4 HelpMeDescribe: automatic rating and feedback of image descriptions

Based on the results of this formative survey and other findings in prior work, I designed HelpMeDescribe, a system to provide real-time quality measurements and feedback for image description authors. I trained a machine learning based classifier to output a discrete rating on this quality scale given a social media post as surrounding context, the included image, and an image description.

5.4.1 Classifying quality ratings

To assess alternative text quality, I use the same 4-point ordinal scale validated in the formative survey. I rated a sample of 1,226 images with descriptions on this scale to serve as the labelled training data. The images were sampled from those that blind users encountered while using Twitter A11y where the original poster had already added a description (Chapter 4). 20% of this dataset was reserved for the test set, and remaining 980 images served as the training data.

For the rating system to accurately judge alt text quality, I derived features that I believed may carry information about the alt text's ability to describe the image. I identified the following possible features based on the aforementioned rubric and prior work.

Alt Text Length: The quality of alternative text is related to the number of words **(Word Count)** or characters **(Character Count)** written. Very short alt text, such as a single word, are unlikely to describe the image fully. Similarly, very long alt text could be overly verbose.

Parts of Speech: The above rubric essentially premises that good alt text will include important **Nouns** present in the image, any relevant actions (**Verbs**), and perhaps additional descriptive words (**Adjectives**). A count of different parts of speech could be informative as to how well the alt text is aligning with this rubric. I include the above three parts of speech as well as **Adposition**, **Adverbs**, **Proper Nouns**, and **Punctuation**.

Congruence with image: The core measure of alt text quality is how well the alternative text describes what is in the actual image. Using object recognition models, we can recognize objects or actions occurring in the image. A feature can then be constructed measuring the similarity between the recognized objects/actions and the alternative text description **(Alt-Image Similarity)**. This is also broken out into congruence with just the nouns **(Alt-Image Noun Similarity)**, verbs **(Alt-Image Verb Similarity)**, and adjectives **(Alt-Image Adjective Similarity)** in the image tags.

Congruence with post text: MacLeod *et al.* found that similarities between the text of the social media post and the alternative text made a more understandable story for blind users. We can measure the word similarities between the two pieces of text to see how congruent they are **(Alt-Post Text Similarity)**. However, low congruence may not indicate poor alt text, as the post text could be simple or non-descriptive.

Recognized text in image: If an image contains text, good alt text should include a transcription of the text. I include a feature called **(OCR Text Length)** to measure the number of words recognized by optical character recognition. I also extract the text in the image and measure if this text is represented in the alternative text **(Alt-OCR Similarity)**.

Grammatical correctness: Some search engine optimization guides recommend that web users place keywords in image alt text fields. As some search engines do utilize alt text for page ranking, this has become popular advice for people looking to raise their site's visibility. As this hurts accessibility, I would like to measure if alt text contains grammatical structure instead of unrelated keywords. A somewhat naive approach to this is to utilize the **"Perplexity"** of a natural language model. This perplexity gives a measure of how likely a specific piece of text is to occur. I also include the **Lexical Density** to measure the complexity of the description, which is the number of non-grammatical lexical words (e.g., nouns, verbs, etc,) divided by the total word count.

Together these 18 features embed information about the image description, social media post, and image contents. I trained the classifier using a random forest ensemble model [47] included in the Python package scikit-learn [79]. On the test set, the classifier achieved an accuracy of 62.6% with a mean squared error of 0.60. Looking at the following confusion matrix Figure 5.3, we see that it failed most on distinguishing between the middle rubric scales.

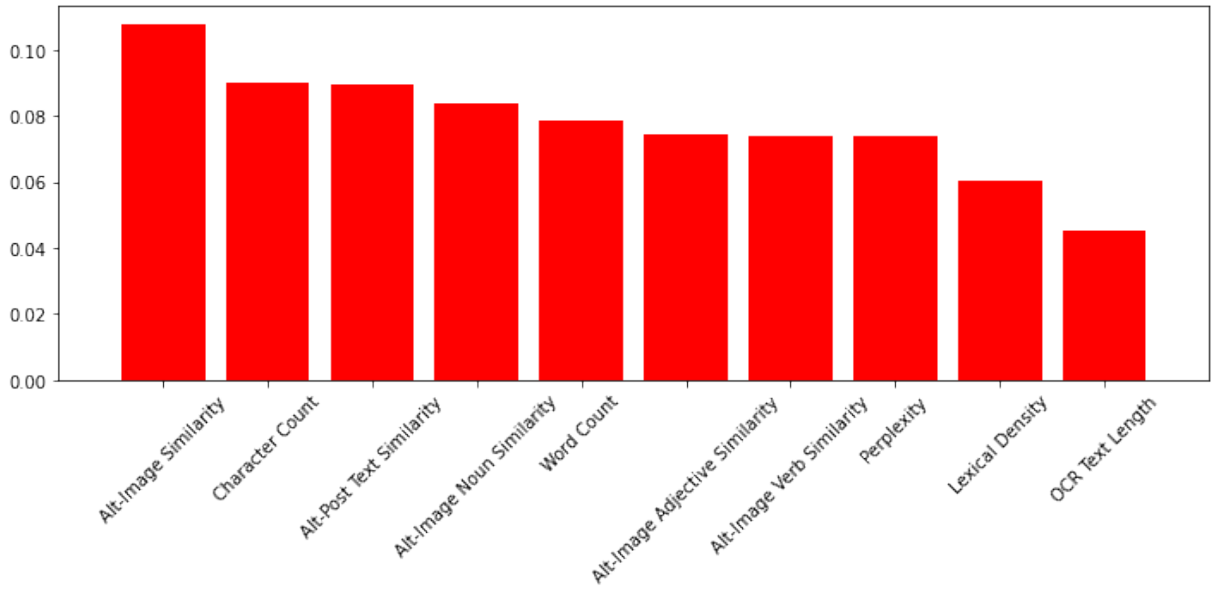


Figure 5.4: These bars represent the relative feature importance in the HelpMeDescribe model for the top 10 features. The importance is approximately the weight each feature contributes to the model.

Figure 5.3: The confusion matrix of the trained classifier. We can see that most mistakes are between the middle two points on the rating scale.

I analyzed the most important features to the classifier, finding that the top 3 important features for discerning quality were the congruence between alt text and recognized image tags, character count of alt text, and congruence between alt text and post text. The feature importance for the top 10 features are in Figure 5.4.

5.4.2 Determining features that a description lacks

A system that rates the quality of alternative text is useful by itself to analyze the accessibility of a social media platform or even to choose between multiple alternative text candidates. However, we seek to employ this as a feedback mechanism to help novice alt text writers improve their accessible descriptions. This feedback mechanism will both output a quality rating, and give users specific recommendations on how to improve their image description.

To do so, I determine which features contributed most to preventing a description from achieving the highest possible rating. For example, if that factor happened to be the length

of the alternative text, HelpMeDescribe says "This description seems a bit short. Why don't you write some more?" The benefit of using the random forest classifier to determine the quality level of a description is that the contributions for each feature to the final classification is easily interpreted. I extract the per-feature continuations for a given example using the Python tree interpreter package [88]. HelpMeDescribe then ranks each feature by the amount of probability mass it is responsible for moving from higher ratings to the classifier's chosen rating.

5.4.3 Recommendations interface

To make this quality assessment and analysis of features that might be lacking in the image description usable to a novice image describer, I designed a simple interaction modeled after password strength meters and requirements [38]. The user is given a text entry box to write the image description in, a four-point star rating scale, and a list of recommended improvements to their description. When the user writes in the image description box, the social media post text, included image, and description are sent to a server to extract features, classify quality, and determine features contributing to low quality as described above.

For each feature that is contributing to low quality, HelpMeDescribe provides a text recommendation to the user. These are tied to the features extracted above, although some share the same recommendation text. Duplicate recommendations are removed before showing them to the user.

Word Count or Character Count This description seems a bit short. Why don't you write some more?

Alt-Image Similarity Items in the image are not described. Try describing more of the objects in the image! Some suggestions: [recognized objects in image]

Perplexity This description seems simple or not grammatically correct. Please use complete sentences.

Lexical Density Your description should describe the image, but otherwise be concise. Make sure you remove redundant words.

Alt-Post Text Similarity The post text mentions other parts of the image. Could you add some of that to your description?

OCR Text Length or Alt-OCR Similarity : There is text in the image. Make sure that text is included in the description!

Nouns What is in this image? Add more nouns!

Verbs Describe more of the action happening in the image. Add some verbs!

Adjectives Be more descriptive about the contents of the image. Add some adjectives!

Punctuation Use proper punctuation and complete sentences.

Alt-Image Noun Similarity Make sure you are describing all of the important people, objects, or other things in the image!

Alt-Image Verb Similarity What is happening in this image? Make sure you are describing what action is occurring, if any.

Alt-Image Adjective Similarity What details are in this image? Be descriptive about any relevant colors, sizes, or textures that are important to know.

Users are given up to three recommendations per classification of their description, although these recommendations will change as the user continues to edit their description. To limit recommendations for features that will not affect the quality much, I empirically chose a threshold of 5% probability. Features that are contributing less than 5% probability to a lower quality rating are ignored for feedback purposes and not shown to the user.

5.5 Comparative evaluation with static instructions

To evaluate if the addition of real-time quality ratings and feedback recommendations improved the quality of written descriptions, I conducted a comparative assessment on Amazon Mechanical Turk (MTurk).

5.5.1 Materials

I selected 5 tweets with images for each MTurk worker to label. The tweets were chosen to present a few different aspects of images that are important to describe: all have some sort of person or animal as a focus, 3 have some sort of text in or overlaid on the image, and at least two have physical actions that can be described. One was a photo of a drawn cartoon, which also represented an image style slightly different than photographs of natural scenes.

I prepared a web page where each of the 5 tweets were shown sequentially in a random order. Next to each tweet and image was a text entry box to describe the image. Two versions of this web page were prepared for two different conditions. The first condition, ***basic instructions***, contained just the tweet and these basic instructions:

- Please describe the images in the following social media posts for someone who is blind or has a vision impairment.
- Good descriptions present what is in the photo accurately, but are also concise.
- Use punctuation and don't mention that you're describing an image.
- Each description that meets our highest quality rating will after review earn a $0.10 bonus (up to $0.50 total for task).

The instructions were based on the default message for image captioning tasks on Amazon Mechanical Turk as well as the instructions given to image description authors on Twitter and Facebook. The bonus was offered to equalize the incentive offered with that in the second condition.

The second condition, ***real-time feedback*** integrated HelpMeDescribe's quality assessment and real-time feedback below the text box. Quality was indicated by a star rating scale consisting of four stars. Up to three pieces of feedback were shown below this. The quality assessment and feedback were updated whenever the user's cursor left the text box or they pressed the enter key. This condition also contained the above instructions, although the bonus instruction was "Use the automated feedback to improve your descriptions. Each description with a 4 star quality

rating before submitting will earn a $0.10 bonus (up to $0.50 total for task)." When a description reached a 4-star quality rating, "+$0.10 bonus" would appear alongside it to indicate the worker had achieved the bonus.

I report the results of this comparison between conditions across two sessions, one where HelpMeDescribe was trained on descriptions from Twitter users, as described above, and one where it was trained on descriptions written by crowd workers.

5.5.2 Session 1: Original poster model

In this session we recruited a total of 87 workers to write image descriptions for the 5 selected images. Workers were required to have completed at least 1,500 tasks previously on the platform and have an approval rate of 98%, as prior experience shows that these requirements significantly reduce spam from new and unverified accounts. Workers who accepted our task first completed an IRB-approved consent form before they were randomly assigned to a condition and moved on to the task. Due to worker drop-off, the conditions were slightly imbalanced, with 42 in the basic instructions condition and 45 in the real-time feedback condition. Workers were compensated $1.50 per task with an additional bonus of up to $0.50 depending on their description quality. Based on prior experience with similar image description tasks, we estimate that may result in an hourly wage between $8 and $11.

To fairly analyze the collected image descriptions and rate them for quality, I randomly sorted all of the descriptions submitted in both conditions. Then, for each of the 5 images used in the study, I rated all of them by applying the quality scale using the provided rubric (in Supplemental Materials) without knowledge of the condition. I chose to assess all of the descriptions for a single image at the same time to consistently apply the rubric to that image.

Even with some of the qualification precautions we took, some workers submitted descriptions that constituted spam. These were either automated accounts that used a search engine to find a relevant piece of text or multiple workers who all submitted the same exact description for all of the five images. Descriptions that fell into this category were marked, and all of the worker's descriptions were subsequently analyzed to determine if they were engaging in these spam behaviors. Spam accounts were removed from the dataset in their entirety, and otherwise all work form all other workers was kept. After this step, the basic instructions condition contained 25 workers (125 descriptions) while the feedback condition contained 21 (105 descriptions).

The descriptions submitted in each condition were roughly the same length, with an average of 22 words in the basic instructions condition and 20 words in the feedback condition. The quality ratings are presented below in Table 5.1.

The real-time feedback condition in Session 1 with HelpMeDescribe seems to have reduced the quality of the resulting descriptions, lowering both the "Good" and "Great" levels by 3 percentage points each. Why would feedback lower the resulting description quality? Analyzing the descriptions submitted by MTurk workers, I found that most of them immediately received a rating of 3 from HelpMeDescribe, which triggered an indication for a bonus payment. The model was therefore encouraging workers to submit their first draft instead of editing it to integrate the feedback. This seemed to be because HelpMeDescribe was trained on image descriptions written by social media users, who have no monetary incentive to write high quality descriptions.

Quality Rating	Basic	Feedback (Session 1)	Feedback (Session 2)
Irrelevant	11%	13%	2%
Relevant	25%	29%	21%
Good	28%	25%	32%
Great	36%	33%	44%

Table 5.1: The distribution of image descriptions from crowd workers in the basic instructions condition and real-time feedback condition in both sessions.

Alternatively, MTurk workers might be comfortable writing image descriptions due to past tasks they have completed, and therefore are able to write a high-quality initial description.

5.5.3 Session 2: Crowd worker model

It would be ideal to evaluate HelpMeDescribe on novice image describers on a social media platform, but any influence due to study participation compensation would likely change the quality of descriptions. Even without compensation, just participating in a study focused on the task of writing image descriptions is likely different than the normal social media context. Instead, on Mechanical Turk the incentive structure is typical, and we can hold it consistent between conditions. Therefore, to improve the quality assessment given by HelpMeDescribe, I retrained the model on image descriptions authored by MTurk workers to tune the quality assessment for the same context as evaluation.

I sampled 423 images with descriptions written by crowd workers during the Twitter A11y study (Chapter 4) and labelled them with the same quality ratings. As above, 20% of this dataset was reserved for the test set, and remaining 338 images served as the training data for the model. This model received a lower initial accuracy score of 50%, with a mean squared error of 0.66.

I then recruited an additional 35 crowd workers to complete the same task in the real-time feedback condition, but using this model instead. 8 workers were removed for spam, leaving a remaining 27 workers and 135 descriptions for analysis. As before, the descriptions submitted in each condition were roughly the same length, with an average of 22 words in the basic instructions condition and 23 words in the feedback condition. The quality ratings are presented above in Table 5.1.

With this model trained on MTurk provided descriptions, there is a modest increase in the feedback condition in the quality ratings. To test if this increase was significant, I performed a Mann-Whitney test between the quality ratings in each condition, finding that quality was greater in the feedback condition (mean $= 2.2$) compared to the basic instructions condition (mean $= 1.9$). This was significant at $p < 0.05$ ($U = 7161$, $p = 0.035$).

The slight decrease in quality in the first session and the increase in quality in the second session suggests that HelpMeDescribe is sensitive to the context it is deployed and the data it is trained on. Using descriptions from social media to train a model for MTurk tasks seems ineffective. This could be due to different styles of description leading to different features emphasized in the model. Future work could examine how much can be transferred between contexts such as document editing and social media. An additional promising area of study

would be to assess other forms of quality rating, such as decomposing overall quality into a set of objective metrics that may be more consistent across environment.

5.6 Future use-cases and improvements for HelpMeDescribe

This work is a preliminary study in the design and use of an automated system to help novice image describers improve their descriptions. However, HelpMeDescribe could be utilized in other assessments of image descriptions, such as automatic auditing of applications, websites, and documents that include alternative text. The quality ratings could also be useful as an alternative quality metric for automated image captioning systems as an external oracle to choose the best caption from a set of suggestions. Even a system like Twitter A11y could utilize the quality ratings to determine which image description to return to the user when attempting multiple methods.

