Adversarial poetry

The tools of resistance

Christian Heck

Abstract
A large number of current "social movement prediction models" are automated using Natural Language Processing (NLP) methods. Poetry presents probably the greatest challenge to this computational way of processing texts in natural language. It has a high density of ambiguity and usually plays by its own rules.
This article aims to provide an introduction to the concept of adversarial poetry, i.e. a practice of subversive resistance by composing politically motivated texts in such a way that they are misinterpreted by common NLP prediction models.
It is not only governments and their secret services that are interested in surveillance through recording and analyzing the structures and individual actors in social movements. By evaluating social media accounts, private companies and platforms such as Twitter or Facebook have also created their own instruments of domination to automatically detect "abnormalities" and pass them on to the relevant authorities in natural language. As a result, activists must create new spaces to communicate on the web under the radar of surveillance. This requires self-determined and self-organized platforms for participation within socio-technical spaces of action. Spaces that are always based on text, since computers are semiotic machines.
The literary currents and movements of the last century were, so

to speak, the forerunners in a development of alternative forms of language that are very difficult for computers to read today. In connection with the social dynamic of hacktivism, strategies can be developed that provide activists with the possibility to play with contemporary instruments of domination.

It is possible—if you know which rules to break and how to break them without bringing down the overall set of rules—to destabilize this socio-technical space of action. How to appropriate this space with poietic rule-breaking will be the subject of the following chapters.

Introduction

Those who change their use of language act and effect differently. "We have become accustomed to the fact that one must speak differently at markets than, for example, at political meetings, that religious speech is inappropriate in court" (Nassehi 2019, 164; translated by the author), and also that a scientific style of expectation must disregard which denomination someone belongs to or the color of their skin. The sociologist Armin Nassehi speaks of "certain forms of social intercourse" from which compact patterns of action emerge (2019, 164; translated by the author). These, in turn, bring with them corresponding specialized knowledge and special language (jargon), forms of reflection and milieus (cf. 165).

Political language, or political speech, constitutes one of these special languages and "this language is not just any instrument of politics, but the condition of its possibility" (2010, 6; translated by the author), as Heiko Girnth, who works at the *German Language Atlas* research center at the Philipps-University of Marburg, introduces "Dossier: Sprache und Politik" (Dossier: Language and Politics). He goes on to claim that "ultimately, anything that is of public interest can become political" (6; translated by the author). Politics can thus permeate all areas of social life.

However, the concept of *public interest* can be used as an example to play through the endless specialized language and everyday language differences in the meaning of words, which all too often lead to communication problems between political actors and citizens. "The everyday language way of reading is based, in some cases, on concrete experience, but it is mostly the result of cultural memory," Girnth continues (6; translated by the author). Political language, in turn, often works with ideological vocabulary, i.e., our value systems, certain thought patterns and concepts such as *freedom, justice or peace*. Buzzwords, such as *democracy* or *terrorism* or *public interest* and word combinations, neologisms, and metaphors are often used to convey political issues more easily against a background of

already familiar experiences.

The use of such linguistic techniques could be described "as social techniques to relieve us of the burden of establishing consensus and agreement" (Nassehi 2019, 206; translated by the author)—they are supposed to facilitate and, in some cases, even relieve us of laborious consensus building and grassroots democratic processes. But they can also be seen as a necessary attempt, i.e., as a form of social life, of social togetherness. Certain intentions are represented by certain individuals and groups (not only in political speech), and these intentions are usually communicated in such a way that an expression of opinion occurs in the sense of an interpersonal understanding.

This communication consists, according to John R. Searle, "in the grasp of my meaning" (1998, 145). Searle speaks of an intention to communicate, namely "the intention that the hearer should recognize my meaning intention. The communication intention is the intention to produce in the hearer the knowledge of my meaning by getting him to recognize my intention to produce in him that knowledge" (145).

We are very concerned to continue to maintain our respective intentions of meaning and understanding for human readers while the authors of adversarial poetry try to destabilize our socio-technical spaces of action by a priori inscriptions of poietic (primarily syntactic) rule breaks into them.

This is where Searle's communication intention meets three general principles for writing adversarial attacks:[1]

— The author wishes to continue to ensure a grammatical fluency of language, albeit ordered according to their own specially created grammatical rules.

— The conventional meaning of the sentences should be preserved as far as possible, which means that when words are replaced by synonyms in pre-trained language models,

1- An adversarial attack, in the context of adversarial poetry, is a perturbative input designed to fool machine learning models in NLP, i.e., to misinterpret or misclassify the original input (text).

for example, they should be in the same *word space*.[2]

— Ultimately, of course, the main goal remains "human prediction consistency" (Jin et al. 2019), i.e., that readers continue to recognize the author's intention, that they do not only express the words and sentences written down but also mean them.

In every single step of adversarial poetry writing, respective meanings have to be preserved for human readers but reinterpreted differently by the prediction system. The interpretation sovereignty, not only over our words but also over the effectiveness of our future actions, should be reclaimed in this manner. For this purpose, direct references are made to a range of current scholarly research in computational linguistics and digital humanities, social and political sciences, and current AI research with a focus on deep learning and NLP.[3] Studying this research reveals the hurdles and obstacles that must be overcome if we are to create linguistic freedom through poetry and work out self-determined and independent movements in society. It also indicates the particular screws we must turn in order to destabilize language models implemented in social movement prediction applications. It points out how little we actually know about the structure and properties of these vector spaces, despite the widespread use of word embedding. Above all, looking at the current research shows that neither the mere transfer of our words into another text genre—in our case poetry—nor that of adversarial hacking to scramble neural word embeddings

2- Word embeddings are based on the idea that, in contrast to formal linguistics and the Chomskyan tradition, contextual information alone is a viable representation of linguistic elements. Depending on the language model, each word is represented as a vector in a semantic vector space (word space) of about 100 to 300 dimensions, based on the textual context in which the word occurs. This technology has become one of the most popular tools in the NLP research communities since the advent of Google's Word2vec (Mikolov et al. 2013), as these embeddings are easy and convenient to use and provide state of the art results. It is an integral part of almost all the applications and research presented in this paper.

3- Natural Language Processing (NLP) is a mixed science. The field consists proportionately of computer linguistics, computer science, and artificial intelligence: the science of algorithmic processing of language, the science of processing data, and the science of artificial intelligent behavior.

in language models can be a panacea in view of state-of-the-art poetry analysis methods.

Numerous approaches to writing adversarial poetry can be found in political poetry and in the experimental literature of the last century. Although their respective lines of tradition were largely separate from one another, both pursued one and the same goal. Both worked to replace, or at least destabilize, the functioning of old and established specialized languages, linguistic customs and systems through the creation and the use of a new language. This objective lies at the heart of the activities of the avant-garde movements and bohemian milieus, poets, politicians, writers, and artists listed in the following two chapters.