The feedback mechanism, if extracted, could additionally be utilized to augment poor descriptions. For example, if an automated image caption described the visual contents of an image but resulted in a low quality description, HelpMeDescribe might indicate that an activity recognition model to describe the missing action or text recognition for missing text content. The integration of surrounding post context would assist these other image description approaches in understanding what aspects need to be described, and what may be avoided or focused on depending on context.

5.6.1 Lessons for other social media formats

The goal of HelpMeDescribe is focused on improving the image descriptions written by people who are novices when it comes to accessible digital media. However, the general goal can be more broadly defined as helping any content creator ensure their human-authored accessible content alternatives are high-quality. Whether that is helping someone create an audio description for a video [78], developers ensure their smartphone applications are screen reader compatible, or augmented reality designers (such as Snapchat filters) integrating non-visual content, the general approach of the previous chapters can be replicated to build similar systems to HelpMeDescribe. Using qualitative and quantitative research methods, one should ensure the know the important factors for accessible content, as well as the motivations and experience of those creating it. They should examine aspects of the created content that may impact the accessibility end-result, as well as surrounding context that may indicate other useful information. And then a feedback system like can be designed to provide real-time feedback to content authors to ensure greater accessibility. HelpMeDescribe likely will perform best in domain-specific areas where the content and accessible alternative (i.e., image description) are more uniform. For example, a system designed specifically to help people describe graphs may provide more tailored and precise feedback than one designed to help all authors on social media. Therefore, any specific genres of content that can be identified and separated should be used to provide domain-specific feedback.

5.6.2 Limitations and future work

While there was an improvement between the conditions in the second crowd evaluation session, the accuracy of HelpMeDescribe and sensitivity to context indicate there is room for future work to improve this approach. First, the list of features chosen for the classification model is limited, especially when image features could be extracted using neural networks. However, doing this robustly would likely require a larger amount of data to train HelpMeDescribe, which would need to be labelled with quality ratings. With the appropriate dataset, deep learning approaches could be pursued, and approaches that learn a shared embedding for the image and description might be best adapted for this task [84]. If a larger dataset was collected, it would be useful to have more fine-grained approximations for quality that could be used as an intermediate measure. For example, "Is the subject of the image described?" and "How much of the image text is present in the description?" may be useful metrics to interpret when choosing which feedback to recommend. Additionally, these metrics may be more consistent for descriptions sourced form different environments (i.e., social media vs. MTurk).

5.6.3 Conclusion

A goal for all social media platforms should be to ensure the images and other content their users create and share is accessible to people with disabilities. While they can encourage unfamiliar users to utilize accessibility features to describe images, the quality of the generated descriptions will depend on the instructions given to the novice describer. I have introduced HelpMeDescribe, which utilizes features of the image, description draft, and surrounding context to provide real-time quality ratings and feedback for authors to improve their descriptions. This real-time feedback increased the quality of descriptions authored by crowd workers on Amazon Mechanical Turk, although HelpMeDescribe was only effective for this group when trained on examples from Amazon Mechanical Turk, suggesting that incentives and training may differ across user environments. This approach offers additional promises for utilizing automatic quality ratings of descriptions alongside automated description generation methods for quality control or description selection. Social media platforms should integrate HelpMeDescribe to more aggressively push alternative text features to unfamiliar users and bring us closer to making images on the web accessible at scale for people with vision impairments.

Part III

Novel Accessible Formats for Social Media Content

Chapter 6

Making Memes Accessible

Through the user interviews with both image description authors and Twitter users with vision impairments, I noticed that some forms of images are common or "native" to social media platforms. While photographs, advertisements, and charts appear all over the web, users more commonly posted things like screenshots or internet memes on social media. As noted by a participant, memes like the one in Figure 3.5a are not straightforward to describe without hindering the humor it is trying to convey. Realizing that this was a challenge, I investigated how one might compose alternative text for memes in a way that preserves their humor or emotional tone.

Work in this chapter was also published as a conference paper. The use of "we" in this chapter refers to all of the authors who contributed to that work. The full citation for that article is:

Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B. Chilton, and Jeffrey P. Bigham. 2019a. Making Memes Accessible. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 367–376. 9781450366762 http://dx.doi.org/10.1145/3308561.3353792

6.1 Memes: an image type native to social media

Increasingly, people communicate on social media networks and in personal chats using visual content (*e.g.*, emojis, memes, and recorded images/videos). However, a large amount of the visual content on social media networks and personal chats remains inaccessible due to a lack of high-quality image descriptions. Social media platforms like Facebook [112], Twitter [101], and Instagram [50] allow users to add alternative text to their images, but most do not use this feature resulting in only 0.1% of images becoming accessible. Because social media platforms and users do not include high-quality alt text with all images, we explore how to exploit repetition in the common content users share over time. A large number of images shared on social media are not original images. In fact, the analysis in Chapter 3 revealed that of a sample of over 1.7 million photos, 80% were retweeted images. In this chapter, I focus on a class of image content which affords opportunities to leverage this repetition – memes.

Broadly, a meme is "an idea, behavior, or style that spreads from person to person within a culture – often with the aim of conveying a particular phenomenon, theme, or meaning repre-

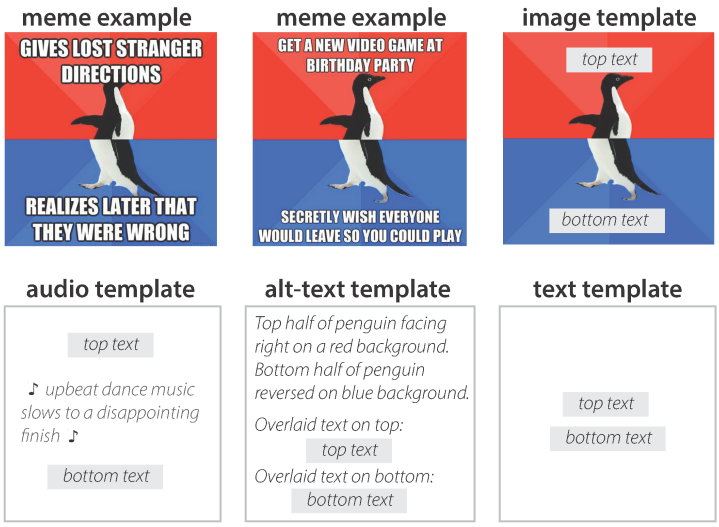


Figure 6.1: Image macro memes feature a meme example that can be described with an image template. We propose alternative forms of meme description including audio, alt-text, and text templates.

sented by the meme"[1]. We focus on image macro memes [27], a common form of image-based meme that features an image overlaid with caption text (Figure 6.2). Sharing an identifiable image macro meme can serve as shorthand for "a phenomenon, theme or meaning". For example, the celebrating toddler image represents "common situations with minor victories" (Figure 6.2A), and the crying woman image represents "first world problems" (Figure 6.2B). However, existing alt text for image macro memes typically describe only the meme text (*e.g.*, "Put candy bar in shopping cart without mom noticing"), dropping the relevant context provided by the template. Without the context recognized through the images, the memes often lose their emotional tone or humorous aspect.

To make memes more widely accessible, we propose 1) an automatic method for applying existing image descriptions to new meme examples, and 2) a non-expert workflow for creating high-quality alt text and audio macro meme templates. Our automatic workflow classifies a meme example with 92% accuracy and recognizes meme example text with a 22% word error rate (9.2% by character error rate).

To understand user preferences for an accessible meme format, we conducted a user study with 10 visually impaired participants comparing 3 different meme formats: meme text only, image description with meme text, and meme text with a unique tonally-relevant background sound (created by a sound designer). While users preferred image descriptions, we find that our traditional image descriptions occasionally fail to efficiently convey the function of the image (*e.g.*, shorthand for tone). For audio, despite quickly conveying tone, a background sound can lack universal accessibility. Based on user performance and preference, we propose structured questions for creating image descriptions for image macro memes.

In summary, our contributions are as follows:

[1]https://www.merriam-webster.com/dictionary/meme

Figure 6.2: Examples of image macro memes from two image templates. Template A represents the "Success Kid" meme and Template B represents the "First World Problems" meme.

- An automatic process to recognize known memes and extract new text,
- An interface for creating accessible memes in alternative text or audio formats with placeholders for the extracted text, and
- Structured questions to be used for alternative description formats for visual image content, specifically memes.

6.1.1 Background: memes and humor

Meme are challenging to describe in alternative text because they contain humor. According to the Semantic Script Theory of Humor [85], what is communicated in humor is *implied* rather than stated directly. According to this theory, jokes have a set-up and a punchline: the set-up leads the listener to expect one thing, but then the punchline violates that expectation and forces the listener to think of a second interpretation that connects both statements. Often the second interpretation involves an insult or an error in logic [62]. For example, in Figure 6.2, the "Success Kid" meme (Template A) has set-up text at the top saying "[I] put candy in the shopping cart", which is a normal thing to do. Then there is a picture of a toddler looking very proud of himself, and a punchline reading "without [my] mom noticing." This implies he did it sneakily and he is proud that his mischievous act was not punished. Additionally, the speaker is exaggerating how big this accomplishment is. It is relatively minor, but the serious look of success on the kids face implies he is treating it as a big accomplishment. This is the error in logic, and perhaps a self-effacing insult that is meant to make it humorous to the reader.

Understanding humor relies on a shared context of the speaker and the listener in order for the listener to infer the correct meaning. This is difficult for both people and computers. Although many computer programs have been trained to detect humor, most struggle to achieve more than 80% accuracy over a 50% baseline [18, 55, 67, 90, 94]. This is likely because of the immense amount of cultural background as well as necessary ability to interpret the hidden meaning that is

required. Additionally, people outside of a culture context often find that culture's humor difficult to understand. A study of people unfamiliar with memes or meme subculture [60] found that memes were very hard to understand. They tested several ways of elaborating or explaining the memes and found the most successful strategy was to provide crowdsourced annotations which explicitly described the implied meaning according to the Semantic Script Theory of Humor. As noted by the common quotation [108], "Humor can be dissected, as a frog can, but the thing dies in the process and the innards are discouraging to any but the pure scientific mind." In this vein, there is a challenge in making the content of a meme more accessible, while still leaving the meaning implied, so that the joke can be enjoyed as intended.

6.2 Making memes accessible

To transform image macro memes into accessible alternative formats, we provide 1) an *automatic method* for converting image macro memes encountered on the web into alternative meme formats, 2) an *authoring interface* for generating meme alt text templates and audio macro meme templates. As each meme template can apply to thousands of instances of the same base meme, our automatic method allows people browsing the web to convert existing image macro memes to preexisting alternative meme template formats (*e.g.*, meme text, alt text, audio meme). Our authoring interface enables non-experts to efficiently produce meme template alternatives.

6.2.1 Automatic method

We automatically convert existing image macro memes encountered in the wild to alternative meme types by: 1) recognizing that an image is a meme, 2) identifying the meme type (*e.g.*, "success kid", "confession bear"), and 3) extracting the text from the meme (Figure 6.3). We then insert the extracted text into the alternative text templates textually or audio macro meme template using text to speech.

Meme recognition

When a user encounters an image on a social media network (*e.g.*, Imgur, Twitter), we first detect whether or not the image is a meme using Google Cloud Vision API's "Detecting Web Entities and Pages" request. For a given image, we obtain a list of web-generated labels (*e.g.* "Meme, Success Kid, Toddler, Brother" for the Success Kid meme) and we check if the keyword "meme" or "internet meme" appears in the list of labels. We evaluated this method with 105 meme images randomly selected from the "Meme Generator Dataset" from Library of Congress's Web Archive [73], and 105 non-meme images (a random subset of the ImageNet database [26]). This method achieves a meme recognition accuracy of 94.4% (100% precision, 89.9% recall). The API typically does not include the "meme" label for new or less prevalent memes.

Meme classification

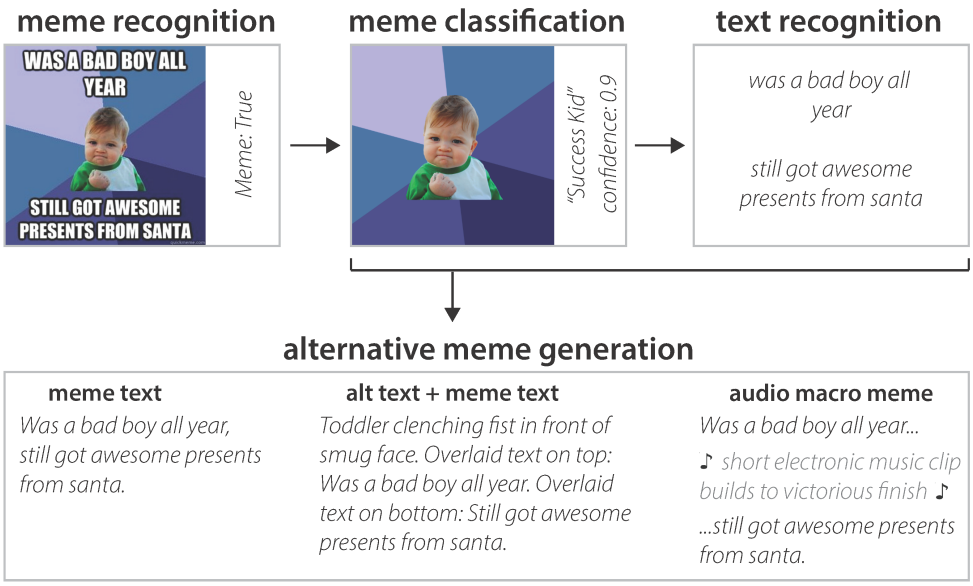


Figure 6.3: Our system first recognizes whether or not the image is a meme. If it is a meme, the system attempts to classify the meme as a representative example of a meme template in our database (e.g., "Success Kid"), and recognizes the text within the meme (e.g. "Was a bad boy all year"). If the meme classification confidence for a match (i.e. image similarity score) reaches a score over a given threshold, we output three formats: meme text only, an alt text + meme text pair, and an audio macro meme. If the confidence falls below that threshold, we output only the text.

We next match the recognized input meme to a meme template in order to identify any corresponding alternative meme representation. We create a dataset of the 137 meme templates from Imgur[2]. To automatically match the input meme image with a database meme template, we first re-size and crop the input meme image to be the same size as the templates in the database. Then, we compute for the input meme and each database meme template: 1) the structural similarity between the input image and the template image, and 2) the color histogram difference between the input image and the template image. To compute structural similarity, we use the Multi-Scale Structural Similarity (MS-SSIM) index [106] that considers the luminance, contrast, and structural similarity between image regions at various zoom levels. To compute the color histogram difference, we divide each image into 5 regions (Figure 6.4) and sum together chi-squared distance between HSV color histograms computed for each region (8 bins for the hue channel, 12 bins for the saturation channel and 3 bins for the value channel) [87]. We define the final image similar-

Figure 6.4: An example of separate regions computed for the color histogram difference measurement.

[2]https://imgur.com/memegen/

ity score between two images X and Y as: αMSSSIM(X, Y)$-\beta$ COLORDIFF(X, Y), where α and β are adjustable parameters that sum to 1. We use $\alpha = 0.15$ and $\beta = 0.85$, determined empirically. We calculate an Image Similarity score for each template with the fixed input meme example, and return the template with the highest similarity score. If the score is below a confidence threshold, we only output the meme text, as it is likely not in the database.

We evaluated meme classification with 385 memes scraped from the "most popular memes of the year" page of Imgur[3]. With the structural similarity (MS-SSIM) score alone, we achieve an accuracy of 79.22%. The structural similarity score method tends to not perform well on images with low resolution or noise, and performs well on photographs with high-contrast. The color histogram difference alone achieves an accuracy of 77.58%. The color histogram difference method often confuses images with similar colors in the same regions (*e.g.*, the nose of a black bear with a black t-shirt). The combined Image Similarity accuracy is 92.25%.

Text recognition

After we match the input meme image to a meme template, we extract the top and bottom caption text of the meme image (Figure 6.2). Given the extracted text and recognized meme template, we can 1) generate the meme's alternative text, and 2) generate an audio meme by using text to speech. We use Google Cloud Vision API's Optical Character Recognition (OCR) feature to detect and extract text from images. Most of the watermarks on memes (*e.g.*, "Imgur.com") appear along image boundaries but do not contribute to the main meme text. So, we remove any text with a bounding box within 5 pixels of the image border.