As we know, (not only) from German history, enforced changes of language use can reduce our possibilities of thought in a devastating way. In order to reduce the possibilities of thought, one must reduce the possibilities of expression and create simple linguistic schemes, preferably clearly evaluated opposites, as linguist Jürgen Schiewe proposes in reference to literary scientist Viktor Klemperer's treatise *LTI*: *Notizbuch eines Philologen* (cf. Schiewe 1998, 213). Klemperer, himself a survivor of the Holocaust, vividly drew "an oppressive picture of the language of the Third Reich, the 'lingua tertii imperii'," or in short: LTI (Schiewe 1998, 213; translated by the author). For him, the LTI was an important part of political domination, a language "that became literally fixed in all its basic features" with the publication of Hitler's *Mein Kampf* in 1925 (Klemperer 1947, 25; translated by the author).

But changes in language use can also expand our possibilities for thinking, by creating a participatory and free space in which society can unfold through collective action. Representatives of both traditions, political poetry as well as experimental literature, were convinced of the idea that a new use of language is the prerequisite for being able to think something new.

Experimental literature

Stein on NLP

> If one wants to understand how art and technology relate to each other in our European tradition, one has to go far afield. Today we are used to seeing intuition and rationality as opposites. The common origin of poetics (poietike - the creative, poetic art) and technology (techné) in Greek poiesis, on the other hand, has been largely forgotten. (Trogemann 2016; translated by the author)

There is probably no more beautiful essayistic approach to *techné* and *poietike* in Western literary history than that of Gertrude Stein in her "Poetry and Grammar," a passionate literary description of writing. What Stein showed in her own unique language is that technology is a way of thinking and doing, for which she was often criticized, particularly, because her literary discoveries always tasted of the scientific laboratory (cf. Brinnin 1964, 177).

Since its origin, modern literature has had a connection, albeit mostly ambivalent, to the exact sciences. It has always referred to them, implicitly or explicitly, be it to assert their otherness or to gain some form of legitimacy. This tendency developed over the twentieth century and practiced its very own exploration of fundamental scientific questions (cf. Maniez, Ludot-Vlasak, and Dumas 2012).

"Poetry is really loving the name of anything" wrote Gertrude Stein, continually doing everything to "creating it without naming it" (Stein 1998, 232, 237). One can love a name, one can feel a name, and one can also know it: "nouns are the names of things and so nouns are the basis of poetry" (234). She proposed that this was already the case in the times of Homer, of Chaucer, and in the writing of the Bible, which were all "drunk with nouns, to name to know how to name earth sea and sky" (233). Today we no longer know this. After hundreds of years and after thousands of poems have been written, we must learn to free ourselves from nouns. But they have remained the basis of poetry. Only differently. Poetry has changed its form, its grammar, its respective use in the word order to give life back to the noun.

Many aspects of our everyday life today are beyond our horizon of

experience. Typically, this concerns ideas and concepts that confront us with insurmountable linguistic boundaries (cf. Duerr 1974, 32). These limits have always had to be overcome. For we constantly get to know these terms anew and experience them by using them, by moving through them (cf. Wittgenstein 1984, 9). And in our everyday life we probably move in the conceptual realm of new technologies more than ever.

To grasp the respective significance of those terms for our everyday lives (at least in public debates) presents us with enormous challenges and requires us to venture into uncharted territory; a terrain in which, astonishingly, precisely these conceptualizations lead to stability problems in language models (cf. Pierrejean and Tanguy 2019). It is hardly comprehensible to the public how little we as researchers know about the structure of these vector spaces. Nevertheless, they are widely and instrumentally implemented in societal interaction. According to Wittgenstein's *Philosophische Grammatik*, the semantic representation within these models we use is not their meaning, but the way this use intervenes in our lives.

In the course of the last century, modernist poets developed their own usages and grammatical rules, many of which were completely unreadable by literary conventions. This led to the point of structural indistinguishability to our everyday language and to utilitarian or literary prose.[4] Hence, it is difficult to align this way of writing with the computational procedures for poetry analysis. Therefore, NLP researchers try to work out techniques and language models which bring these new rules, especially created by respective poets, into a strict form or to recognize a new grammar within it—for example, equating poems with the syntax constructions of prose texts (cf. Barakhnin and Pastushkov 2019) with the help of chunks and syntax groups (Noam Chomsky theory). For the writing of adversarial poetry, this would mean trying to avoid strict word order.

4- Texts written for a specific purpose are classified as utilitarian prose: the speech, the conversation, the letter, the article or the factual text (legal texts, instructions for use, etc.). Literary prose refers to texts that are commonly referred to as narratives or stories.

For this purpose, much of the current research in this area enters a space between language combinatorics and algorithmically processed language: "If we define a text as a combination of elements (letters, words, lemma, interpunction, POS, n-grams, etc.) we can count these elements and compare texts to find a pattern which might be characteristic for authorial style or genre," explains Nanette Rißler-Pipka from the field of digital humanities in her paper "In Search of a New Language: Measuring Style of Góngora and Picasso" (2019). Here she tries to extract countable structures from the poems of both authors in order to identify possible hidden rules behind the structure of the texts.

Pablo Picasso in particular was known for his combinatorial play with language. Like Stein, he knew how to play with new syntaxes, creating new words and trying out new grammatical orders. Just as Stein herself was inspired in her writing by the impressionism of Cézanne and, above all, by Picasso's cubism, her experimental writing techniques naturally inspired her closely acquainted painter friends as well;[5] techniques that Stein developed from her preference for diagramming sentences. This also holds true for her unusually direct reference to the concrete world of things. Stein constantly attempted to recreate concrete objects through unconventional new names and to depict their visual characteristics within the framework of individual sentence structures or object related rhythms. Through these rhythmic-syntactic operations she led her readers to the semantic level of meaning only (cf. Kirchner 2001).

Those who write in rhythm

Numerous artists and writers of the last century studied Stein's work and her way of writing. Among them were representatives

5- Gertrude Stein and her partner Alice B. Toklas, together with Stein's brother Leo, presented their collection of works of modern art by Paul Cézanne, Henri Matisse, and Pablo Picasso, among others, at their apartment at 27 Rue de Fleurus (1903–1938). Writers and artists such as Georges Braque, F. Scott Fitzgerald, Guillaume Apollinaire, Pablo Picasso, James Joyce, Thornton Wilder, Ezra Pound, Francis Picabia, and Henri Matisse met there regularly on the renowned Saturday evenings.

of the *Stuttgarter Schule* such as Max Bense, Ernst Jandl, and Reinhard Döhl.[6] They became familiar with Stein's writing techniques primarily through Helmut Heißenbüttel's public presentation of her work.

Heißenbüttel began his Georg Büchner Prize Speech 1969 as follows: "Eine Rede ist eine Rede. Eine Rede ist eine Rede heißt eine Rede ist eine geredete Rede das heißt sie muß geredet das heißt gehalten werden" (1970, 42). In doing so he introduced Steinese, that repetitive, sparingly punctuated literary language, into the German poetic repertoire (cf. Melin 1985, 497).

The *oulipotique* writing style of the Oulipo circle of authors was also strongly influenced by Stein's combinatorics.[7] Among other things, Oulipo set itself the task of examining the formal aspects of Lettrism, which in turn explicitly dealt with the smallest units of our natural language in terms of content: letters. Isodore Isou, the founder of Parisian Lettrism in 1942, deconstructed poetry into mere sequences of letters, which he recited in Parisian bars provoking numerous scandals at the time.