We evaluated our this recognition approach using the "Meme Generator Dataset" from Library of Congress's Web Archive [73] that contains 57,000 memes along with the top and bottom text. For each ground truth and prediction pair, we calculate word error rate (WER) or the number of substitutions, deletions and insertions in an edit distance alignment over the total number of words [105]. We achieve a word error rate of 22.1% and a character error rate of 9.2%. We find two common types of errors: 1) a word includes only a few mistaken characters ("OET" instead of "GET"), and 2) two words are recognized as one word ("ANDTWO" instead of "AND TWO"). When a word is not recognized, a screen reader either pronounces the word phonetically or spells out the word. In the case of combined words, the phonetic pronunciation is typically correct. We explored applying a simple spell-checker to the resulting OCR text. While it did correct many 1-character mistakes, it often incorrectly changed the combined words. We chose not to use the spell-checker, but in future work we will explore more approaches to reduce the WER, such as spell checkers with more advanced language models or OCR fine-tuned for fonts typically used in image macro memes.

6.2.2 Authoring alternative meme templates

Our authoring interface (Figure 6.5) lets users generate alternative templates including alt text templates and audio meme templates to add to the database.

[3]https://imgur.com/memegen/popular/year

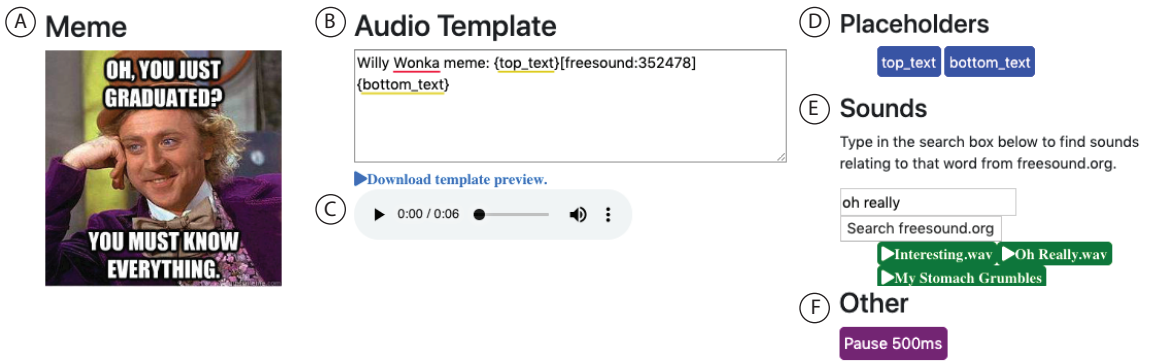


Figure 6.5: The meme template creation interface displays (A) a reference meme example, (B) the constructed meme template so far, (C) preview and output in text and audio formats, and then a series of tools to construct the meme template. To create an alt text template, a user can drag the (D) top/bottom text placeholders to the meme template box then write alt text in relation to where it should occur with the placeholders. The system then exports the template as text to be applied by the automatic method. To create an audio macro meme, a user can input placeholders then click and drag (E) sounds from a library accessed via search to place sounds in relation to the placeholders. Finally, users can optionally place (F) pauses for comedic timing.

The authoring interface accepts an input example meme (Figure 6.5A) and parses the meme using the automatic pipeline to identify the top or bottom text. To create an alt text template, a user drags the (Figure 6.5D) top/bottom text placeholders to the meme template box and writes alt text in relation to where it should occur to the placeholders. The system then exports the template as text such that the automatic method can later apply the template to new examples. To create an audio macro meme, a user can place top/bottom text placeholders then click and drag (Figure 6.5E) sounds from a library accessed via search to place sounds in relation to the placeholders. Finally users can optionally place (Figure 6.5F) pauses for comedic timing.

The authoring interface is the same for creating either alt text or audio meme templates, except that sounds and pauses are unavailable for alt text meme templates. Authoring of the meme template occurs for the general instance of that meme, so users cannot edit OCR results that will eventually fill the placeholder. However, they can preview their alt text or audio templates with an example.

Once a user has created and submitted their new alt text or audio template, it is reused for any user after a meme example is matched to that base meme template. The system currently chooses just the most recent template, but future work may involve a measure of popularity or voting to assign a default alt text or audio template to a meme.

The authoring interface itself is not currently accessible to screen readers, as it is designed to translate visual content, and also relies heavily on drag-and-drop interactions. In future work, we intend to explore accessible interfaces for designing audio-first or alt text-first memes, in addition to translating image macro memes.

ID	Age	Gender	SM years	Level of vision	Level of vision years	Screen reader
P1	41	M	12	None	10	NVDA
P2	23	M	12	Peripheral, 2 percent central	2	Voiceover, NVDA
P3	53	M	10	None	52	Voiceover, NVDA
P4	45	M	14	None	45	Voiceover, Jaws, NVDA, Narrator
P5	19	M	7	None	19	Voiceover, NVDA
P6	25	F	4.5	None	25	Jaws, NVDA
P7	32	M	12	None	32	Jaws, Voiceover, NVDA
P9	22	F	6	Low vision to total blindness (fluctuates)	19	Voiceover, NVDA, Talkback
P10	19	M	6	Light perception	19	NVDA
P11	39	F	11	None	39	Voiceover

Table 6.1: Demographics of participants who participated in the online study including age, gender, years on social media (SM years), level of vision, screen reader, and years at the designated level of vision (level of vision years). Note that P8 was unable to complete the study and is excluded here.

6.3 Meme format evaluation

We conducted a user study and interview with 10 blind or low-vision participants to understand their experiences with internet memes and compare different media formats to make them accessible. Eleven participants were recruited on the Twitter platform, and participated in our study remotely over online voice chat or phone. One participant (P8) was unable to complete the study due to issues with audio on their computer, so their data is excluded from these results. Participant ages ranged from 19 to 53, with an average age of 31.8. Three participants were female and seven were male. All participants accessed Twitter using a screen reader. All participants reported they had encountered memes before. But, due to accessibility issues with memes, only two participants reported experiencing memes in more depth: P6 reported friends explaining memes, and P9 experienced accessible memes on sites like Instagram. Further participant demographics can be found in Table 6.1.

6.3.1 Meme formats

The participants in our study were asked to interact with meme examples sourced from Imgur and Meme Generator's list of popular memes [31]. There were 9 different meme types (Appendix C.1), with 5 examples of each, for a total of 45 meme examples. The participants experienced 15 examples of these memes in the following three conditions:

1. **Text Only:** As a baseline, the simplest media format was the text-only results from an automatic OCR pass of the meme. These were HTML images that contained alternative

text of only the overlaid text. If memes have any alt text at all, it is common for it to only be the overlaid text that the meme generator automatically added. This also represents a completely automatic solution without human involvement, but the visual elements from the image are lost in these descriptions.

2. **Meme Description:** The alternative text in this condition contained a description of the visual content of the image and the overlaid text. The text was separated by the top an bottom of the image, so the participant could tell how they were visually separated.
3. **Audio Macro Memes:** Visual memes intend to provoke an emotional reaction, often some form of humor, that is lost in a pure textual description read by a screen reader. Audio macro memes, a sound analog to image macro memes, include background sound that can carry the emotional affect the meme creator intended. These were sound files that contained background audio customized to each meme type. Text-to-speech rendered the overlaid text in the meme. We hired a professional sound producer to create these audio versions, attempting to convey the emotional tone of the visual meme.

The examples we presented (Appendix C.1) represented a best case scenario in quality of meme examples. For all of these memes, we corrected the OCR results before generating each example, in order to ensure participants were evaluating the meme formats, not the OCR results. Members of the research team who were familiar with alternative text wrote the image descriptions for the alt text format. We hired a professional sound designer to create background audio for the audio memes, instead of picking from a sound effect library. In future work we would want to additionally evaluate the memes created by novice users.

6.3.2 Study procedure

Each participant completed a tutorial, listening to the same meme in each format using the screen reader or playing the audio file for the audio macro meme. Then, they were assigned an ordering of the media conditions which were balanced across participants. The meme types (see Appendix C.1) were randomized for each condition, and examples within each set of five examples were also randomized. They listened to all 5 examples of one meme type, then were asked two questions:

1. To what extent do you agree with the statement "I feel I understood the meme" where 1 is Strongly Disagree, 3 is Neutral, and 5 is Strongly Agree?
2. Please describe the meme template (*i.e.* common joke format) to us.

After answering these questions, they completed the same task for two sets of 5 more examples. After completing all 3 meme types for that format condition, they completed the other two conditions. In total, the participants experienced 45 meme examples from 9 meme types. They answered the questions above for each meme type.

6.3.3 Results

The first question posed above seeks to measure the participant's confidence in their understanding of the common joke format for 5 examples of the same meme. We present the average response for each media format by participant in Table 6.2. Participants were more confident

with alt text memes (mean = 3.95), and confidence levels for the text-only (mean = 3.55) and audio macro (mean = 3.52) media formats were similar.

ID	Text Only	Alt Text	Audio Macro	All Conditions
P1	2.67	3.67	3.33	3.22
P2	3.33	4.00	4.67	4.00
P3	4.83	3.83	3.67	4.11
P4	5.00	4.00	3.00	4.00
P5	2.33	2.33	2.83	2.50
P6	5.00	4.67	4.67	4.78
P7	1.33	3.00	1.00	1.78
P9	4.33	5.00	4.67	4.67
P10	2.33	5.00	4.00	3.78
P11	4.33	4.00	3.33	3.89
All	3.55	**3.95**	3.52	3.65

Table 6.2: The average agreement with "I feel I understood this meme." for each participant by meme format.

The second question we asked after each 5 meme examples was to measure the participants' accuracy of understanding the joke format. Three members of the research team individually wrote the target joke formats, extracting the common elements important to the joke across all of the visual meme examples. These three interpretations of the joke format were combined into a rubric for each example. Two members of the research team redundantly coded a random subset of 20 participant meme templates as either correct or incorrect, and inter-rater reliability was estimated using Cohen's kappa $= 0.7$, which can be interpreted as substantial agreement [57]. One of the team members continued to rate the remaining participant templates. Participant answers were marked correct if they partially or fully matched that meme's rubric, or if they mentioned the name of the meme directly. For example, the rubric for the Success Kid meme was "Victory/outcome/success (especially minor)", and a participant's response of "Little triumphs, little minute triumphs" was rated correct, while "Something bad and then something good." was not specific enough to the form described in the rubric and marked incorrect.

Overall, participants accurately stated 63% of the joke formats after hearing 5 examples in various media conditions. The results across conditions were close, with audio memes having an accuracy of 70%, alt text memes an accuracy of 63%, and text-only memes an accuracy of 57%. Due to the small number of participants, it was not appropriate to perform a statistical analysis on these results, but a larger follow-up study may be able to examine if there is a statistically significant difference between media formats.

6.3.4 Post-study interviews

We interviewed each participant about the memes and media formats they experienced after they finished listening to all 45 examples and answering the questions above. Here, we summarize some of their responses and the trade-offs between the different formats.

Format preferences

The overwhelming majority of participants (8 of 10) preferred the alternative text memes, primarily because it gave them access to a visual description of the content. Several participants noted that this description helped them understand the meme better, particularly if the emotions or facial expressions of the character in the meme were described. Participants often called these "characters" and believed they might be the "speaker" of the meme text. As P3 said regarding the First World Problems meme:

> *It gives you "head in hands, crying". I could get the emotion, but the reason for the emotion appears in the text. – P3*

On the other hand, many participants noted that the images were not always clearly connected with the meme template, and they were confused why it was included.

> *It's a little confusing, because I'm like "Why is a bear saying this?" or "Why is a penguin saying this?" – P6*

This sometimes lead participants to be overly specific about the joke format, such as "Ways the toddler is prevailing over life." for Success Kid, even though a meme example was parking a car, which is an activity not performed by most toddlers.

Participants raised specific concerns about the audio meme format, as it did not use the standard accessibility features (*i.e.* alternative text). This meant the participants did not hear the memes in their preferred voice and speed. Additionally, one participant noted that audio memes are not universally accessible, whereas alternative text or text only memes are available to deaf-blind users or those who use Braille displays.

The participants who preferred formats other than alt text (P6, P9) also reported the most in-depth meme experience in the pre-interview. P6 and P9 noted they found formats other than alt text to be more efficient. While P9 preferred audio memes because the audio quickly conveyed the meme tone (*e.g.*, "dark memes", "sarcasm"), P6 preferred text alone.

Willingness to share and create memes

As many of the participants had not experienced a large number of internet memes before, we asked them if they would have posted any of the 45 examples they experienced during the study. Nine of the participants had at least one they might post, but several would only do so with friends, not publicly. P9 was very enthusiastic about sharing memes in general – just not the ones we chose as examples:

> *I would probably consider posting them because they were strictly made in an accessible format, [But] my friends would think "Why are you posting things from 2011?" – P9*

Three participants said they would definitely create memes themselves if they had tools to do so.

> *I certainly want to be part of the culture. There are circumstances where I think the message I am trying to convey would be done better by visual memes than verbal or writing. It's so easy and it's so efficient to share when a picture can convey a message. – P1*

Three participants were not confident they would be able to create memes without sight, as the visual component is important. Four participants stated they were not interested in creating memes themselves, but would like to view them.

6.4 Recommendations for composing meme alt text

Our interviews and user studies with the ten Twitter users with vision impairments revealed a number of opinions and preferences about meme media formats.

Primarily, the users sought access to the same information provided to sighted users: a description of the visual image and the overlaid text. In some cases this helped the participants understand the humor or other sentiment in the meme (*e.g.*, First World Problem), although in a few cases it was confusing (*e.g.*, Confession Bear). The users stated the audio and text memes did not provide enough context to understand the meme, and this is reflected in their confidence ratings for these conditions. However, the users had similar accuracy scores for memes in these conditions, indicating there might be a divide between confidence and actual understanding of the different formats.

Some of the stated concerns with the audio memes may be due to its unfamiliarity. They were not integrated with screen readers, so they did not automatically play on focus like the alternative text. They also did not use preferred voices or speaking rates. Close integration with screen readers could alleviate these problems with audio memes, but other issues, such as lack of universal accessibility, are inherent to the media format. As the system can produce text-only, alt text, and audio memes, we can create accessible content in multi-modal formats, allowing users to select their preferred formats.

We followed established guidelines for creating meme alt text [89]. Still, our alt text did not always highlight information users needed to understand memes. Specifically, users requested more information about the character in the meme and their emotional state. In addition, several users mistook the image style of memes when reporting what they imagined the meme to look like (*e.g.*, reporting the images to be low-effort drawings or stick figures instead of photographs). Based on prior work [89] and our study results, we propose a condensed, meme specific set of structured questions for writing alt text of memes:

- Who are the character(s) in these memes?
- What actions are the characters performing, if any?
- What emotions or facial expressions do the character(s) exhibit in these examples?
- Do you recognize the source of the image (TV show, movie, etc)? If so, what is it?
- Is there anything notable, or different about the background of the image?

Meme descriptions that provide this type of context remain consistent with the fact that much of the humorous effect comes from a character acting out a scenario rather than simply describe it [48, 104]. By describing who is acting out the meme text, and what the image indicates about their background, we may be able to give viewers the intended experience.

6.4.1 Re-use of human annotations for rich media on other platforms

Memes are an interesting case of digital media on social networks, as the overall universe of content is much smaller and more constrained than the amount of content shared on platforms. While memes do not exist in a canonical set like emoji, there exist set visual templates that are reused repeatedly and may change only sporadically. Because of this, we have leveraged an approach that relies on high-quality and creative human descriptions or audio alternatives which may not be possible for all generic images on social media. This approach of re-using human-authored accessibility metadata can be replicated on other platforms, as the creation and usage imbalance is true for other mediums. As an example, augmented reality filters on various apps are created by few contributors, yet used widely. Additionally, TikTok provides an interesting example where every video may be slightly unique, but the re-used audio tracks encode a common theme similar to memes. The exact methods to make these accessible may differ, but this approach leverages the ability of humans to describe emotive qualities better than machines while still providing a scalable solution.

6.4.2 Limitations and future work

In the user study with Twitter users with vision impairments, we presented meme examples that were crafted by members of the research team. These examples represent some of the best case scenarios for each format. Word errors in the OCR results were corrected, alt text was written with best practices in mind [89], and the background audio in the audio memes were created by a professional sound designer. Online volunteers or crowd workers may not generate alternative meme templates of the same quality, although prior work demonstrates that this is true in the case of alternative text [89].

We operated from a known set of historical memes curated by Imgur and Meme Generator, but in reality new memes are always being created or modified. These examples may not exist in our database, or they may be similar enough to another meme to match, but have a different semantic meaning. Future work should explore how quickly a new meme in the wild can be recognized, and how many examples of the meme are needed before it can be transformed into an accessible format.

Internet memes are so commonly associated with visual content that most participants did not imagine audio memes beyond accessible versions of images. We believe that memes generated as audio first by people with vision impairments may be interesting as a standalone non-visual media, especially for other blind users. This may open up opportunities to explore multi-modal representations of memes and online content. In addition to static memes, participants mentioned they would like access to GIFs that are commonly posted on Twitter as reactions to tweets. Audio descriptions of GIFs could be similar to those provided for accessible videos.

6.4.3 Conclusion

Memes may not always be vehicles for conveying serious content, but they remain an important part of online discourse, whether that is public or in small groups with friends. Creators of memes typically do not include alternative text, rendering almost all of them inaccessible to

people with vision impairments. We have presented an automatic method to recognize known memes, extracting the overlaid text, and rendering that text into a more accessible format, such as alternative text or an audio meme template. Because many memes are repeated images with new text, this results in a scalable solution to make a large number of online memes accessible just by creating alternative text or audio versions of the base meme template.

In a study with 10 Twitter users with vision impairments, we found that they preferred the alternative text memes due to their inclusion of visual context, compatibility with screen readers, and universal accessibility. The study also reveals that people with vision impairments are eager to share accessible memes, as they are a part of culture and communication online. Based on their responses, we propose a short set of structured questions for alternative text authors to answer when describing memes. These can assist the authors using our system to not only make memes trivially accessible, but also preserve the emotional tone or humor embedded in the meme. Even the participants who were not as interested in "silly" memes noted that their lack of alternative text was a source of significant accessibility issues on social media.

> *I think [memes] could become a way to generate a lot of useless content very quickly. But if there has to be a lot of useless content out there, it ought to be accessible. – P4*

Chapter 7

Making GIFs Accessible

Through these prior studies of accessibility on Twitter, especially looking at the accessibility of memes, I uncovered another format that was typically inaccessible on social media: short animations known as GIFs. Unlike static images, GIFs contain action and visual indications of sound, which can be challenging to describe in alternative text descriptions. I, along with my colleagues, examined a large sample of inaccessible GIFs on Twitter to document how they are used and what visual elements they contain. In interviews with 10 blind Twitter users, I discussed what elements of GIF content should be described and their experiences with GIFs online. The participants compared alternative text descriptions with two other alternative audio formats: *(i)* the original audio from the GIF source video and *(ii)* a spoken audio description. From these interviews and my understanding of what kinds of GIFs are shared online, I recommend that social media platforms automatically include alt text descriptions for popular GIFs (as Twitter has begun to do), and content producers create audio descriptions to ensure everyone has a rich and emotive experience with GIFs online.

Work in this chapter was also published as a conference paper. The use of "we" in this chapter refers to all of the authors who contributed to that work. The full citation for that article is:

Cole Gleason, Amy Pavel, Himalini Gururaj, Kris Kitani, and Jeffrey Bigham. 2020. Making GIFs Accessible. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 24, 10 pages. 9781450371032 `http://dx.doi.org/10.1145/3373625.3417027`

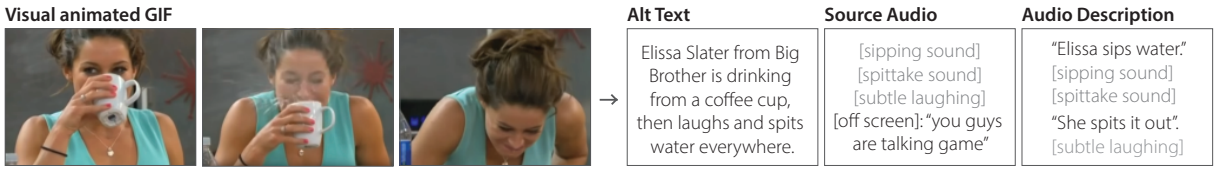


Figure 7.1: The most popular GIF we observed on Twitter was one of a spit-take. We converted GIFs like this into three alternative formats: alternative text, the source audio from the original video, and an audio description recorded over the source audio.

7.1 GIFs embody expression, but only for sighted people

While the accessibility of images online has been long discussed but infrequently addressed, the rise of animated GIFs as a method of communication has been a more recent phenomenon, and accessibility guidelines and features are just catching up to their widespread adoption by sighted people. GIFs are primarily used on social media to either embody the emotion of the poster or react to another poster's content [96]. If people with vision impairments are unable to understand the visual content of a GIF in a conversation, they miss key channels of emotional tone and information, if not derailing the conversation entirely.

The primary approach to make images accessible is via alternative text [20], which some social networks have recently begun to support for static images as discussed in Chapter 3 Twitter extended this capability to GIFs on their platform as of January 2020. However, GIFs are more than static images: the visual content over multiple frames often conveys action and contains visual elements that imply sound. Can alternative text adequately describe the emotional tone or meaning that is being visually conveyed? We collected a large sample of popular GIFs on Twitter to examine what kinds of content they contained and how they could be described.

To gather the perspective of blind people on important visual elements to describe, we interviewed 10 Twitter users with vision impairments about their prior experience encountering GIFs. In a second session, they compared three alternative formats for GIFs: *alternative text* descriptions, *original source audio* if the GIF was excerpted from a longer video, and spoken *audio descriptions* of action occurring that overlay the source audio. In both interviews, participants stressed that they viewed alternative text as a minimum accessibility requirement. However, depending on both the visual content of the GIF and the original source audio, participants suggested that some audio descriptions presented a more emotive and enjoyable experience of viewing GIFs.

In summary, our contributions are:

1. An analysis of GIF usage on Twitter, including how many have alternative text;
2. Findings from interviews with 10 Twitter users with vision impairments regarding their past experiences with GIFs; and
3. Preferences for accessible alternative formats for GIFs.

In February 2020, few GIFs (0.04%) contained alternative text on Twitter, as the ability to add alternative text to GIFs was new. Therefore, most of our participants had not experienced accessible GIFs on social media, while some participants knew that GIFs were present but undescribed. Based on their experiences with GIFs during our study, many participants were eager to have accessible GIFs on social media – with both alternative text and more expressive audio descriptions.

This work suggests that social media platforms should seek to automatically include alternative text for GIFs on their platforms. In May 2020, Twitter started to include short alternative text descriptions of GIFs taken from their titles on GIF aggregation sites (*e.g.,* GIPHY). They also made it easier for users to add alternative text in general by removing the requirement to enable a special setting, which was noted as holding back alternative text adoption (Chapter 3). Based on our investigation of important visual elements in GIFs and discussions with participants, social media platforms should create libraries of descriptive alternative text and automatically include

them when users re-use GIFs. Additionally, they should push the accessible experience further by working with content creators to develop rich audio descriptions to convey the emotion in GIFs.

7.2 A short history of GIFs online

The Graphics Interchange Format (GIF, pronounced "jif" [41]) began in 1987 as a format designed to bundle multiple images at a time for later viewing as sequential frames. But the format grew over time with the advent of the World Wide Web and acquired new features: a timed delay between images, transparent backgrounds, and automatic looping of the animation [28]. These features led to widespread use of GIFs on web sites to display animated icons, but the modern emergence of GIFs seen on social media is due to their use on the Tumblr and Reddit platforms.

Tumblr, a microblogging platform, supported GIF uploads from its inception, leading its community to share a significant number of GIFs that were excerpted from TV shows or movies [28]. Fans used these excerpt GIFs to talk about their favorite characters and moments, while spreading these out-of-context actions and dialogue (overlaid as a visual caption). Others re-used the visual context from the TV show, but added their own text to give the GIF a new meaning [44]. Reddit popularized the "reaction GIF", which contain actions or gestures (especially facial expressions) that convey an emotional reaction to a scenario. The original creators of these GIFs may have intended to convey a specific meaning, but interpretations may vary based on the separate understandings of the GIF poster and viewer, their prior knowledge of the source material, and their relationship [52]. GIFs that are shared on most social media platforms and text messaging services today resemble those that spawned on Tumblr and Reddit, and they are typically either act as a response someone else's post, or as a supplement to text posted by the author to embody an action [96].

The initial uses of GIFs on these two social media platforms demonstrate the two core abilities of GIFS: performance of affect and conveyance of cultural knowledge [68]. They are more engaging than other forms of media due to this and their technical constraints [54]. But these constraints limit GIFs as a visual-only medium, which is a disservice to people with vision impairments who will miss out on emotional tone on social media [36] and be unable to share GIFs themselves.

7.3 Audio descriptions of video content

Like GIFs, longer videos often contain visual content expressed over time. Although videos are not silent like GIFs, they often feature visual content that is inaccessible from the audio track alone. Audio descriptions are the primary method for providing viewers information about this content via a narration track overlaid on top of the video [83]. In the past decade since instating the Twenty-First Century Communications and Video Accessibility Act, audio descriptions have become increasingly common on TV and movies [76, 82], especially with the advent of streaming platforms that add audio descriptions to new content such as Netflix. Audio descriptions are challenging to produce because an author must fit all of the necessary visual content into a

limited time provided [78, 107], and are most often professionally produced.

However, audio descriptions are exceedingly rare for online user-generated content for reasons including: lack of video author awareness, challenge of crafting descriptions, and a lack of platform support. Prior work proposed methods to make audio description easier to create including using text-to-speech instead of human narration [56], creating task-specific authoring tools [13, 78, 80], offering methods to add audio descriptions on embedded YouTube videos [1, 80], and hosting audio descriptions [80]. Such tools rely on proactive video authors and third party volunteers, and are challenging to scale. We instead consider the space of GIFs, where we can leverage the resources of centralized GIF creation, and the repetition of the medium in order to make them more accessible.

In our consideration of audio descriptions for GIFs, we analyzed several audio description guidelines often written by and in collaboration with blind authors [4, 25, 74]. While such guidelines primarily offer guidance for long stories, we apply key principles (*e.g.*, describe important visual content, avoid overlapping dialog and key sounds, start general then add detail) in the case of providing audio descriptions for the extremely short medium of GIFs.

7.4 Existing usage of GIFs on Twitter

To explore how GIFs are used on Twitter and what types of content they contain, we used the Twitter API to collect a large, random sample of approximately 108 million tweets continuously from February 26 - March 13, 2020, containing 791,600 GIFs (0.7%). This sample was filtered to remove tweets that Twitter automatically tagged as containing possibly sensitive (*i.e.*, pornographic) material, deleted tweets, retweets, and non-English tweets. After filtering, 303,874 GIFs remained, and only 126 of these (0.04%) contained alternative text. However, the ability to add alternative text to GIFs was launched only 1 month prior to our sample collection, so it may not yet have widespread adoption.

In May 2020, Twitter introduced a feature that automatically included short alt text for GIFs if they were taken from GIF aggregation sites. These are the titles of the GIFs present on these sites, and Twitter added them if users shared a GIF and did not include alternative text themselves. For example, the GIF in Figure 7.1 had the title "Big Brother Elissa Slater GIF". While this title includes the name of the person in the GIF and the TV show she appeared on, it fails to describe the visual content of the GIF and the spit-take action occuring. When these titles did describe the action, such as "Oprah Shrug GIF" for Figure 7.3, it did not include much detail. Twitter also made it easier for users to add alternative text in general by removing the requirement to enable a setting before seeing the interface to add alternative text. In light of these changes, we collected a smaller sample of 31,000 GIFs in June 2020, and found 47.4% included alternative text with these automatic GIF titles. Because they follow a specific format (short titles ending in "GIF"), we estimate that almost all (99.3%) of the GIF alternative text is automatic titles. Excluding those, 0.3% of GIFs have alternative text likely added by the GIF poster. The remaining analyses in this section are not concerned with the alternative text already on Twitter, and therefore are base don the larger GIF sample.

Prior work has noted two common ways to use GIFs: to supplement your own post or to react to another post [96]. We see this behavior in our large sample as well: 23% of the GIFs

were included in original posts and 77% were in reply to other tweets. Notably, of those that were original posts, 89% contained additional text, whereas only 33% of reply GIFs accompanied text. This indicates that someone using a screen reader or Brialle display may glean some information from the text content of original tweets with GIFs, supposing the GIF is not the central element. Two-thirds of GIF replies would read as completely blank.

7.4.1 Determining unique GIFs

When online memes use repeated visual elements, it becomes easier to make them accessible as portions of alternative text can be reused between images (Chapter 6). We were interested to see if GIFs were often reused, and if so, how many unique GIFs might need to be described. We analyzed the first frame from each GIF to output a perceptual image hash [15]. To verify this method, 10 instances of 25 GIFs were manually examined to ensure they correspond to the same GIF, excluding minor changes due to compression or resolution differences. It is possible that some GIFs could be incorrectly marked as unique if they had significantly different first frames, but the likelihood of this is small as many were shared from aggregator websites and contain the same set of frames. The total number of unique GIFs that were tweeted at least once is 127,916 (42%), and the remaining GIFs were repeated. Several (187) of these GIFs exceeded 100 uses, and the remainder form a long tail of usage distribution (Figure 7.2). This suggests that accessible formats could be reused for the most popular GIFs.

7.4.2 Visual elements of GIFs

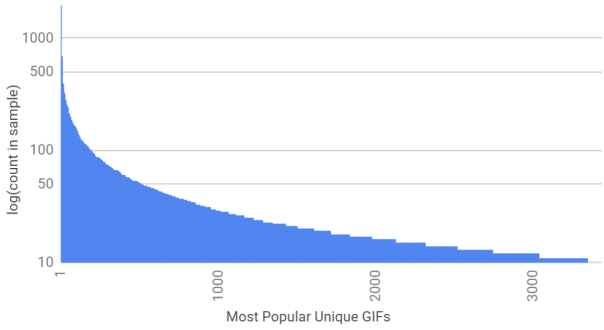


Figure 7.2: Histogram of the all of the most popular GIFs in our sample (used at least 10 times). The y-axis shows how often each unique GIF was used, on a logarithmic scale.

We randomly sampled 97 of the most popular 1,000 GIFs to understand the kind of visual content they contained. Popularity was determined by the number of unique times a GIF was shared, not retweeted or liked. To first identify the important visual elements of GIFs, two members of the research team iteratively coded three small, seperate random samples of GIFs (30 at a time) to describe them textually and add open codes. They met frequently to discuss their codes, which were based on elements relied on to textually describe the focus of this GIF (*e.g.*, number of characters, text captions, is the character performing an action) and other elements of composition that differed between GIFs (*e.g.*, live-action versus animation, shot length). Once the kinds of visual elements were agreed upon, the researches then proceeded to code the 97 popular GIFs to describe the frequency of various visual elements, which are reported below:

Original or Excerpt: 87 of the GIFs were excerpted from a longer video, while 10 seemed to be created just to share as a GIF.

Animated or Live-Action: 77 GIFS were live-action content, while 20 were animated. All 10 of the original GIFs mentioned above were animated.

How many characters?: 75 of the GIFs contained only 1 person or character, whereas 14 displayed 2 or more characters, and 8 contained none.

If there is text, is it dialogue?: 14 GIFs contained text, and 7 of these were lines of dialogue from the original source. The others were either overlaid by the GIF author or original GIFs that displayed text only.

Are there visual indications of sound?: 37 GIFs contained some visual indication of sound, with 11 being dialogue, 11 vocalizations that were not speech, and 18 sound effects (e.g., clapping). A GIF could contain more than one indication of sound.

Is the character(s) face important?: 85 of the GIFs had at least one face present, and we identified 58 of them as being important visual context (*i.e.,* the face was the focus).

Is the character performing an action?: 53 GIFs contained the character performing some action or gesture, including clapping, walking, dancing, etc.

Camera Angle Shot: 36 of the GIFs were close-up shots of a person's face, 36 were medium-length shots of someone's torso and head, and 16 were full-body shots of someone from a distance.

This analysis gives us a good understanding of the kind of visual content that might need to be described in a GIF. Most are excerpted from longer, live-action videos and contain characters. About a third contain visual indications of sound, meaning many gestures or actions may be non-verbal. In around 60% of the GIFs, a character's face is the focal point, indicating facial expressions will be critical for understanding GIF content.

7.5 Formative interviews with blind users on important visual information

This analysis of a large sample of tweets gave us insight into the quantitative nature of GIFs on Twitter, but we desired a qualitative perspective from people with vision impairments to assess the impact on accessibility. To do this, we conducted a formative study with 10 Twitter users who had a vision impairment. The participants were equally split between men and women, and they averaged 36.2 years old (min = 20, max = 52). Only one participant (P1) used their vision to access content on Twitter, but she often used a screen reader as her level of vision can fluctuate. All participants had used Twitter for at least 5 years, except P1 who used it for 3 years. More detailed demogrpahics are available in Table 7.1.

In our formative interview, we asked participants about encountering GIFs on Twitter or other social networks, showed them examples of alternative text that we wrote for 10 GIFs, and solicited their feedback on what information to include in accessible GIFs. The interview questions are available in Appendix D.1.