Kurt Schwitters, a Dadaist poet, painter and spatial artist, had taken decompositional Lettrism—avant la lettre—to the extreme a few years earlier in his "i-Gedicht" ([1922] 1974). In that poem, name and thing merge into one through his play with the levels of signified and signifier. His elementary material had always been letters, numbers, colors, and notes, and by analyzing them he came across the elements of language. But the Dadaists were not only the intellectual fathers of the Lettrists[8] —their artistic work always opposed the ruling political system. Dadaist

6- The *Stuttgarter Schule* included Max Bense, Helmut Heißenbüttel, Reinhard Döhl, Ludwig Harig, Franz Mon, Ernst Jandl and several visual artists and musicians. They published concrete poems. Language did not serve them primarily to describe an event, but the words and letters were used as a means of visual and also acoustic expression. Their artistic production was closely linked to scientific research in the field of literature, sign and information theory (cf. Rosen 2004).

7- The French, Italian, US-American, and Transylvanian writers' circle Oulipo (Ouvroir de littérature potentielle) was founded in 1960. The members of Oulipo pleaded for a new formal poetry and experimented with a procedural or rule-governed poetics, often creating elaborate numerical constraints that a particular text had to follow.

poetry—following formal instructions—is also in this tradition of experimental literature, with its constant attempt to create a construct consisting of a combination of randomness and control (cf. Paul 2015, 11–13).

There are some notable Beatnik concepts as well: in his 1970 essay "The Electronic Revolution," which helped Gilles Deleuze to develop his idea of the society of control, William S. Burroughs put forward concepts of how we can grammatically scramble the dominant form of society in order to unscramble the syntax of control. These included instructions like deleting the copula "is/are" to disrupt fixed identities or replacing definite articles like "the" with indefinite articles "a/an" to avoid reification. In replacing "either/or" with "and," Burroughs ignored the law of contradiction (cf. 1970, 33–35).

A few years before Oulipo's literary experiments in France, Jack Kerouac introduced the term Beat Generation to the New York literary scene. The authors who identified themselves primarily with this self-imposed generation not only named themselves in analogy to the Lost Generation (F. Scott Fitzgerald, Ernest Hemingway, Gertrude Stein and many others), they were also often called "those who write in rhythm," because they, like Stein, poetically appropriated language technologies of their time. Their poetry seemed to be aware of what the political theorist Hannah Arendt described a few years later in *Vita Activa:* "that man must have already become accustomed to this rhythm of the machines, as it were, when he even conceived such a thing as a machine in his mind" (1981, 136; translated by the author). For as long as one writes with and through machines, their mechanical processes and their discrete units of time also take the place of our own body rhythm. It was therefore necessary to be able to write language that carries "all the history of its intellectual recreation" within itself (Stein 1988, 238).

"Whoever wants poetry must also want the typewriter," Arno

8- The Lettrist Michèle Bernstein explained in 1983: "Everyone is the son of many fathers. There was the father we hated, which was surrealism. And there was the father we loved, which was Dada. We were the children of both" (Marcus 2009, 175).

Schmidt wrote in his monumental work *Zettel's Traum* (1970; translated by the author). The typewriter rarely replaced the handwriting of writers, just as rarely as Gutenberg's printing press did before or our computers do today. Instead, it "rather displaces and redirects it toward writing, inventing, and thinking about other things" (Dick 2013, 86).

Today it is the code poets who, just like the Lettrists in the past, deal with the smallest units of our natural language, but they do so in the digital realm. They are the ones who read and write the natural language texts as well as the encoded texts, who think about all these abstract intermediate levels between things and our thinking. They read texts that they cannot see and make them appear through their work.

Code poets are constantly moving between what was and what is to come—in other words, the a priori knowledge that is inscribed in digital technologies. This knowledge is inextricably linked with early industrial history, namely with the abstract modeling of social groups, algorithmic problem solving, and statistical prediction procedures. In short, purposive rational thinking (cf. Trogemann 2019). Code poets move through these texts, first a posteriori with the aid of formal scaffolding, into these language models with neural embeddings, then on to the input representation, until they finally arrive at the respective vectors of meaning, the word embeddings.

They also weigh up and estimate, at best, the respective consequences of their writing: the potential byproducts of the literary code. This means, in the writing of poetry, to clearly differentiate between syntactic codes, which according to Umberto Eco represent the knowledge of the constructions, and the semantic codes, which refer to their respective function (Eco 1994, 329).

Code poets assign their cultural value to the code when writing. If they were to remain solely with the semantic codifications of their texts, they would be unable to provide readers with anything they were not already prepared for. The text would only give solutions worked out in a predefined form. Finally, it would lock any free space for the participation and imagination of the

reader or the activist in the respective space of action: precisely in that space in which the algorithms to be used are embedded (cf. Trogemann 2010).

What code poets do can be called code poetry or code literature. There is another alongside these two relatively young genres: conceptual literature. The best-known representatives are the author and literary scholar Hannes Bajohr in the German-speaking world and the writer and conceptual artist Kenneth Goldsmith in the English-speaking world. Goldsmith himself explicitly places his writing in the Steinian tradition and often refers to reading her monumental work *The Making of Americans* in one go as "like trying to read the Web linearly" (2011, 305). In his book *Uncreative Writing*, Goldsmith elaborates a concept which he describes as "the art of managing information and representing it as writing" (446). This is a concept that is (not only) forced upon us by our communication in and through social networks in which we now continuously parse, sort, forward, channel, tweet and retweet expressions. Goldsmith suggests that "what we're experiencing for the first time" in our everyday lives "is the ability of language to alter all media, be it images, video, music, or text" and thus the social habitualizations accompanying its use (65).

Brain works

Gertrude Stein studied in what was, at the time, the emerging fields of psychology and brain science at Radcliffe College (1893–1897) at Harvard University and then at Johns Hopkins School of Medicine (1897–1902). In her studies of brain science, Stein, together with Florence Sabin (the first professor of medicine at Johns Hopkins Medical School), produced repetitive brain models and diagrams over and over again in detailed handwork (cf. Stein and Barnes 1950, 148). As a member of the School of Medicine's inaugural class of 1897, she joined the first generation of students to learn about a new experimental focus in medical education. Under the direction of Professor Franklin Mall, in whose laboratory Gertrude Stein later conducted independent research

(1901–1902), Johns Hopkins Medical School was the first American medical institution to teach anatomy in the dissecting room rather than the lecture hall. This was a groundbreaking shift from a descriptive teaching method to an experimental one of analyzing, observing, treating, viewing, tabulating and classifying (cf. Mall 1896, 86).

Stein and other experimental writers, such as William Carlos Williams, who was also a practicing doctor during his time as a writer, integrated many concepts and models from early neuroscience and brain research into their experimental literary works and created entirely new approaches to language through their poetic language techniques. Central to this are Stein's well-known stylistic devices of repetition and abstraction (Farland 2004, 118).