ID	Age	Gender	Level of vision	Level of vision years	Years on Twitter	Other social media	Screen readers
P1	23	F	Low-vision	Since age 3	3 years	Instagram, Pintrest, Facebook Messenger	NVDA, VoiceOver, Select to Speak
P2	25	F	Light perception	Since birth	5.5 years	Facebook	VoiceOver, NVDA, JAWS
P3	39	F	Light perception	Since birth	12 years	Facebook	VoiceOver
P4	33	M	Totally blind	Since birth	13 years	None	NVDA, JAWS
P5	41	M	Totally blind	Since age 26	13 years	Facebook, Instagram	NVDA, VoiceOver, Talkback
P6	54	M	Totally blind	Since age 1	11 years	None	NVDA, Voiceover
P7	46	M	Totally blind	Since birth	13 years	Facebook, LinkedIn	JAWS, NVDA, Narrator, VoiceOVer
P8	29	F	Totally blind	Since age 17	7 years	Facebook, LinkedIn	JAWS, NVDA, VoiceOVer
P9	20	M	Light perception	Since birth	7 years	Facebook, Youtube	NVDA, VoiceOver
P10	52	F	Totally blind	Since birth	10 years	None	JAWS, NVDA, Narrator

Table 7.1: The demographics of the participants who engaged in both online interviews, including age, gender, level of vision, years at the designated level of vision, years using Twitter, other social networks used, and screen reader software used.

7.5.1 Prior experience with GIFs

We asked our participants about their prior experiences encountering GIFs on social media or the web as a whole. Five of the participants stated they commonly encounter GIFs online, and the others either saw them sporadically or not at all. Three participants used the TWBlue client to access Twitter [22] which does not notify the user when they encounter a tweet with a GIF included, so those three participants were not very aware of GIFs. Two participants who frequently encountered GIFs noted that it was typically in replies to other tweets or in comments for posts on Facebook. Five participants stated that when they encounter GIFs, they are not sure if they are missing content that is important to the conversation, while five participants stated they mostly ignore GIFs because they are inaccessible. Four participants had experiences where the use of inaccessible GIFs interrupted conversation, with P1 relating how it interrupted an interpersonal relationship:

> About three years ago I was talking to this guy who only reacted in reaction GIFs, and I could never tell what emotion they were feeling about a particular question. [. . .] I think he assumed that there was a lot more accessibility available for GIFs than there actually was. Because I couldn't see almost anything he was sending me and I

ended up just like, 'You know what? We're done. We're not talking.' – P1

The participants stated that they did not usually share GIFs because interfaces to select GIFs on their mobile phones or social network applications did not provide enough information about the GIF to choose one.

In terms of workarounds, 4 participants explicitly stated they used the surrounding textual content, if available, to guess at what a GIF might contain. P1 was the only participant who reported using external software, such as Microsoft Seeing AI [64], to describe GIFs. Four other participants said it was too much work, as the GIF might not be very interesting and they must take a screenshot to extract a single frame from the GIF to get a description. Of course, this is unlikely to fully describe a GIF, as they contain action over multiple frames. P1 recounted this:

> My brother [said] "Hey, watch, this garden hose turned into a snake!" So we had to do it frame by frame so I could figure out what was going on. – P1

Three participants said friends would verbally describe GIFs they wanted to share in person, or send text descriptions in online messages, but this was infrequent. Six participants had seen people online describe a GIF posted by someone else at least once, but P5 noted that asking others to describe this content either in person or online has high social barriers:

> Oh, you know, I don't want to wear out my welcome. It's a socially awkward. But at the same time, I feel like I need some access to that culture. – P5

7.5.2 Information to include in GIFs

To elicit feedback on what information to include in GIFs, we prepared alt text for 10 GIFs and read each to the participants during the formative interview. The GIFs were selected by sampling 100 random GIFs and manually choosing 10 that roughly spanned the visual elements presented in Section 7.4.2. This formed a diverse sample to elicit discussion about important information.

After each GIF, we asked what elements of the alt text participants thought were important and which they they might remove. We attempted to include a lot of information in the alternative text descriptions, so that participants were aware of the majority of the visual elements. The alt text and GIFs are available as supplemental material.

All participants noted that the most important elements of the GIF descriptions were: the people or characters present and what they actions they are taking. If there was not a definite character or person in the GIF, then the focus should be described. All participants wanted to know what text said, if it was present. When text is present, care should be taken to distinguish if it is dialogue from the GIF source video or not. One GIF was a clip from Saturday Night Live with unrelated text overlaid, similar to an image macro meme [27], and participants were unsure if the text was dialogue from the SNL skit.

If a GIF was taken from a movie or TV show, participants wanted to know information about the character, actor/actress, and the work they appeared in. There was some disagreement between participants about which of these three was important to include. Three participants thought the character was most important as the action or dialogue might be more closely linked with the character. Others mentioned that different actors can play the same character (as in a GIF for The Batman), and sighted people viewing GIFs may recognize the actor or actress even if they never saw the film or show. P8 suggested:

Figure 7.3: A popular reaction GIF of Oprah Winfrey shrugging. She turns to look to the camera, glances to the side, stares at the camera, then shrugs with her palms up.

> So you've got 'Princess Diaries', 'Princess Mia', and 'Anne Hathaway', right? Having two out of those three I think is probably good. – P8

Participants wanted most of the information to be present, but also alternative text to be concise. When alternative text mentioned the clothing of the character or person in the GIF, most participants were not personally interested but were reluctant to suggest removal in case others may be interested. Some participants already knew pieces of information in GIFs (*e.g.*, P2 and P8 were aware Michael Jordan played for the Chicago Bulls), but thought others might benefit from it. The only information that the majority of participants felt comfortable suggesting to remove was information about overall GIF coloring such as "It is very dark and red" or redundant information that appeared elsewhere in the description. In one case, the alternative text described Michael Jordan performing a "reverse one-handed dunk" and included a more lengthy description of the same action. Participants wanted one or the other to make the GIF more concise. Four participants stated that length was not their primary concern, and that the description needs to be proportional to the amount of action occurring:

> It's long, unfortunately. I know you want to keep these brief, but I think sometimes for sake of being complete, it just takes as long as it takes. – P10

7.5.3 Stated preferences for accessible formats

Both before and after hearing example alternative text for GIFs, we asked participants about their thoughts on how to make GIFs accessible. Before hearing the alt text descriptions, almost all participants suggested that GIFs be accompanied by alternative text on sites like Twitter. Specifically, participants wanted Twitter to make it easier to add alternative text to GIFs on mobile devices and make users more aware of alternative text. P5 suggested that alternative text be turned on by default for everyone, something that has been suggested in Chapter 3. Three participants wondered if GIFs could be automatically captioned as they were used to from applications like Microsoft Seeing AI. P1 wanted human-authored descriptions to be added to all of the GIFs that Twitter and others offer in their GIF libraries:

> Just put alt text in across all the GIF libraries, because I feel like other users aren't going to take the time to know what alt text is or how to write it. – P1

After experiencing the alt text descriptions for 10 GIFs and recognizing that many were extracted from other videos, seven participants brought up audio formats as another possibility. Two participants suggested that the source audio by itself would not have enough context, but five suggested that audio descriptions could be recorded or extracted from the original video if

it was described. However, all participants were confident they still wanted alternative text for GIFs as a minimum accessibility requirement. Alt text is quicker and less disruptive as it can be read in the screen reader's voice and speed. It is also universally accessible to people who browse social media with a Braille display. P10 said she sometimes struggles to hear audio descriptions over background noise and music. So these participants noted that they would like to have source audio and preferably audio descriptions if available, but that alternative text always needs to be there to fall back on situtationaly or for more context.

7.6 User perceptions of alternative GIF formats

Based on the formative interviews with participants, we developed some sample accessible alternatives for GIFs and asked participants to examine them in a second 30-minute session as a means to understand their perceptions of these alternative formats.

7.6.1 Materials

We determined that there were three likely formats for alternative representation of GIFs that could be more inclusive: alternative text, original source audio, and audio descriptions. Alternative text drew on existing best practices for describing images online and audio descriptions were based on best practices for accessible movies or TV shows. We also experimented with only the source audio, bringing in the audio context from the original source material if the GIF was excerpted from other media.

From our prior sample of 97 popular GIFs, we chose a representative 15 (Figure 7.4) that covered different aspects of their composition (*e.g.*, facial expressions, action, source material). 13 were excerpted from longer videos, and two contained dialogue with text. One had additional text overlaid, and another was just a GIF of text. The chosen GIFs, alternative text, and audio files for the below alternative formats are all included in the supplemental material for this article.

Alternative text

Alternative text was a natural choice for an accessible alternative format for GIFs, as it is the existing standard for making images accessible online, and GIFs on the web and social media may already include alternative text (although this is uncommon on most social media sites). Most of our participants would prefer alternative text descriptions to make GIFs accessible as a minimum requirement, and expect sites like Twitter to support their inclusion. We composed alternative text descriptions for all 15 of the popular GIFs we selected. Based on prior conversations with participants, we ensured the GIF described the person or characters, actions occurring, and setting of the GIF (if important). If the GIF was from a known television property (which many were), we varied which descriptions included the character's first name, last name, and TV show name, as a way to provoke more discussion on the topic. Our composed alternative text averaged 15.9 words (min = 10, max = 20). An example for Figure 7.3 is "Oprah Winfrey turns to look straight at the camera, shifts her eyes sideways and then back to center, then shrugs."

Source audio

For GIFs that are excerpted from TV shows, we hypothesized that some GIFs could be accessible with the source audio alone, as if a video clip had been shared instead of the GIF. To evaluate this, we found the original source audio for as many of the 15 GIFs as possible. Two of the GIFs were not excerpted from a video, and we were unable to find the source audio for another three GIFs, as they did not contain enough identifying information or the video had since been deleted. We trimmed the recovered audio for the remaining 10 GIFs to be representative of the visual content. However, some of the source audio has additional dialogue that was not in the original visual GIF. Our source audio files were on average 5.0 seconds long (min = 2.0, max = 9.7). An example for Figure 7.3 is audio of someone talking off screen, saying "I always look back at that and say, you know, when I feel like I'm hungry"

Audio description

Finally, our conversations with participants revealed that audio descriptions might be a viable way to make GIFs accessible, as it is a common method to describe longer videos. GIFs, as a sequence of frames, are a format somewhat in between a static image and a video. Therefore, as audio descriptions often describe action and accompany sound, we developed short audio descriptions for each GIF with source audio. One drawback with audio descriptions is that there is often very little space to add the description audio between music, sound effects, and other dialogue in the original video. We did not attempt to ensure that the entirety of the alternative text fit into the audio description, and instead focused on brevity and conveying the most important information according to audio description guidelines [19]. We also sometimes extended the amount of source audio to allow the audio descriptions to fit, but we were careful to ensure this did not give additional context that was outside the scope of the original visual GIF. Our audio descriptions for the 10 GIFs with source audio averaged 7.9 words (min = 3.0, max = 16.0) and 5.4 seconds (min = 2.0, max = 9.7). An example for Figure 7.3 is a narration track over the original audio with the script "Oprah looks at us, to the side, and back at us, shrugging with her palms up."

7.6.2 Procedure

All of our participants from the formative interview returned for a second 30-minute session in which they listened to the alternative formats for the 15 GIFs. Participants were engaged over an online voice call using Zoom, and they were compensated $20 via an Amazon or Paypal gift card.

Because of discussions in the formative interview about how alternative text was a critical minimum requirement for accessibility, all participants heard the formats in the order of: alt text, source audio, and audio description. After hearing all available formats for a GIF example, a member of the research team asked the following questions:

1. How would you (or someone else) use that GIF on social media?
2. (If multiple formats:) Which format did you prefer and why?

The first question ensured the participant felt confident in the meaning of the GIF, and that the understood meaning from the accessible alternative was similar to the meaning interpreted visually. The second question on format preference elicited whether a particular format excelled or failed for a specific GIF, as the content in the GIF or source audio affected which format participants preferred. After listening to all examples, participants answered questions (listed in Appendix D.2) about their overall format preferences for GIFs .

7.6.3 Study scope and limitations

The purpose of the second session was for the participants to experience the source audio and audio description formats alongside the format they heard in the formative interview (alt text). This would help them compare the formats and provide qualitative feedback about the preferred format and included information.

Our participants highlighted in the formative interviews that alternative text was critical, so we chose to explicitly highlight the comparison in the second session as a preference, not a mutually exclusive choice. Because of this, we did not randomize the ordering of the formats, as someone listening to the formats would likely always here alternative text first. Therefore, we do not make statistical claims about the stated preferences of the participants. As participants only heard 15 GIFs, it is possible that participants might develop different preferences with exposure to more GIFs of different content.

As the same 10 participants were present in both the formative interviews and evaluation of alternative representations, our findings cannot represent all users with vision impairments. Longer term evaluations with larger cohorts may be necessary to solidify or confirm these results.

7.6.4 Findings

Six participants were confident in the meaning of all but 1-2 of the 15 total GIFs, and their descriptions were similar to a visual interpretation of the same GIF. Three participants reported that they were unsure how to use at least 3 of the GIFs, often the GIFs with the least context present in the source audio or unclear visual expressions. P4 was unsure how to use 8 of the 15 GIFs or what they meant. These are reported by GIF in Table 7.2.

The GIFs that presented the most confusion sometimes had sound that could be interpreted multiple ways or was hard to discern, such as the spit-take clip from Big Brother (Figure 7.1). In the source audio for this clip, another contestant starts talking right after the on-screen Elissa Slater performs a spit-take. Participants were confused who was speaking, and if the spit-take was meant to imply laughing or indignation. A GIF of Oprah shrugging (Figure 7.3) was confusing to participants because the action was entirely visual, yet another woman is speaking in the source audio, leading to additional context that is not important to the visual GIF.

Subtle character actions proved difficult to describe. A GIF of the character Stringer Bell from the show The Wire involves subtle facial expressions like a “side eye”. Participants were not sure what this gesture implied. P8 suggested that nonverbal gestures that are well-known may require the description author to editorialize more to describe the meaning.

Figure 7.4: The first frame of all 15 GIFs we used in our second session. Their source is annotated below each GIF.

Format preference

Overall, six of the 10 participants stated they preferred the audio description format as the best way to experience GIFs, with the caveat that most participants expected alternative text to be present as a fallback option if the audio description was hard to understand or they were not able to listen to audio files at the moment. Three participants preferred to use alt text, and P9 preferred to use a combination of the alternative text and source audio to understand the GIF content.

Source audio by itself was viewed as the most inaccessible format, as it often did not describe the action in the scene or was too noisy to pick apart distinct sounds in the clip due to background music, dialogue, or laugh tracks. For example, the audio for a GIF of character Stringer Bell from The Wire had a mostly silent audio track, as he sits in silence while the GIF focuses on his expression. Eight of the 10 participants stated source audio was their least favorite format, while P1 and P9 disliked audio descriptions:

> I don't like the audio descriptions because at that point I would have already looked at the alt text to know what was going to happen. So I would be more paying attention to the [source] audio. – P9

Seeking out GIF conversations

All participants said they were unlikely to specifically seek out conversations that contained accessible GIFs, but most would be more engaged when they encountered them their existing social media accounts. P8 noted:

> One of the biggest bummers is if I'm reading through social media and [. . .] the post is accessible, and then I'm reading the comments and it'll say like, "comment with a GIF". I'm like, "Damn, that really sucks". – P8

GIF Source	Understood	AT*	SA	AD
Spongebob	9	4 (+2)	1	5
Big Bang Theory	9	5 (+2)	2	3
Judge Judy	10	N/A	N/A	N/A
The Office - No!	8	1 (+2)	4	5
Brooklyn 99	10	1 (+2)	3	6
The Wire	5	9 (+1)	1	0
Utah Jazz	5	N/A	N/A	N/A
Big Brother	6	4 (+2)	2	4
Full House	9	4 (+2)	1	5
Original GIF (Text)	9	N/A	N/A	N/A
Obama's Address	10	4 (+2)	3	3
Ryan Gosling	8	N/A	N/A	N/A
The Office - Party	10	2 (+2)	1	7
Original GIF (Cats)	10	N/A	N/A	N/A
Oprah's Next Chapter	5	4 (+2)	3	3

Table 7.2: Participant understanding and format preference for each GIF. From left to right the columns are: GIF source (Figure 7.4), the number of participants who understood that GIF, the number who prefered Alt Text (AT), Source Audio (SA), and Audio Descriptions (AD). Note: * P5 and P9 always responded that they would prefer to use alt text in combination with other formats. Their preference for alt text is represented by the (+X) notation in this column, and they are also counted among the other format they preferred.