At the beginning of the twentieth century both the disciplines of modernist poetry and modern neuroscience discovered a new space to assemble fragments into meaningful arrangements to replace what they believed to be the obsolete systems of the nineteenth century (Ambrosio 2018).

A few decades later, William S. Burroughs, probably the most ambivalent Beatnik, also incorporated his very own forms of cognition from the neurosciences into his work. The *Dreamachine* Burroughs developed in collaboration with artist Brion Gysin, for example, made direct reference to the experiments of neurophysiologist and roboticist William Grey Walter, who studied the stimuli triggered by perception of a strobe and its direct influence on the electrical activities in our brains (cf. Walter 1953).

Through discovering synaptic spaces, the syntax of our formal technical languages, especially that of artificial neural networks, thus entered a new millennium, hand in hand with early poetic language techniques and experiments.

Political poetry

Anarchistic decision making (ADM)

Poetry is dynamite for all orders of this world!
Heinrich Böll

The meetings at Café Stefanie in Munich organized by the satirical magazine *Simplicissimus*[10] were to the *Schwabinger Bohème* what the famous Saturdays in Gertrude Stein's salon in Paris were to the avant-garde of the early twentieth century. Among the bohemian milieu of Munich-Schwabing at the turn of the century were intellectuals, expressionists, Dadaists, cabaretists, poets, publicists, and anarchists such as Heinrich Mann, Emmy Hennings, Frank Wedekind, Franz Blei, Paul Klee, Otto Gross, Erich Mühsam, Gustav Landauer, Ernst Toller, and Kurt Eisner. The latter four were to become important political players a few years later. They attributed great revolutionary power to the poetic word, art, and education for social change, and to this end they took political office as members of the Munich council republic in 1918 and 1919.[11] The achievements of this bloodless revolution, which led to the end of monarchy, not only included the main goal, the installation of democracy, but also women's suffrage, the eight-hour workday, and other milestones on the way to social equality and justice.

For lyricist Gustav Landauer in particular, poetry and poetic forms of expression were a prerequisite for the creation of communitarian, free and just societies. Only through a concomitant linguistic reconstruction did it seem possible to him to destroy the dominant, ruling language with its strict and rigid terminology. He believed deeply in the political capacity of poetic language (cf. Mokrohs 2018). For him, it created something new by allowing for blur, paradox, and contradiction. Landauer derived the creation of a new, anarchistic model of society from these precise blurs of poetry. Individuals were meant to destroy fixed

10- *Simplicissimus* was a satirical weekly magazine with editorial headquarters in Munich, published from April 4, 1896 to September 13, 1944.

11- The Munich or Bavarian Councilors' Republic was proclaimed on April 7, 1919 and represented an attempt to establish a socialist councilors' republic in the Free State of Bavaria, which had been founded shortly before. It lasted about four weeks.

terms and given concepts, come to an understanding of themselves and join similarly minded people to establish self-governing communities of production and life (Friedmann 2019, 5). Landauer advocated a form of anarchism understood as the absence of coercion, domination, and hierarchy, which could only be fought for collectively and without violence: a liberation from egoistic individualism in order to be able to develop as independent and self-reliant individuals.

Hannah Arendt's grassroots-democratic, council-political concept of power also held that politically free action could only come about through collective action in the public sphere. For Arendt, the (communicative) power potential of emancipatory movements, which can oppose repressive instruments of domination, led, in the sense of her *Vita Activa*, through collective action in public space towards empowerment. The movements that emerged from this period are not only an integral part of modern societies, they have also shaped them in their present form. They oscillate between urban and media space and march on data highways and paved streets. Networking in socio-technical environments that are familiar to them, they open up new social and political spaces for action.

However, these social and political spaces also present new challenges that must be overcome in order to create a public space from this media space. To put it more concretely: textualizing and co-defining utopias and political goals can only happen in anarchic moments and grassroots democratic processes in the plurality among people and in the recognition of the needs of the respective counterpart as well as in the ambivalence of existence (cf. Arendt 1981). This entails accepting others in their otherness and beginning to understand them in their own way.

In socio-technical networks we are someone different than out there on the street. Each one of us is many: many digital identities. We are algorithmic narratives that move and act according to codes and laws different to those in the analog world. Some have to adapt their language a little less there; indeed, behavior and uses of language are even promoted algorithmically (cf. Fielitz and

Marcks 2020). Others, as in the following example, are made unreadable by rating and ranking algorithms.

The occurrences in the US city of Ferguson in 2014 and 2015 did not appear in the virtual space of Facebook.[12] This is not because these incidents were not "spectacular"; on the contrary, the riot control after the murder of Michael Brown was martial. But Facebook's Edgerank algorithm filtered out the topic because, according to Facebook, it edits news according to personalized relevance. The Black Lives Matter protest in the aftermath of Michael Brown's murder thus became virtually invisible in this socio-technical network (cf. Tufekci 2014).

Especially in times of social unrest, socio-technical networks and spaces of action are increasingly observed, measured, and controlled, not only for the sake of functioning or for their own business interests, but also to measure the reactions of the public and to estimate the duration and severity of the associated protests. Thus, according to some social media researchers, the data extracted from social network analysis is quasi-isomorphic to the organizational structure of social movements. Given the central role of the Internet in social structures or movements such as in the early days of the Arab Spring or the Zapatista Movement in Mexico,[13] a map of network connections is, in effect, a map of the social and organizational relationships that constitute the most significant part of those movements (cf. Garrido and Halavais 2003, 2). Still others attempt to empirically capture

12- 18-year-old Michael Brown was shot and killed during a police check in Ferguson on August 9, 2014. A police patrol stopped him because he was walking on the road instead of on the sidewalk. During the discussion, a shot was fired from the patrol car. Brown fled and was shot in the back by a police officer. Michael Brown was unarmed and was black. Citizens of the city gathered for a vigil the very next day, most of them black. They were confronted by 150 police officers in armored gear. The atmosphere heated up and the situation got out of control. Street fighting and looting broke out. On August 11 and 12, the police used armored vehicles, stun grenades, smoke bombs, tear gas and rubber bullets against the angry crowd.

13- The Zapatistas in Mexico are an insurgent group of predominantly indigenous people who rose up against the government's neoliberal and neocolonial policies in 1994 and created an alternative autonomous space, characterized by self-management, grassroots democracy, collective ownership and distribution, gender justice, and a sustainable approach to nature.

the differences between offline and online movements, by examining the activities of formal and informal organizations and identity groups involved in protests for example (cf. Fowler and Steinert-Threlkeld 2016).

Messages from the jungle

When Colombian writer Gabriel García Márquez asked Subcomandante Marcos, the self-proclaimed spokesman for the Zapatista Army of National Liberation (EZLN) about the place of literature in his life, the latter replied that as a child he thought of language "not as a way of communicating, but of building something" (2001). Marcos conceived of his poetry and prose as weapons to mobilize readers for the rights of Mexico's indigenous population but also more generally against neoliberal economic policies and for autonomous self-government. Like Landauer, he relied on the power of the lyrical in political language. One of the ways he practiced this was in the way he created his hybrid literary language to explain the identity of the guerrillas and their goals, for example, by using colorful symbolism to describe the actors, but also to describe intertextual threads woven between his stories and the *Popol Vuh*, or poems and texts of García Lorca and Jorge Luis Borges.

Until his public farewell in May 2014, the Subcomandante communicated in irregular intervals with articles, letters, and poetic communiqués, which still circulated around the world by fax in the early days of the EZLN. Quite quickly, however, the ELZN networked digitally, reaching the Western industrialized nations as well as other emancipatory movements and NGOs in Mexico and around the world via the Internet. As an emancipatory movement, their organizational forms and goals have made them a model for many transnational social movements, such as the anti-globalization movement, the counter summits against G20 meetings, Anonymous, Reclaim the Streets, the São Paulo Forum, and many others (cf. Zimmering 2020).

The poetology of the Zapatistas emerged at a very interesting historical moment—not least in view of the mainstreaming of

the Internet. They set a technological infrastructure and also reflected it very consciously with their messages from the jungle (Woznicki 2020).
Spanish sociologist Manuel Castells wrote in his study *The Information Age* that what made the Zapatistas special, was their use of information technology to build an international network of solidarity (1997).

Black Lives Matter

As I began writing this article, riots spread from Minneapolis to cities across America. Yesterday, a police officer in Minneapolis killed George Floyd. An African American man who was paying at a kiosk with a counterfeit $20 bill. The owner called the police. Four officers came. One of them knelt on Floyd's neck until he stopped breathing. He died of the consequences a short time later.
The murder of George Floyd was certainly not planned by anyone. Like so much that happened in Minneapolis and the United States in 2020, it was not planned, but was somehow predictable. Yes, even predicted, and yet nothing was done about it. Because it was not foreseen.
In 2018, the information science research paper *Using Linguistic Cues for Analyzing Social Movements* attempted to use the emancipatory movement Black Lives Matter as a case study to analyze the relationship between traditional media such as news articles and the communication trajectories in Twitter that began with the hashtag #BLM and #Ferguson (Rezapour 2018). For this, the researcher Rezvaneh Rezapour accessed a dataset from the Center for Media and Social Impact in Washington (Freelon, McIlwain, and Clark 2016). In her methodology, Rezapour used an interesting aspect to ultimately apply the process of sentiment analysis to her dataset.[14] She assumed that people

14- Sentiment analysis (opinion mining) is a computational analysis of the views, attitudes, opinions and emotions of people on a subject or object. It is a subfield of text mining and refers to the automatic evaluation of texts with the aim of identifying an expressed attitude as positive or negative.

connect to movements and events by expressing their feelings and changing their language. She acknowledged, although she did not directly refer to current research, the research on linguistic alignment in text-based communication, "that people tend to adjust their language use to one another both in terms of word choice and sentence structure" (Wang, Reitter, and Yen 2017). She also drew on diverse studies from psychology (Campbell and Pennebaker 2003). Rezapour pursued the hypothesis that "individuals mostly use I and *my* in their everyday life to describe events or express their opinions. However, to emphasize their participation, people use more we and *our* to show their involvement in the process of change or movement" (2018).

In writing adversarial poetry, this would mean, in direct contrast, using "we/our" more often as pronouns to express one's political opinion and "I/my" to emphasize participation in a movement process.

According to Rezapour's research findings, the use of pluralism is widespread in the passages that have far-reaching effects or are more generally focused on society such as:

"'You think we WANT to protest? Nah. We wanna live. We protest because we are being slaughtered. #ferguson,' anonymous3" (2018). Individualism is more common in posts that present opinions or emotions: "'I'm in tears sitting here at my desk #ferguson,' anonymous4" or "'I can't believe what I'm reading... R.I.P #AntonioMartin #policebrutality #civilrights #BLACKLivesMatter,' anonymous5" (2018).

Four years earlier, the United States Department of Homeland Security was also interested in recording and analyzing the structures and individual actors in this emancipatory movement. According to reports by the research platform *The Intercept*, it evaluated social media accounts on services such as Facebook, Twitter, and Vine in order to obtain information about protesters' whereabouts (Joseph 2015). In this way, instruments of domination are networked and linked in botnets that automatically detect abnormalities and pass them on to the relevant authorities in natural language. Actors in emancipa-

tory movements in these new spaces therefore also need new strategies and tactics to protect themselves and their counterparts from possible repression. To communicate with one another they need to use anonymization and encryption tools, e.g., Tor[15] and PGP[16] or their own (poetic) symmetric encryption method,[17] at irregular intervals. This means consciously changing the common use of language by using alternative codes and techniques. Repressive instruments of domination preventively limit the political power of emancipatory movements, not only in the virtual world, but also on the streets, up to the point of the absolute political inability to move.

With the advent of information technologies and computer-assisted communication, there has also been a revival in the analysis of social networks. In disciplines such as the life sciences, economics, psychology, or the digital humanities, this computational modeling of text analysis processes has become a central technique for deriving reliable insights about social phenomena from data obtained through observation. These predominantly computational linguistic insights flow unobtrusively into applications for the private sector, government agencies, intelligence agencies, police departments, and the military. Quite often they are used for speech biometric procedures. At the turn of the millennium, it was precisely these procedures that successfully merged with cognitive technologies such as deep learning to extract people and political movements from socio-technical networks, identify them, and predict their next moves.

15- Tor is an anonymization tool for improving privacy and security on the Internet. It can be used to prevent services from tracking you, or to connect anonymously to news sites or instant messaging services that are blocked by local ISPs. Tor's services allow users to post to websites, blogs, forums, etc. without revealing their location or digital identity. The network has proven crucial for emancipatory movements around the world.

16- PGP (Pretty Good Privacy) is a program developed by Phil Zimmermann in 1991 for encrypting and signing data. His goal was to allow citizens and civic movements especially to securely exchange encrypted messages (even from access by intelligence agencies).

17- German political prisoners on the Isle of Men, for example, used Schwitter's Ursonate as an encoder to exchange information during World War II.

Today this must always be kept in mind when one is forced to move outside of predefined norms and rules, i.e. when one does not or cannot follow cultural-technical rules, or when one is consciously dedicated to not following and playing with these rules, the social, and the machine codes. Nevertheless, the wave of protest movements in the twenty-first century grows year after year. Despite all the new technologies of domination emerging in our millennium, protesters do not seem to stop reinterpreting them as a basis for hacking and appropriation, i.e., inscribing their own rules and their own language in technologies. The Black Lives Matter movement, for example, have attempted to establish an anti-sexist and antiracist vocabulary. However, the meaning vectors of pre-trained models are primarily trained with the broadest possible textual material from the Internet and thus are not attuned to the nuances of this vocabulary. They therefore cannot interpret out-of-the-box language usages that correspond to these new cultural norms (Hao 2020).