While we focused the conversation on large social media networks, P2 and P3 both mentioned they would engage more with content posted on their workplace communication platforms (*i.e.,* Microsoft Teams and Slack) as GIFs are common there. P5 wondered if means of making GIFs accessible could be extended to short videos on Instagram or TikTok, as they were interested in trying out those platforms that remain mostly inaccessible non-visually. P8 echoed this about TikTok more negatively:

> That whole app is not even accessible. I've given up on trying new social medias. – P8

7.7 Recommendations for deploying accessible GIFs at scale

In both sessions, our participants made it clear that alternative text must always be available for GIFs on social media. Alt text is what people are familiar with on the web, it works well with screen reader software, and can be customized to be read in a preferred voice or speed. It does not vary in volume, and can be skimmed quickly, as well as being universally accessible to a user with a Braille display. The kinds of visual information present in GIFs that is needed to write alternative text is similar to that of images, although a user must also describe action occurring over time. While a still image might be described as "Oprah shrugs", the GIF in

Figure 7.3 may include several distinct actions to describe such as "Oprah turning to look at the camera, shrugging with her palms up in the air, giving a sly smile, and turning back to the speaker". Participants reported a tension between an "objective" account of visual action versus a shorter description that gives a subjective interpretation like "Oprah shrugs as if to say 'I told you so"'. They discussed that the description length should be proportional to the amount of action occurring, and this tension is more clear when listening to audio descriptions, where space is very limited.

Once alt text is present for GIFs on social networks, the majority of participants were interested in additional modalities to describe GIFs, as the audio descriptions or (in 2 cases) source audio can give a more rich, emotive experience. Just as sighted people utilize GIFs to embody actions or expressions in supplement to text, people with vision impairments should have that option with audio GIFs. A caveat here is that the original GIF author may not have considered the audio content when designing the visual GIF. Thus, the source audio could be useless (with purely visual actions like a shrug) or be discordant with the visual meaning (*e.g.,* Big Brother contestant talking over the spit-take). Not all GIFs may benefit from the inclusion of source audio alone or with an audio description, but those that are centered on dialogue or vocalizations would benefit from these additional formats. Additionally, our participants disagreed about the information that should be included in the audio descriptions, as they all heard the longer alternative text and knew what was excluded from the briefer audio description. For now, we would recommend audio description best practices to decide what information to include, but future research could explore modular audio descriptions that allow people with vision impairments to choose what information is most important to them. As alternative text should always be present and contain all of the information, this decision is not as important as it is for longer media where audio descriptions are the only accessible format.

When talking about making images accessible on social media, research largely focuses on automatic solutions to scale the problem [61, 91, 112] or human-written descriptions. Automatic approaches can scale human-written descriptions for viral memes that changes, as long as the visual content remains the same [36]. Like image memes, GIFs are often used repeatedly online, so information about the origin of a GIF may help convey meaning. Current efforts to document a meme's origin and spread on sites like Know Your Meme rarely describe the visual content, as they assume a sighted audience. Future work may investigate integrating this information along descriptions of the visual content of GIFs to better convey their thematic meaning.

Recent GIFs seem less likely to be modified and remixed compared to memes, as many excerpted from TV shows are produced and distributed by the television networks [92]. If TV production and network companies are producing this content, they could make it accessible before distributing it to GIF collection website or keyboard applications. In fact, content produced for broadcast TV may already have produced audio descriptions that are sufficient for the excerpted GIFs, depending on the script. For user-generated GIFs that are not made accessible by their creator, third-party volunteers or crowd workers could generate alt text or audio description templates similar to the proposed solution for memes.

This work has primarily focused on the consumption of GIFs on social media posts, but half of participants said they would like to share GIFs if accessible formats were available. The addition of alternative text or audio descriptions to GIF keyboards and GIF search engines would aid people with vision impairments in selecting the perfect GIF. Further research may need to

explore accessible tooling to assist people with vision impairments in the creation of new GIFs, such as excerpting video clips. We focused on GIFs specifically, as these were common on social networks like Twitter, Facebook, or Reddit. But one participant mentioned they would like to see an extension of this work to short videos, such as those popularized on Vine or Tik Tok. As those videos prominently feature audio, audio descriptions seem like a promising solution, but may need additional tooling to support creation by all social media users.

7.7.1 Expressive media libraries on other platforms

The study of animated GIFs broadens the overall work of this thesis from primarily static image content to short animations that contain possibly other forms of visual information. Our findings show that some forms of visual information seem to be important regardless of the medium, such as included textual information and the importance of describing people that are the focus. However, the animated nature of GIFs has highlighted the need to describe action occurring, and it is possible that GIFs will require much more description compared to static images. Like memes, GIFs are often used to convey an expressive tone, and they are often re-used, showing that the combination of human annotations and automatic application may be appropriate again. Other platforms should take inventory of the forms of expressive media they provide their users, and determine how they might make libraries of this content accessible from the outset. As Twitter has tried to provide standard descriptions for GIFs from aggregation websites, they realize that they can address this problem differently than describing every other piece of user-generated content on their platform. Similar opportunities will exist on other platforms where creation effort is high but reuse is easy. This analysis of GIFs also focuses mostly on the consumption of GIFs, but it is clear that the lack of accessible alternatives is limiting the expressively of people with disabilities, as they struggle to choose the appropriate GIF to share. Therefore, other social media sites should consider how to encourage not only accessible consumption but also sharing and creating of digital media.

7.7.2 Conclusion

GIFs are a common and expressive way to display emotional reactions or embody physical actions on social media. In the words of P10, they are "supporting actors" for posts, but as a visual medium, they are inaccessible to people with vision impairments. In this work, we examined just how prevalent GIFs are, and how many were made accessible by GIF posters (0.04%-0.3%) on Twitter. In formative interviews with blind participants, we discussed prior accessibility issues with GIFs online, leading to the development of three accessible alternative formats. This lead to a second interview with probes of the accessible alternative formats. Our participants stressed the importance of alt text as a minimum requirement, but also enjoyed the expressiveness of audio descriptions when it fit the GIF well. We recommend that platforms continue efforts to include alternative text with GIFs on their site, and consider more expressive formats, such as audio descriptions, for the most popular GIFs.

Chapter 8

The Future of Accessible Social Media

In this work, I have laid out the state of accessibility challenges on social media for people with vision impairments, automated approaches to address the lack of high-quality image descriptions, and richer accessible alternatives for native social media content like memes and GIFs. My focus has primarily been on the Twitter platform, which happens to have many blind users that are being gradually excluded as visual media has and will continue to become more prevalent [72]. The state of accessibility varies slightly on other platforms: LinkedIn and Facebook both have automatically generated captions that are met with mixed reviews [61, 112], while Reddit and others still have not implemented basic features to add alternative text to images. This work has also focused primarily on accessible images and animations for people with vision impairments, but accessibility struggles are mirrored for other media formats and user groups with disabilities. Regardless of the specific platforms and media formats, there is certainly stark problem of access on social media, and we must develop better philosophies for building technology platforms that can support and encourage inclusion for everyone with a disability.

I would now like to discuss the general approaches to take make digital content accessible on social media platforms: understanding the divide between content and accessibility knowledge and choosing between support for accessible content created by end users and retrofitting inaccessible content.

8.1 Access knowledge vs. content knowledge

The necessary knowledge and related skills to make a piece of content accessible can be thought of as mostly content knowledge or mostly accessibility knowledge. I define these as:

Content Knowledge Information about the contents of a piece of media, such as the things, actions, or places in an image. This can include specific information like names of people or additional context such as when the photo was taken or what happened right before the photo was captured.

Access Knowledge Understanding the information that is most important to describe for someone with a disability, such as whether to describe the image style (“a photograph”) or whether to go into intricate detail. For animated GIFs, access knowledge might include information on what an audio description is and what should be included in the voice over

Content Type	Access Knowledge	Content Knowledge
Photograph	Knowing the important elements to describe depending on image context and how to write a concise description.	Who or what is in this image? Is there a setting, action, or text to describe?
Meme	Is the visual aspect important to describe? Will screen reader users recognize this meme by name?	What emotion or tone does this convey? What is the structural template of this meme?
Animated GIF	What format should this be in: audio description or alternative text?	What is this GIF from? Where can I find the source audio? Who is in this GIF?
Audio Snippet	Proper transcription practices, including timings if that is relevant.	Understanding the language, accents, names of speakers, and vocabulary of the recording.
Video	Knowing what elements to include in an audio description and finding spaces to insert narration well.	Understanding what is in the video and surrounding context, such as where it was filmed or who is in it.

Table 8.1: Examples of access knowledge versus content knowledge in accessible media creation.

of 3-second audio clip.

To create truly accessible content, whether that is image descriptions or video captions, one must have adequate levels of both content knowledge and accessibility knowledge. Accessibility professionals, such as those that record audio descriptions for movies, have the expertise in accessibility knowledge and acquire content knowledge from clients, a video script, or their own research. Typical end-users who have yet to write image descriptions have plenty of content-knowledge of the photos they are about to post, as they likely know the most about their intent and the visual contents, but they lack the accessibility knowledge gained through understanding guidelines or lived experiences.

Platforms should pursue efforts to bring pools of content and access knowledge together to create quality accessible digital media. But what methods should they pursue?

8.2 Should platforms invest in access or accommodations?

There is an interest from social media platforms to invest in automated technologies that, once deployed, will provide image captions (Facebook, Instagram, and LinkedIn) or video captions (YouTube) to all content on their platform. Unfortunately, this does not solve the accessibility problem, as these methods are often inaccurate and may be over-trusted by people with disabilities [61]. While these issues stem from the issues of accuracy and robustness, it also comes from a lack of both accessibility and content knowledge in these systems. As these systems are rarely

trained specifically for accessibility use cases (e.g., the Microsoft COCO captioning dataset does not contain captions validated by people with vision impairments [97]) they cannot encode accessibility knowledge without fine-tuning or additional engineering. If these systems were designed with accessibility use-cases in mind and were impeccably accurate, they would improve but still lack the content knowledge that the human photographer or content author holds. We must recognize that automated approaches to accessibility are scalable and useful, and yet most of them will remain *accommodations* for people with disability.

Across accessibility and disability research in both the digital or built environment, we make the distinction between "accessibility" and an "accommodation." To quote one possible definition of accommodation [42]:

> Accommodation means that some aspect of a system—for example a document or facility—has been adapted or modified to meet the needs of a specific individual or group. Accommodations are patches or fixes, applied retroactively to overcome barriers in the environment or system.

These automated systems are an accommodation, a retrofitting of existing infrastructure to have the appearance of accessibility. Just as some buildings retrofitted with a wheelchair ramp may still be inaccessible inside, digital content retrofitted with automated image captioning models will be unable to present an accessible experience for all content. However, they are necessary to address the wide swath of existing inaccessible content and users who are unfamiliar or unwilling to create accessible content.

Instead, we must imagine what a platform might do to create an accessible experience for content on their platform, rather than an accommodation. The first is to make it possible for accessible content to be created by users who already have the requisite access knowledge. While sites like Twitter support this for image descriptions, other platforms like Reddit lack support for image descriptions and YouTube lacks audio descriptions. Users are able to find workarounds to create accessible content if they desire, but the lack of structural support inhibits what could be wider adoption.

Twitter has implemented the structural support necessary for image descriptions, but users lack the access knowledge to find this feature and create accessible content. Instead platforms must seek to further support and encourage end-users to acquire and utilize access knowledge. One possible approach is a system like HelpMeDescribe, which seeks to train users in creating accessible image content. Another might be systems that reduce the monotony of creating accessible content, such as optical character recognition to recognize long text passages or automated speech recognition to produce a rough draft of captions, as Youtube provides for their videos. Finally, platforms can seek to change the incentives for end users to produce accessible content. Some participants in my studies have suggested that image descriptions be required before posting, for example, but social media platforms could also attempt to remind users to add an image description before posting. Alternatively, positive incentives seem to have aligned to increase the amount of captioned videos on sites like Facebook [30], seemingly because so many viewers are situationally impaired and browse withvideo sound muted. Just as website developers are encouraged to build accessible websites for better search engine ranking [70], social media content could be ranked higher in algorithmic news feeds if platforms knew it was more inclusive of users with disabilities.

Overall, platforms must invest in both accommodation approaches that scale quickly to remedy a rapidly growing problem and build experiences that train and encourage content authors to become accessible content authors. The relative prioritization of the two philosophies will likely depend on the specific type of content common on the platform and the types of investments their content creators are already making in digital media production. YouTube creators are aware of the inadequate nature of automatic captions [77], so further tooling the help them edit and refine them to high-quality standards would be beneficial. Twitter, on the other hand, might be best served by integrating automatic optical character recognition to recognize text in screenshots commonly shared on their site. In the end, this is a large and longstanding accessibility problem and no one approach is likely to solve it alone.

8.3 Accessibility of future social media

We can explore how these philosophies of retrofitting and designing accessible-first experiences might apply beyond photographs into the future of social media platforms. As an example, some platforms like Facebook are popularizing 360-degree photos and photos that contain depth information. Sighted users can pan their smartphones in the physical world to examine different parts of the photos. The alternative text models we have currently designed for images with a fixed field of view may suffice to describe these new forms of images, but if sighted people are exploring new interactions and more visual information, could we create better non-visual experiences as well? My work with memes and GIFs suggests that we start with simple alternative text as a minimum accessibility requirement, but explore better non-visual interactions to convey the same sort of experience that sighted people are communicating when they share these photos.

While much of this work examines static images and the slightly-more dynamic GIFs, videos are already prominently featured on many platforms. My conversations with participants with vision impairments over the course of this work has indicated that these are a source of accessibility issues as well, although the presence of audio often mitigates this accessibility issue slightly, depending on the contents of the video. Similar to image descriptions, videos might be produced with more audio descriptions if platforms have explicit support for it and provide tooling that help novice users acquire the access knowledge needed to create good audio descriptions [78]. Retrofitting of inaccessible videos may only be possible with the often-inaccurate automated video descriptions.

Future social media platforms are already moving on to even richer mediums including augmented reality filters and overlays (e.g., Snapchat and TikTok), 360-degree videos (Facebook), and virtual avatars (Bitmoji, Memoji, and Facebook Avatar). We can also see that platforms intend to expand into virtual reality for social interactions online, such as the Facebook Horizon project. Because the number of people currently creating these experiences is low, the time is ripe to push for the inclusion of accessible interactions and alternative formats from the start. The Canetroller project [114] demonstrates how interaction might be enabled non-visaully with a different input device for white-cane users. Retrofitting virtual reality experiences has also been recently explored to adapt these environments for people with low vision [115]. Virtual reality developers may pursue integrating these approaches deeply into the experiences they create while platforms seek to encourage this behavior and provide retrofitting accommodations for

those that do not.

8.4 Shared responsibility for accessible technology platforms

As I mentioned in the prior chapter, social media platforms need to pursue dual investment in accessible experiences at the point of creation by content authors and retrofit the content on their platforms to be accessible by people with disabilities. They must do this because they have a *shared responsibility* for the content that they enable to be uploaded and shared. Just as we might expect sites like Facebook or Twitter to have some role in the moderation of harassment or graphic violence on their platform, we should expect them to ensure that content is not actively excluding the participation of people with disabilities. The content author should, ultimately, be the one to ensure the piece of media they create is accessible, as they have the most content knowledge. But platforms must enable and encourage this behavior. This understanding of shared responsibility can be broadened to other relationships between technology platforms and the production of user generated content.

8.4.1 Who is the "user" in "user-generated" content?

On social media platforms it seems clear that the user is anyone who logs on to a platform to view or share content. But there are, in fact, many different subgroups that may primarily post media, consume media, create media for others to share, create advertisements, or build new applications around the platform. All of these user-groups have relationships of power and dependency. People with vision impairments rely on content creators to make their content accessible who in turn rely on the platform to enable this behavior. Users who use a third-party application, such as TWBlue for Twitter [22]. may rely on the application developers to surface accessibility information, who in turn rely on Twitter to expose that information in the Application Programming Interface (API). They may also depend on the developers that create software they depend or build on top of, such as desktop/smartphone operations systems or web browsers.

The content creator uploading a photo or video may hold the ultimate responsibility to make their piece of content accessible, but every other actor in this chain of dependencies is acting as the "end-user" to some other platform they depend on. Therefore, they must all share some responsibility in ensuring that the ultimate consumers of the user experience, including people with disabilities, have an accessible experience. Sometimes, platforms attempt to support just their direct end-users, for example smartphone operating system platforms encouraging application developers to ensure their images are labelled for screen reader users. To truly fulfill this responsibility to the ultimate user experience consumers, however, sometimes requires retrofitting to a further extent. As an example, Google Chrome now has built-in image captioning for all images that users may encounter in their web browser. Similarly, Apple has introduced image captioning on their iOS devices, retrofitting images added by both application developers or social media content creators that lack accessibility information. The accessibility responsibility between Google Chrome/Apple and blind people consuming content is now more clear, even if it is still mediated by the website or application developer.