As Alicia Garza, a central figure of the BLM movement wrote "Hashtags do not start movements—people do" (2020, 8). It was Garza herself who made sure the #blacklivesmatter hashtag became popular on social media after the acquittal of the neighborhood watch coordinator George Zimmerman, who murdered Trayvon Martin in Sanford, Florida on February 26, 2012. In her recent book *The Purpose of Power: How We Come Together When We Fall Apart*, she writes about her own active politicization and how grassroots work in black neighborhoods in San Francisco.

When might unrest occur?

We can see from the history of global protest and resistance movements that most started on a small scale, very locally, with interpersonal events on streets and in cafes, in squats or schools, in cities and in forests. If we take the global Fridays for Future movement as an example, we find it is inseparable, at least in German-speaking countries, from the first forest occupations in Germany and the later professionally organized peaceful resistance actions of the

civil disobedience movement Ende Gelände. This included regular concerts, readings, VoKüs (people's kitchens), peace camps and forest walks attended by thousands of protesters who came to Hambacher Forst in North Rhine-Westphalia and planted trees and collected acorns with their children.

But, as is usually the case with new movements, only a few excerpts and snapshots exist from these early developments. Occasionally, songs and poems, fanzines, manifestos, minutes of meetings and collected notes emerge. Then there are magazines and brochures and countless other pieces of literature, some grey, some illegal, books, also leaflets and posters, badges, stickers and other devotional items. There are tweets, Facebook posts, WhatsApp, Telegram, forum posts, mails and mailing lists, blog posts, articles in alternative and encrypted channels, video clips, podcasts, talk shows and other audiovisual recordings, GPS data, timestamps, log files, and traces and information collected by the smartphones and computers of protesters while writing.

Archives and processed documents of protests and social movements of all kinds are created from this material. These virtual documents also constitute material to inspire our models of future society. Most of these archives have emerged from the movements themselves. They are more or less directly connected to them and based on the voluntary work of individuals or groups of people who collect them during the movement's activities or painstakingly reconstruct them afterwards.[18]

Other archives and databases have been created by intelligence agencies and the military as the basis of so-called crisis early

18- The Bibliotheks-Verbundkatalog antifaschistischer Archive is an example of such an archive in the German-speaking countries (http://bibliothek.antifa-archiv.org/). Other include Verzeichnis Freier Archive, Bibliotheken und Dokumentationsstellen in Deutschland (http://afas-archiv.de/verzeichnis-freier-archive/) or Portal der deutschen Umweltbibliotheken (http://www.umweltbibliotheken.de/). Others worth mentioning are Archiv des Informationszentrums Dritte Welt (https://www.iz3w.org/projekte/das-dritte-welt-archiv) in Freiburg im Breisgau, Antifaschistische Pressearchiv und Bildungszentrum (apabiz) in Berlin (https://www.apabiz.de/) and of course Bibliothek der Freien (https://www.bibliothekderfreien.de/), the largest library on anarchism in the German-speaking world.

warning systems. They are called automated event databases: ICEWS (Integrated Crisis Early Warning System),[19] a project funded by DARPA, and GDELT (Global Database of Events, Language and Tone)[20] by Kalev Leetrau (Yahoo) and Georgetown University are two such systems.

EMBERS AutoGSR is another system for creating automated event databases worth mentioning (Saraf and Ramakrishnan 2016). EMBERS (Early Model Based Event Recognition using Surrogates) was designed by ten institutions and over seventy academics to provide "anticipatory intelligence" in support of US national security decision making. The system was supported by the IARPA (Intelligence Advanced Research Projects Activity) OSI (Open Source Indicators) program. The objective of EMBERS was to forecast (predict) population-level changes using open-source data feeds, such as tweets, news/blogs, Wikipedia, and others. It has gone through various iterations over the years and expanded from monitoring Latin America to also covering countries in the Middle East and North Africa. EMBERS was "deployed" by security agencies for several years, but only as a research activity. In many cases, the system was able to predict a high percentage of civil unrest events, such as the impeachment of the president of Paraguay in 2012, the World Cup protests in Brazil in 2013, and the violent student protests in Venezuela in 2014 (Muthiah et al. 2016). The EMBERS system has produced forecasts since November 2012, "automatically emailing them in real-time to IARPA upon generation, which have been evaluated by an independent test and evaluation (T&E) team (MITRE)" (Ramakrishnan et al. 2014). Using human analysts, MITRE corporation compiled a gold standard report (GSR), a monthly catalog of events, by surveying newspapers for reports of civil unrest in 10 Latin American

19- ICEWS (2007) focuses primarily on monitoring, accessing, and predicting events of interest to military commanders and is now being further developed by Lockheed Martin Corporation.

20- GDELT focuses on capturing a large dataset of events in terms of both categories and geographic distribution. The goal is to capture a large number of events without worrying about false positives.

countries: Argentina, Brazil, Chile, Colombia, Ecuador, El Salvador, Mexico, Paraguay, Uruguay, and Venezuela.

AutoGSR, part of the EMBERS project, was developed to automate the validation of riots and civil unrest by using minimal human effort. The system processed data in Spanish, Portuguese, and English 24/7 for six months and encoded unrest in the Latin American countries listed above. It used ranking algorithms to detect new actor roles and for automated updates of their actor dictionaries. After these respective actors were recognized using the named-entity recognition method,[21] an algorithm placed them into the training dataset of a Word2vec language model. Based on this, a kind of role recommendation determined whether a news article (or other texts) might contain a future unrest in code, i.e., certain features are recognizable, such as "When might an unrest occur?" "Where?" "With whom?" and "Why?"

Literal untargeted adversarial black box attack

Adversarial hacking

> As we move into a world where all social, economic and political systems are to some extent technological, we need to extend our way of thinking. Come learn how to hack—and then defend—society's core systems: elections, the market economy, lawmaking, tax policy, journalism and more. (Schneier 2020)

Since adversarial poetry gradually focuses on established technologies (Word2vec) in NLP, it is possible to hack a relatively wide range of prediction applications from the outside without needing information such as the particular parameters or even having access to the respective model architecture. Instead, the authors can apply a so-called black-box attack. This type of attack does not manipulate (in our case) the model architecture in which the respective criteria for the machine production of meaning are anchored, but it changes the respective input x, i.e. the characters of our natural language texts. Likewise, in adversarial poems the interest does not aim at guiding the output of the respective model towards a specific goal but rather tries to scramble its machine interpretation so that next steps are incorrectly predicted based on the underlying data material (the text). It is therefore called an "untargeted attack."

However, not needing to know the respective parameters does not mean not having to know the structure of these models. As we have seen in the previous examples, we are well served by having at least a rudimentary knowledge of how AI prediction systems work; a knowledge of these models and algorithms in general, but also of their use in particular. In adversarial attacks, whether a black box attack or white box attack, it is necessary to know the structure of the models to be hacked very precisely. In this sense, the term "black box" should not be confused with its classical meaning, in which nothing is known about the interior, in our case the artificial neural structure of the system in question. We need to know the system and we get

to know it mainly through research papers: if, for example, an institute for machine learning shows us in their research that there are significant differences in the semantic representation of concrete nouns (in short, things that can be seen, touched, and felt) and abstract nouns (things that we can only think), we can exploit this knowledge (Pierrejean and Tanguy 2019). We can see in this specific research result that the increased use of abstract nouns, or mental terms in texts, destabilizes the word embedding space of language models. The poetic potential of these techniques was explored in previous subchapters. Employing these insights on the code level, specifically towards specific needs and against the functionality of the language model under attack, is called adversarial hacking or attack.

Research on adversarial attacks began to gain popularity through neural computer vision systems (Szegedy et al. 2014). However, adversarial poetry focuses on natural language texts. These types of hacks are also called paraphrasing attacks. They turn out to be harder than adversarial attacks on images as the model input usually consists of words that inherently form a discrete space. This means that a particular input x usually consists of discrete symbols such as characters or words. Hence, we cannot just take ten percent more or less of the word "anarchy" in a sentence, but we can play with its meaning and concept on several levels. Since there is no simple metric between two utterances, one of the biggest difficulties is replacing words using computational methods in such a way that the fluency of the language remains and the reader is still able to grasp the intended meaning. To do this, we need functions c in our code that guarantee that both expressions, i.e., the original input x and the misleading expression we insert, have the same meaning (semantics). Also, the original syntactic properties must be preserved (writing style, sentence structure, etc.), i.e.: $c(x,x')$. There are various research approaches and studies on how this can be accomplished. We will take a closer look at one of them in the following.

Thought vectors

A team of researchers at Stanford University devised a method that replaces words in a sentence that are in the same embedding space, in such a way that they produce large perturbations in the embedding space but not in the flow of the sentence, in short "Greedy search and counterfeited embedding swap" (Kuleshov et al. 2018). These perturbation effects change the predicted class of the text and can cause its complete reversal. To maintain correct grammar, only words that do not significantly change the probability of the sentence under a language model are changed.

The researchers first wrote a semantic identification method to detect semantic similarities in the text using thought vectors. Thought vectors act like word embeddings, only that they map sentences instead of words with similar meaning close to each other. These thought vectors were then transferred as average values of the individual word embeddings. Accordingly, the original thought vector v had to be similar or equal to the vector v' to be replaced. However, thought vectors do not capture syntactic similarities, only semantic ones. For example, all words in the sentence can be rearranged and still result in the same word vector average. To preserve syntactic similarity, the team superimposes different language models like LSTM (Hochreiter and Schmidhuber 1997) over the vector representations to guarantee that a grammatically diced sentence in x, for example, is also used in x' in the same diced way. Once the syntactic similarity is set, a special optimization algorithm finds and replaces the words that should ultimately mislead the system. For this purpose, the researchers implemented the Greedy Optimization Method in their hacking model.

Thus, a technical description of adversarial poetry could be as follows: "literal untargeted adversarial black box attack." While experimental literature and political poetry open up a syntactic playground, primarily in natural language and its social codes, adversarial attacks show us how we can use these language games specifically to hack repressive technologies implemented in society.

Conclusion

In the twenty-first century we have seen that as soon as social subjects and social movements aim for political visibility, they become inscribed in opaque layers of new technologies, whether this is their intention or not. The tactic of adversarial poetry presented in this article aims to show that this does not necessarily mean submitting to the laws of these technologies. In this sense, it serves as an interface between aesthetic practice and technological and political critique. Adversarial poetry is a space for individual and collective positioning vis-à-vis constitutive inequalities and the patterns of domination inscribed in modern liberal democracies and their instrumentalities.

But writing adversarial poetry requires experimentation and expertise on multiple levels to maintain the respective intentions of meaning and understanding for human readers. Not only because poetry is not the most common genre for political writing but also because the linguistic formation of spaces of meaning to be filled by the reader can no longer be found on the phenomenological level (cf. Heilbach 2000). At best, adversarial poetry opens up cultural spaces as places of expression where those to whom political spaces remain largely closed have an opportunity to articulate themselves: spaces of emancipation, empowerment, and visibility.

However, there is no question that without appropriate interfaces, this writing process will never be truly viable. For further elaboration, much more than just dialogue between political activists, cultural workers, and the respective research communities, such as AI research, computational linguistics, digital humanities, social network and movement research is needed. A tableau for this kind of collaboration needs to be actively formed. A tableau to initiate further debates about approaches of this kind of political subversion in think tanks and research labs as well as in alternative cultural centers, in theaters, and on the streets.

Perhaps in this kind of collaboration we can find a way towards a more participatory open space within and outside socio-technical spaces of action.

Reference List

Ambrosio, Chiara. 2018. "Gertrude Stein's modernist brain." In *Imagining the Brain: Episodes in the History of Brain Research: Progress in Brain Research* 243, edited by Chiara Ambrosio and William MacLehose. Cambridge MA: Academic Press.

Arendt, Hannah. 1981. *Vita activa oder Vom tätigen Leben*. Munich: Piper.

Bajohr, Hannes. 2016. "Das Reskilling der Literatur." In *Code und Konzept: Literatur und das Digitale*, edited by Hannes Bajohr. Berlin: Frohmann Verlag.

Barakhnin, V B, and I S Pastushkov. 2019. *Word reordering algorithm for poetry analysis*, J. Phys.: Conf. Ser. 1405 012009. Accessed April 8, 2022. https://iopscience.iop.org/article/10.1088/1742-6596/1405/1/012009

Brinnin, John Malcolm. 1964. *Die dritte Rose*. Tübingen: Henry Goverts Verlag.

Burroughs, William S. 1970. *The Electronic Revolution*. Göttingen: Expanded Media Editions.

Campbell, Sherlock R., and James W Pennebaker. 2003. "The secret life of pronouns: Flexibility in writing style and physical health." *Psychological science* 14(1): 60–65.

Castells, Manuel. 1997. *The Power of Identity, The Information Age: Economy, Society and Culture Volume II*. Cambridge: Blackwell Publishers.

Deleuze, Gilles. 1990. "Postscript on the Societies of Control." *L'Autre journal*, no. 1. Accessed April 8, 2022. https://theanarchistlibrary.org/library/gilles-deleuze-postscript-on-the-societies-of-control

Dick, Stephanie. 2013. "Machines Who Write." *IEEE Annals of the History of Computing* 35, no. 2, 88–87. Project MUSE. Accessed April 8, 2022. muse.jhu.edu/article/522043.

Duerr, Hans Peter. 1974. *Ni Dieu – ni mètre. Anarchische Bemerkungen zur Bewußtseins- und Erkenntnistheorie*. Frankfurt am Main: Suhrkamp.

Eco, Umberto. 1994. *Einführung in die Semiotik*. Munich: Fink.

Farland, Mary. 2004. "Gertrude Stein's Brain Work." *American Literature* 76 (1): 117–148.

Fielitz, Maik, and Holger Marcks. 2020. *Digitaler Faschismus. Die sozialen Medien als Motor des Faschismus*. Berlin: Duden-Verlag.