This shared responsibility could be expanded beyond the direct dependency relationship between technology platforms and their direct or indirect users to include other community members. Takagi et al. explored the community coming together to retrofit inaccessible websites with Shared Web Accessibility [93]. Similar approaches have been explored for users to repair inaccessible user interfaces for smartphone applications and share their efforts among the community [113]. Finally, Brady et al. has explored the idea of asking social media users, both friend and strangers, to volunteer their time to describe images lacking alternative text [11, 12]. Platforms could embrace this idea of community-sourcing to make content accessible on their platforms, especially if they could match volunteers and community members to leverage the most content or accessibility knowledge. As an example, assigning followers of popular accounts to describe content they are most familiar with as fans or colleagues. These approaches of shared accessibility broaden the responsibility of maintaining an accessible online space among a larger portion of the community, instead of placing the burden primarily on people with disabilities to advocate and content creators to produce accessible content.

Chapter 9

Conclusion

Social media platforms are now an important space for both public and private communication, and people with disabilities deserve access to these online spaces. The deluge of inaccessible visual content on these platforms, from images and videos to augmented reality interactions, is excluding people with vision impairments from full participation. Social media platforms, in conjunction with accessibility advocates and researchers, must design technology to reduce this vast accessibility gap.

My work has shown that enabling users to add accessible alternatives, such as image descriptions, does not solve this problem as very few people know how to enable or use these features. On Twitter, no matter what subgroup of users is examined, alternative text is present in only a few percentage points of image content on the platform (0.1% of a random sample). Even the accounts that we might expect to be best equipped, celebrities and politicians, fail to add image descriptions to the vast majority of their tweets. This may keep users with disabilities from participating in everything from the latest viral funny moment online to a critical safety update about an ongoing pandemic.

Instead, we must design technology to help users create accessible content easily, encourage widespread adoption, and retrofit inaccessible content when necessary. Social media platforms can make accessibility features more prominent and explicitly promote them, encouraging content creators to add accessible alternatives like image descriptions, instead of hiding the feature in various sub-menus. Once users notice and attempt to create accessible content, tools like HelpMeDescribe can give automated, real-time feedback to users. This feedback improves the quality of the image description or other accessible alternative while simultaneously training users with accessibility knowledge to improve future descriptions.

When users fail to make content accessible, either because they choose not to or are unable to do it well, social media platforms can deploy technologies to serve as accommodations in their stead. I compared various methods for achieving this with Twitter A11y, using automated image captioning, text recognition, and crowdsourcing to provide image descriptions for every image encountered by our blind participants. This analysis showed promise in using automated methods as a way to get some visual understanding of an image, but only human authors, especially the original content creator, will have the appropriate content knowledge to make this content accessible. Therefore, we must strive for accessibility solutions before resorting to automated accommodations.

The power of automated approaches to scale widely can be used to amplify and reuse accessible alternatives authored by humans. Both memes and animated GIFs, types of visual media spread frequently on social media, are difficult to describe with automatic captioning models, text recognition, or other fully-automated approaches. Yet, because they are reused often, they are primary candidates for automatic recognition and re-use of pre-written descriptions or other prepared accessible alternatives. Through interviews with social media users with vision impairments, I have determined the important visual elements to describe in memes and GIFs. Simultaneously, I have compared different formats such as the familiar alternative text to more novel audio interpretation of memes. Social media platforms and meme/GIF aggregation sites should develop human-authored alternative text libraries for the most popular content, and record additional audio versions to give a richer non-visual experience when encountering this content. These libraries of accessible alternatives can be automatically matched to instances of the same meme or GIF around the web, enabling human-quality accessibility and content knowledge to be scaled efficiently.

The responsibility to advocate and push for accessible social media content has fallen mostly on people with disabilities, which is unfortunately typical in disability activism. This responsibility to encourage adoption of accessibility features and actually create accessible content instead must be shared by technology platforms, content creators, and wider community members. Platforms must incentivize their users to understand and use accessibility features, while providing accommodation backstops to ensure at least minimal access to content on their platforms. Content creators must be given tools and knowledge to ensure they can include people with disabilities in their audience, and then be held accountable if they fail to do so. The wider community of content consumers, accessibility advocates, and technology researchers must contribute in every feasible way to reducing this widening access gap online. Only through shared and overlapping responsibility to rebuild social media with accessibility in mind will we ensure that people with disabilities have equal access to online communication.

Chapter 10

Bibliography

[1] 3PlayMedia. 2020. 3PlayMedia. (2020). https://www.3playmedia.com/

[2] Adobe. 2018. Add alternative text. (2018). https://www.adobe.com/accessibility/products/acrobat/pdf-repair-add-alternative-text.html

[3] Khaled Albusays and Stephanie Ludi. 2016. Eliciting Programming Challenges Faced by Developers with Visual Impairments: Exploratory Study. In *Proceedings of the 9th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE '16)*. Association for Computing Machinery, New York, NY, USA, 82–85. DOI: http://dx.doi.org/10.1145/2897586.2897616

[4] Nina Reviers Aline Remael and Gert Vercauteren. 2020. Pictures painted in Words: ADLab Audio Description Guidelines. http://www.adlabproject.eu/Docs/adlab%20book/index.html. (2020).

[5] Apple. 2018. Supporting VoiceOver in Your App. (2018). https://developer.apple.com/documentation/uikit/accessibility/supporting_voiceover_in_your_app

[6] Chieko Asakawa and Takashi Itoh. 1998. User Interface of a Home Page Reader. In *Proceedings of the Third International ACM Conference on Assistive Technologies (Assets '98)*. Association for Computing Machinery, New York, NY, USA, 149–156. DOI: http://dx.doi.org/10.1145/274497.274526

[7] Julian Ausserhofer and Axel Maireder. 2013. National politics on Twitter: Structures and topics of a networked public sphere. *Information, Communication & Society* 16, 3 (2013), 291–314.

[8] Brooke E. Auxier, Cody L. Buntain, Paul Jaeger, Jennifer Golbeck, and Hernisa Kacorri. 2019. #HandsOffMyADA: A Twitter Response to the ADA Education and Reform Act. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 527, 12 pages. DOI:http://dx.doi.org/10.1145/3290605.3300757

[9] Tim Berners-Lee and Dan Connolly. 1995. HTML 2.0 Specification. Internet Requests for Comments, *W3C: http://www. w3. org/MarkUp/html-spec* 34, 1866 (Nov. 1995), 1–77.

DOI:http://dx.doi.org/10.17487/RFC1866

[10] Jeffrey P Bigham, Ryan S Kaminsky, Richard E Ladner, Oscar M Danielsson, and Gordon L Hempton. 2006. WebInSight: Making Web Images Accessible. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility - Assets '06 (ASSETS '06)*. ACM, New York, NY, USA, 181. DOI:http://dx.doi.org/10.1145/1168987.1169018

[11] Erin Brady, Meredith Ringel Morris, and Jeffrey P. Bigham. 2015. Gauging Receptiveness to Social Microvolunteering. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1055–1064. DOI:http://dx.doi.org/10.1145/2702123.2702329

[12] Erin L. Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P. Bigham. 2013. Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*. ACM Press, New York, New York, USA, 1225. DOI:http://dx.doi.org/10.1145/2441776.2441915

[13] Carmen J Branje and Deborah I Fels. 2012. Livedescribe: can amateur describers create high-quality audio description? *Journal of Visual Impairment & Blindness* 106, 3 (2012), 154–165.

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. DOI:http://dx.doi.org/10.1191/1478088706qp063oa

[15] Johannes Buchner. 2020. A Python Perceptual Image Hashing Module. (2020). https://github.com/JohannesBuchner/imagehash [Online; accessed 6-May-2020].

[16] C-SPAN. 2018. @cspan/Members of Congress on Twitter. (2018). https://twitter.com/cspan/lists/members-of-congress [Online; accessed 2-November-2018].

[17] Diagram Center. 2019. Specific Guidelines: Art, Photos, and Cartoons. (2019). http://diagramcenter.org/specific-guidelines-final-draft.html [Online; accessed 6-Oct-2020].

[18] Lei Chen and Chong Min Lee. 2017. Convolutional Neural Network for Humor Recognition. *CoRR* abs/1702.02584 (2017).

[19] ADI AD Guidelines Committee. 2003. Guidelines for Audio Description. (2003). https://www.acb.org/adp/guidelines.html [Online; accessed 6-May-2020].

[20] W3 Consortium. 2018. Web Content Accesisbility Guidelines (WCAG) 2.1. (2018). https://www.w3.org/TR/WCAG21/

[21] W3 Consortium and others. 1998. *HTML 4.0 specification*. Technical Report. Technical report, W3 Consortium, 1998. http://www. w3. org/TR/REC-html40. http://www.w3.org/TR/REC-html40

[22] Manuel Cortez. 2019. TWBlue. (2019). https://twblue.es/ [Online; accessed 6-May-2020].

[23] Cameron Cundiff. 2015. alt-text-bot: automatic text descriptions of images on Twitter. (2015). https://devpost.com/software/alt-text-bot

[24] Daniel Dardailler. 1997. *The ALT-server ("An eye for an alt").*

[25] DCMP. 2020. Description Key. https://dcmp.org/learn/captioningkey/624. (2020).

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of 2009 Computer Vision and Pattern Recognition (CVPR '09)*.

[27] Marta Dynel. 2016. "I has seen Image Macros!" Advice Animals memes as visual-verbal jokes. *International Journal of Communication* 10 (2016), 29.

[28] Jason Eppink. 2014. A brief history of the GIF (so far). *Journal of Visual Culture* 13, 3 (2014), 298–306. DOI:http://dx.doi.org/10.1177/1470412914553365

[29] Amber Ferguson. 2016. The #CripTheVote Movement Is Bringing Disability Rights To The 2016 Election. (2016). Retrieved August 20, 2018 from https://www.huffingtonpost.com/entry/cripthevote-movement-2016-election_us_57279637e4b0f309baf177bd.

[30] Facebook for Business. 2016. Capture Attention with Updated Features for Video Ads. https://www.facebook.com/business/news/updated-features-for-video-ads. (2016).

[31] Meme Generator. 2019. The Most Popular Memes of All Time. (2019). https://memegenerator.net/memes/popular/alltime

[32] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning* 63, 1 (2006), 3–42.

[33] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's Almost like They're Trying to Hide It": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 549–559. DOI:http://dx.doi.org/10.1145/3308558.3313605

[34] Cole Gleason, Amy Pavel, Himalini Gururaj, Kris Kitani, and Jeffrey Bigham. 2020. Making GIFs Accessible. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 24, 10 pages. DOI:http://dx.doi.org/10.1145/3373625.3417027

[35] Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B. Chilton, and Jeffrey P. Bigham. 2019a. Making Memes Accessible. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 367–376. DOI:http://dx.doi.org/10.1145/3308561.3353792

[36] Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B. Chilton, and Jeffrey P. Bigham. 2019b. Making Memes Accessible. In *The 21st International ACM SIGACCESS*

Conference on Computers and Accessibility (ASSETS '19). Association for Computing Machinery, New York, NY, USA, 367–376. DOI:http://dx.doi.org/10.1145/3308561.3353792

[37] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M Kitani, and Jeffrey P Bigham. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, USA, 12. DOI:http://dx.doi.org/10.1145/3313831.3376728

[38] Maximilian Golla and Markus Dürmuth. 2018. On the Accuracy of Password Strength Meters. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS '18)*. Association for Computing Machinery, New York, NY, USA, 1567–1582. DOI:http://dx.doi.org/10.1145/3243734.3243769

[39] Google. 2016. Cloud Vision API. https://cloud.google.com/vision/docs/. (2016). Accessed 2019-07-10.

[40] Google. 2018. Content labels. (2018). https://support.google.com/accessibility/android/answer/7158690?hl=en

[41] Doug Gross. 2013. It's settled! Creator tells us how to pronounce 'GIF'. (5 2013). https://www.cnn.com/2013/05/22/tech/web/pronounce-gif/index.html [Online; accessed 21-July-2020].

[42] Big Ten Academic Alliance I.T. Accessibility Group. 2017. Accessibility vs. Accommodation. https://uiowa.instructure.com/courses/40/pages/accessibility-vs-accommodation. (2017).

[43] Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. 2018. Caption Crawler: Enabling Reusable Alternative Text Descriptions Using Reverse Image Search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 518, 11 pages. DOI:http://dx.doi.org/10.1145/3173574.3174092

[44] Ödül Akyapi Gürsimsek. 2016. Animated GIFs as vernacular graphic design: producing Tumblr blogs. *Visual Communication* 15, 3 (2016), 329–349. DOI:http://dx.doi.org/10.1177/1470357216645481

[45] Stephanie Hackett, Bambang Parmanto, and Xiaoming Zeng. 2004. Accessibility of Internet Websites Through Time. In *Proceedings of the 6th International ACM SIGACCESS Conference on Computers and Accessibility (Assets '04)*. ACM, New York, NY, USA, 32–39. DOI:http://dx.doi.org/10.1145/1028630.1028638

[46] Xiaodong He and Li Deng. 2017. Deep Learning for Image-to-Text Generation: A Technical Overview. *IEEE Signal Processing Magazine* 34, 6 (Nov. 2017), 109–116. DOI:http://dx.doi.org/10.1109/MSP.2017.2741510

[47] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1. IEEE, 278–282.

[48] Sally Holloway. 2010. *The Serious Guide to Joke Writing: How To Say Something Funny*

About Anything. Bookshaker, Great Yarmouth, UK. 207 pages.

[49] Web Accessibility Initiative. 2019. Informative Images. (2019). https://www.w3.org/WAI/tutorials/images/informative/ [Online; accessed 6-Oct-2020].

[50] Instagram. 2018. Creating a More Accessible Instagram. (2018). https://instagram-press.com/blog/2018/11/28/creating-a-more-accessible-instagram/

[51] Glenda M Jessup, Elaine Cornell, and Anita C Bundy. 2010. The treasure in leisure activities: Fostering resilience in young people who are blind. *Journal of visual impairment & blindness* 104, 7 (2010), 419–430.

[52] Jialun "Aaron" Jiang, Casey Fiesler, and Jed R Brubaker. 2019. "The Perfect One": Understanding Communication Practices and Challenges with Animated GIFs. *arXiv* 2, CSCW (2019), 1–20. DOI:http://dx.doi.org/10.1145/3274349

[53] @JohnMu. 2018. Alt text is extremely helpful for Google Images – if you want your images to rank there. Even if you use lazy-loading, you know which image will be loaded, so get that information in there as early as possible & test what it renders as. (4 Sept. 2018). https://twitter.com/JohnMu/status/1036901608880254976

[54] Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, Juan Pablo Hourcade, Saeideh Bakhshi, David A Shamma, Lyndon Kennedy, Yale Song, Paloma de Juan, and Joseph 'Jofish' Kaye. 2016. Fast, Cheap, and Good: Why Animated GIFs Engage Us. (2016), 575–586. DOI:http://dx.doi.org/10.1145/2858036.2858532

[55] Chloé Kiddon and Yuriy Brun. 2011. That's What She Said: Double Entendre Identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2 (HLT '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 89–94. http://dl.acm.org/citation.cfm?id=2002736.2002756

[56] Masatomo Kobayashi, Trisha O'Connell, Bryan Gould, Hironobu Takagi, and Chieko Asakawa. 2010. Are synthesized video descriptions acceptable?. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*. 163–170.

[57] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[58] Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. 2007. What Frustrates Screen Reader Users on the Web: A Study of 100 Blind Users. *International Journal of Human–Computer Interaction* 22, 3 (2007), 247–269. DOI:http://dx.doi.org/10.1080/10447310709336964

[59] Veronica Lewis. 2018. How to Write Alt Text for Memes (on Veronica With Four Eyes). (2018). https://veroniiiica.com/2018/11/29/how-to-write-alt-text-for-memes/ [Online; accessed 6-Oct-2020].

[60] Chi-Chin Lin, Yi-Ching Huang, and Jane Yung jen Hsu. 2014. Crowdsourced Explanations for Humorous Internet Memes Based on Linguistic Theories. In *Proceedings of AAAI Conference on Human Computation and Crowdsourcing (HCOMP '14)*.

[61] Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 5988–5999. DOI:http://dx.doi.org/10.1145/3025453.3025814

[62] Daniel C. Dennett Matthew M. Hurley. 2011. *Inside Jokes: Using Humor to Reverse-Engineer the Mind*. The MIT Press, Cambridge, MA, USA. 376 pages.

[63] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.

[64] Microsoft. 2017. Seeing AI — Talking camera app for those with a visual impairment. (2017). https://www.microsoft.com/en-us/seeing-ai/

[65] Microsoft. 2018. Add alternative text to a shape, picture, chart, SmartArt graphic, or other object. (2018). https://support.office.com/en-us/article/add-alternative-text-to-a-shape-picture-chart-smartart-graphic-or-other-object-44989b2a-903c-4d9a-b742-6a75b451c669

[66] Microsoft. 2019. Bing Web Search. https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/. (2019). Accessed 2019-09-20.

[67] Rada Mihalcea and Carlo Strapparava. 2005. Making Computers Laugh: Investigations in Automatic Humor Recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 531–538. DOI:http://dx.doi.org/10.3115/1220575.1220642

[68] Kate M Miltner and Tim Highfield. 2017. Never Gonna GIF You Up: Analyzing the Cultural Significance of the Animated GIF. *Social Media + Society* 3, 3 (2017), 205630511772522. DOI:http://dx.doi.org/10.1177/2056305117725223

[69] Valerie S. Morash, Yue-Ting Siu, Joshua A. Miele, Lucia Hasty, and Steven Landau. 2015. Guiding Novice Web Workers in Making Image Descriptions Using Templates. *ACM Trans. Access. Comput.* 7, 4, Article 12 (Nov. 2015), 21 pages. DOI:http://dx.doi.org/10.1145/2764916

[70] Lourdes Moreno and Paloma Martinez. 2013. Overlapping factors in search engine optimization and web accessibility. *Online Information Review* (2013).

[71] Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. 2018. Rich Representations of Visual Content for Screen Reader Users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 59, 11 pages. DOI:http://dx.doi.org/10.1145/3173574.3173633

[72] Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P. Bigham, and Shaun K. Kane. 2016. "With most of it being pictures now, I rarely use it". In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems -*

CHI '16. 5506–5516. DOI:http://dx.doi.org/10.1145/2858036.2858116

[73] Library of Congress. 2012. Homepage — Meme Generator. (2012). https://www.loc.gov/item/lcwaN0010226/

[74] American Council of the Blind Audio Description Project. 2020. Guidelines for Audio Describers. https://www.acb.org/adp/guidelines.html. (2020).

[75] Abiodun Olalere and Jonathan Lazar. 2011. Accessibility of US federal government home pages: Section 508 compliance and site accessibility statements. *Government Information Quarterly* 28, 3 (2011), 303–309.

[76] Jaclyn Packer, Katie Vizenor, and Joshua A. Miele. 2015. An Overview of Video Description: History, Benefits, and Guidelines. *Journal of Visual Impairment & Blindness* 109, 2 (2015), 83–93. DOI:http://dx.doi.org/10.1177/0145482X1510900204

[77] Becky Parton. 2016. Video captions for online courses: Do YouTube's auto-generated captions meet deaf students' needs? *Journal of Open, Flexible, and Distance Learning* 20, 1 (2016), 8–18.

[78] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20)*. Association for Computing Machinery, New York, NY, USA, 747–759. DOI:http://dx.doi.org/10.1145/3379337.3415864

[79] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[80] Charity Pitcher-Cooper. 2017. YouDescribe. (2017). https://www.ski.org/project/youdescribe

[81] KR Prajwal, CV Jawahar, and Ponnurangam Kumaraguru. 2019. Towards Increased Accessibility of Meme Images with the Help of Rich Face Emotion Captions. (2019).

[82] Audio Description Project. 2020a. Master AD List. (2020). https://acb.org/adp/masterad.html

[83] Audio Description Project. 2020b. What is Audio Description? (2020). https://acb.org/adp/ad.html

[84] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. (2020).

[85] Victor Raskin. 2009. *The Primer of Humor Research*. De Gruyter, Berlin, Germany. 673 pages.

[86] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A Dataset for Movie Description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[87] Adrian Rosebrock. 2014. The complete guide to building an image search engine with

Python and OpenCV. (Dec. 2014). https://www.pyimagesearch.com/2014/12/01/complete-guide-building-image-search-engine-python-opencv/

[88] Ando Saabas. 2015. Random forest interpretation with scikit-learn. https://blog.datadive.net/random-forest-interpretation-with-scikit-learn/. (2015).

[89] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. 2017. Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind. *Aaai Hcomp 17* Hcomp (2017), 147–156. www.aaai.orghttps://www.microsoft.com/en-us/research/wp-content/uploads/2017/08/scalable

[90] Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside Jokes: Identifying Humorous Cartoon Captions. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 1065–1074. DOI:http://dx.doi.org/10.1145/2783258.2783388

[91] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. DOI: http://dx.doi.org/10.1145/3313831.3376404

[92] Louis Staples. 2019. Are Memes Ruining Television? (10 2019). https://www.gq.com/story/are-memes-ruining-television [Online; accessed 6-May-2020].

[93] Hironobu Takagi, Shinya Kawanaka, Masatomo Kobayashi, Takashi Itoh, and Chieko Asakawa. 2008. Social Accessibility: Achieving Accessibility Through Collaborative Metadata Authoring. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility (Assets '08)*. ACM, New York, NY, USA, 193–200. DOI:http://dx.doi.org/10.1145/1414471.1414507

[94] Julia M. Taylor and Lawrence J. Mazlack. 2004. Computationally recognizing wordplay in jokes. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci '04)*, Vol. 26.

[95] @todd. 2016. Accessible images for everyone. (29 March 2016). https://blog.twitter.com/official/en_us/a/2016/accessible-images-for-everyone.html

[96] Jackson Tolins and Patrawat Samermit. 2016. GIFs as Embodied Enactments in Text-Mediated Conversation. *Research on Language and Social Interaction* 49, 2 (2016), 75–91. DOI:http://dx.doi.org/10.1080/08351813.2016.1164391

[97] Kenneth Tran, Xiaodong He, Lei Zhang, and Jian Sun. 2016. Rich Image Captioning in the Wild. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 434–441. DOI:http://dx.doi.org/10.1109/CVPRW.2016.61

[98] Zeynep Tufekci and Christopher Wilson. 2012. Social Media and the Decision to Par-

ticipate in Political Protest: Observations From Tahrir Square. *Journal of Communication* 62, 2 (03 2012), 363–379. DOI:http://dx.doi.org/10.1111/j.1460-2466.2012.01629.x

[99] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm* 10, 1 (2010), 178–185.

[100] Twitter. 2018a. About your Twitter timeline. (2018). https://help.twitter.com/en/using-twitter/twitter-timeline

[101] Twitter. 2018b. How to make images accessible for people. (2018). https://help.twitter.com/en/using-twitter/picture-descriptions

[102] Twitter. 2018c. Post, retrieve and engage with Tweets. (2018). https://developer.twitter.com/en/docs/tweets/post-and-engage/api-reference/get-statuses-lookup.html

[103] Twitter. 2018d. Sample realtime Tweets. (2018). https://developer.twitter.com/en/docs/tweets/sample-realtime/guides/recovery-and-redundancy

[104] John Vorhaus. 1994. *The Comic Toolbox: How to Be Funny Even If You're Not*. Silman James Press, Los Angeles, CA. 191 pages.

[105] Ye-Yi Wang, Alex Acero, and Ciprian Chelba. 2003a. Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '03)*. IEEE, 577–582.

[106] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003b. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2. Ieee, 1398–1402.

[107] Kelly Warren. A day in the life of an audio describer. https://dcmp.org/learn/277-a-day-in-the-life-of-an-audio-describer. (????). [Online; accessed 6-May-2020].

[108] Elwyn Brooks White. 1954. Some remarks on humor. *The Second Tree from the Corner* (1954), 173–181.

[109] Wikipedia contributors. 2018. List of most-followed Twitter accounts — Wikipedia, The Free Encyclopedia. (2018). https://en.wikipedia.org/w/index.php?title=List_of_most-followed_Twitter_accounts&oldid=866718146 [Online; accessed 2-November-2018].

[110] Shaomei Wu and Lada A. Adamic. 2014. Visually impaired users on an online social network. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. ACM Press, New York, New York, USA, 3133–3142. DOI:http://dx.doi.org/10.1145/2556288.2557415

[111] Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Who Says What to Whom on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. ACM, New York, NY, USA, 705–714. DOI:http://dx.doi.

org/10.1145/1963405.1963504

[112] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1180–1192. DOI:http://dx.doi.org/10.1145/2998181.2998364

[113] Xiaoyi Zhang, Anne Spencer Ross, and James Fogarty. 2018. Robust Annotation of Mobile Application Interfaces in Methods for Accessibility Repair and Enhancement. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. Association for Computing Machinery, New York, NY, USA, 609–621. DOI:http://dx.doi.org/10.1145/3242587.3242616

[114] Yuhang Zhao, Cynthia L. Bennett, Hrvoje Benko, Edward Cutrell, Christian Holz, Meredith Ringel Morris, and Mike Sinclair. 2018. Enabling People with Visual Impairments to Navigate Virtual Reality with a Haptic and Auditory Cane Simulation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. DOI:http://dx.doi.org/10.1145/3173574.3173690

[115] Yuhang Zhao, Edward Cutrell, Christian Holz, Meredith Ringel Morris, Eyal Ofek, and Andrew D. Wilson. 2019. SeeingVR: A Set of Tools to Make Virtual Reality More Accessible to People with Low Vision. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. DOI:http://dx.doi.org/10.1145/3290605.3300341

[116] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2017. The Effect of Computer-Generated Descriptions on Photo-Sharing Experiences of People with Visual Impairments. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 121 (Dec. 2017), 22 pages. DOI:http://dx.doi.org/10.1145/3134756

Appendices

Appendix A

Twitter A11y Interviews

A.1 Pre-study interview questions

1. Demographics information:
   (a) Age
   (b) Gender
   (c) Years using Twitter
   (d) Visual ability
   (e) Any other disabilities?
   (f) Years of vision disability
   (g) What screen readers do you use?
   (h) What other social networks do you use?
2. How accessible do you find Twitter?
3. What are major barriers for using the site?
4. What percent of images that you encounter on Twitter do you think have an image description?
5. Do you find it easy to understand tweets that include an image with no description?
6. Does accessibility of images affect which accounts you follow?
7. What about other forms of media, such as videos and GIFs; how accessible are those?
8. What changes would you make to Twitter to make it more accessible for people with vision impairments?
9. Do you use any other tools to make Twitter more accessible?

A.2 Post-study interview questions

1. What was your experience using the tool? Did you enjoy it?

2. What percent of image do you think were accessible when you used the tool and browsed Twitter?
3. We used many methods to make Twitter images accessible. Which did you find the best? The worst?
4. Were you ever uneasy or felt like you didn't trust a description?
5. Would you continue using a tool like this?
6. What could be most improved about this tool? What worked well?
7. Do you think this tool would translate well to other social networks you use?
8. What else should we focus on to make social networks more accessible?

Appendix B

Alt Text Quality Survey Questions

1. Demographics information:
   (a) What is your current age?
   (b) What is your gender?
   (c) How would you describe your level of vision?
   (d) Since what age have you had that level of visual ability?
   (e) Are you fluent in English?
   (f) Do you use a screen reader to access content on the Internet?
   (g) Which screen readers do you use on your computer or mobile devices?
2. 20 Social Media Posts with 2 Answers Each
3. Wrap Up:
   (a) We are researching what aspects make an image description high quality. Based on your past experiences, what aspects of image descriptions should we focus more time on improving?
   (b) In what order do you think information should be prioritized or presented for image descriptions?
   (c) Please add any additional comments or suggestions here:

Appendix C

Making Memes Accessible

Figure C.1: An example of each meme template. In the study, we used 5 example memes for each meme template for 45 total memes.

C.1 Meme templates

In our study, we used nine different visual memes (Figure C.1) with five examples for each. The names of the memes we used, are listed here:

A Awesome Awkward Penguin

B Success Kid

C Philosoraptor

D Bad Luck Brian

E Most Interesting Man in the World

F Confession Bear

G Awkward Moment Seal

H First World Problems

I Futurama Fry

We include the alt text template for each meme (Table C.1) and a meme example for each (Figure C.1).

Base meme	Alt text template
Confession Bear	Baby black bear staring into space with paws on a tree branch. Overlaid text on top [top text]. Overlaid text on bottom [bottom text].
Success Kid	Toddler clenching fist in front of a smug face. Overlaid text on top [top text]. Overlaid text on bottom [bottom text]
Awkward Moment Seal	Close up of a seal's face with wide eyes and a straight face. Overlaid text on top [top text]. Overlaid text on bottom [bottom text].
Interesting Man	A man with gray hair in a nice shirt and jacket smirking while leaning on one elbow. A bottle of Dos Equis beet is in front of him. Overlaid text on top [top text]. Overlaid text on bottom [bottom text].
Philosoraptor	A drawing of a green dinosaur raptor with a claw to it's chin and mouth open as if it is contemplating something. Overlaid text on top [top text]. Overlaid text on bottom [bottom text].
First World Problems	Close up on a woman with her eyes closed head in one hand and a stream of tears running down her cheek. Overlaid text on top [top text]. Overlaid text on bottom [bottom text].
Awesome Awkward Penguin	Close up of a seal's face with wide eyes and a straight face. Overlaid text on top [top text]. Overlaid text on bottom [bottom text].
Bad Luck Brian	A young kid in an awkward school photo. He is wearing a plaid vest and has an open smile where you can see his braces. Overlaid text on top [top text]. Overlaid text on bottom [bottom text].
Futurama Fry	Fry from the show Futurama a cartoon man with orange hair squinting his eyes as if he suspects something. Overlaid text on top [top text]. Overlaid text on bottom [bottom text].

Table C.1: Names of memes (base meme) with the corresponding alt text template for each meme. When the template lists [top text] and [bottom text], we replace the placeholders with the example meme text. Audio meme templates included in the supplemental materials.

Appendix D

GIF Interview Questions

D.1 Session 1

1. Collection of demographic information.
2. How often do you encounter GIFs? In what contexts?
3. Do you remember the last time you encountered a GIF? What cues do you use to interpret it?
4. Have you encountered GIFs elsewhere on the web? Are they accessible there?
5. Have you encountered GIFs where people add informal alt text (in the original post or in the comments)?
6. Has not being able to access the visual content of a GIF prevented you from understanding something in the past?
7. Have you had any experience of someone helping you access a GIF? What was the context?
8. What would you do to make GIFs accessible?

D.2 Session 2

1. Which format did you most prefer? Why?
2. Which format did you least prefer? Why?
3. Given a tool that could provide all three formats for popular GIFs, which do you think you would enable at least some of the time? Why?
4. Given accessible alternatives for popular GIFs, do you imagine you would seek out more conversations that contain GIFs? If so, where would you look?

Appendix E

Related Publications and Awards

The following publications and awards either contributed directly to the completion of this dissertation or my Ph.D. degree:

E.1 Publications

[C.5] **Cole Gleason**, Stephanie Valencia, Lynn Kirabo, Jason Wu, Anhong Guo, Elizabeth J. Carter, Jeffrey P. Bigham, Cynthia L. Bennett, Amy Pavel. 2020. Disability and the COVID-19 Pandemic: Using Twitter to Understand Accessibility during Rapid Societal Transition. *ASSETS 2020*

[C.4] **Cole Gleason**, Amy Pavel, Himalini Gururaj, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Making GIFs Accessible. *ASSETS 2020*

[C.3] **Cole Gleason**, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. *CHI 2020* **[Best Paper Honorable Mention (Top 5%)]**

[C.2] **Cole Gleason**, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B. Chilton, and Jeffrey P. Bigham. 2019. Making Memes Accessible. *ASSETS 2019*

[C.1] **Cole Gleason**, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. *WWW 2019*

[J.2] **Cole Gleason**, Alexander J. Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2018. FootNotes: Geo-referenced Audio Annotations for Nonvisual Exploration. *IMWUT 2018*

[J.1] **Cole Gleason**, Dragan Ahmetovic, Saiph Savage, Carlos Toxtli, Carl Posthuma, Chieko Asakawa, Kris M. Kitani, and Jeffrey P. Bigham. 2018. Crowdsourcing the Installation and Maintenance of Indoor Localization Infrastructure to Support Blind Navigation. *IMWUT 2018*

E.2 Awards

- ACM Student Research Competition - Graduate Student Finalist
- NSF Graduate Research Fellowship Program - Fellow
- NSF Graduate Research Fellowship Program - Honorable Mention
- The Paciello Group - Web Accessibility Challenge Delegates' Award
- IBM - Web for All People with Disabilities Award