Fowler, James, and Zachary Steinert-Threlkeld. 2016. "Online and Offline Activism in Egypt and Bahrain." Accessed April 8, 2022. https://www.iie.org/Research-and-Insights/Publications/DFG-UCSD-Publication

Freelon, Deen, Charlton D. McIlwain, and Meredith D. Clark. 2016. *Beyond the hashtags: #Ferguson, #Blacklivesmatter, and the online struggle for offline justice*, Washington. Accessed April 8, 2022. https://cmsimpact.org/resource/beyond-hashtags-ferguson-blacklivesmatter-online-struggle-offline-justice/

Friedmann, Reto. 2019. *Musik- und Sprachperformance zur anarchistischen Utopie Gustav Landauers "Die Trommel passt sich zornig an."* Accessed April 8, 2022. http://www.textxtnd.de/theater/trommel/trommel_4.html.

Garrido, Maria, and Alexander Halavais. 2003. "Mapping networks of support for the Zapatista movement: Applying Social Network Analysis to study contemporary social movements." In *Cyberactivism: Online Activism in Theory and Practice*, edited by Martha McCaughey and Michael D. Ayers, 165–184. Abingdon: Routledge.

Garza, Alicia. 2020. *The Purpose of Power: How We Come Together When We Fall Apart*. New York: Random House Publishing Group.

Girnth, Heiko. 2010. *Dossier: Sprache und Politik*. Bundeszentrale für politische Bildung.

Goldsmith, Kenneth. 2011. *Uncreative Writing: Managing Language in the Digital Age*. New York: Columbia University Press.

Hao, Karen. 2020. *We read the paper that forced Timnit Gebru out of Google. Here's*

what it says. December 4, 2020. Accessed April 8, 2022. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/
Heilbach, Christiane. 2000. *Transformation – Lesertransformation, Veränderungspotentiale der digitalisierten Schrift*. Accessed April 8, 2022. http://www.dichtung-digital.de/2000/Heibach/30-Mai/
Heissenbüttel, Helmut. 1970. "Georg-Büchner-Preis-Rede 1969." Text + *Kritik* 25.
Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long short-term memory." *Neural Computation 9*, issue 8, 1735–1780. Accessed April 8, 2022. https://www.researchgate.net/publication/13853244_Long_Short-term_Memory
Jin, Di, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. *Is bert really robust? natural language attack on text classification and entailment*. Accessed April 8, 2022. https://arxiv.org/abs/1907.11932.
Joseph, George. 2015. *Feds regularly monitored Black Lives Matter since Ferguson*. Accessed April 8, 2022. https://theintercept.com/2015/07/24/documents-show-department-homeland-security-monitoring-black-lives-matter-since-ferguson/
Kirchner, Jutta. 2001. "Gertrude Steins 'Namenssprache' in Tender Buttons." PhiN - Philologie im Netz, Ausgabe 16/2001, 12–40. Accessed April 8, 2022. http://web.fu-berlin.de/phin/phin16/p16i.htm.
Kleiner, Marcus S. 2013. "On the Poetics of Pop Literature. Part 2: Burroughs, Fiedler, Brinkmann," pop-zeitschrift.de. Accessed April 8, 2022. https://pop-zeitschrift.de/2013/03/10/zur-poetik-der-pop-literaturteil-2-burroughs-fiedler-brinkmann-von-marcus-s-kleiner10-03-2013/
Klemperer, Victor. 1947. *LTI: Notizbuch eines Philologen*. Berlin: Aufbau Verlag.
Kuleshov, Volodymyr, Shantanu Thakoor, Tingfung Lau, and Stefano Ermon. 2018. *Adversarial Examples for Natural Language Classification Problems*. Accessed April 8, 2022. https://openreview.net/forum?id=r1QZ3zbAZ
Mall, Franklin. 1896. "The anatomical course and laboratory of the Johns Hopkins University," *Johns Hopkins Hospital Bulletin* 7. Accessed April 8, 2022. https://hdl.handle.net/2027/coo.31924069247371
Maniez, Claire, Ronan Ludot-Vlasak, and Frédéric Dumas, eds. 2012. *Science and American Literature in the 20th and 21st Centuries: From Henry Adams to John Adams*. Newcastle upon Tyne: Cambridge Scholars Publishing.
Marcus, Greil. 2009. *Lipstick Traces*. Hamburg: Rowohlt.
Márquez, Gabriel García, and Habla Marcos. 2001. *Cambio*, March 26, 2001. No longer available. http://www.revistacambio.com/
Melin, Charlotte. 1985. "Gertrude Stein and German Letters: Received, Recovered, Revised." *Comparative Literature Studies* 22, no. 4 (Winter): 497–515.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. *Efficient estimation of word representations in vector space*. Accessed April 8, 2022. http://arxiv.org/abs/1301.3781.
Mokrohs, Laura. 2018. *Dichtung ist Revolution. Kurt Eisner, Gustav Landauer, Erich Mühsam, Ernst Toller. Bilder – Dokumente – Kommentare*. Regensburg.
Morris, John, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin and Yanjun Qi. 2020. *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP*. Accessed April 8, 2022. https://arxiv.org/abs/2005.05909.
Muthiah, Sathappan, Patrick Butler, Rupinder Paul Khandpur, Parang Saraf, Nathan Self, Alla Rozovskaya, Liang Zhao, Jose Cadena, Chang-Tien Lu, Anil Vullikanti, Achla Marathe, Kristen Summers, Graham Katz, Andy Doyle, Jaime Arredondo, Dipak K. Gupta, David Mares, Naren Ramakrishnan. 2016. *EMBERS at 4 years: Experiences operating an Open Source Indicators Forecasting System*. arXiv.org. Ac-

cessed April 8, 2022. https://arxiv.org/abs/1604.00033
Nassehi, Armin. 2019. *Muster – eine Theorie der digitalen Gesellschaft*. Munich: C.H.Beck.
Paul, Christiane. 2015. *Digital Art*, London: Thames & Hudson Ltd.
Pierrejean, Benedicte and Ludovic Tanguy. 2019. "Investigating the Stability of Concrete Nouns in Word Embeddings." In *Proceedings of the 13th International Conference on Computational Semantics - Short Papers*, edited by Association for Computational Linguistics, 5. Gothenburg. Accessed April 8, 2022. https://www.aclweb.org/anthology/W19-0510
Ramakrishnan, Naren, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, Chris Kuhlman, Achla Marathe, Liang Zhao, Ting Hua, Feng Chen, Chang-Tien Lu, Bert Huang, Aravind Srinivasan, Khoa Trinh, Lise Getoor, Graham Katz, Andy Doyle, Chris Ackermann, Ilya Zavorin, Jim Ford, Kristen Summers, Youssef Fayed, Jaime Arredondo, Dipak Gupta, David Mares. 2014. *'Beating the news' with EMBERS: Forecasting Civil Unrest using Open Source Indicators*. arXiv.org. Accessed April 8, 2022. https://arxiv.org/abs/1402.7035
Regner, Freihart. 2006. "Zur Bedeutung Hannah Arendts für die (psychosozial-therapeutische) Menschenrechtsarbeit. Eine kritisch einführende Hommage." *Zeitschrift für Politische Psychologie* 14, no. 1+2: 141–170.
Rezapour, Rezvaneh. 2018. *Using Linguistic Cues for Analyzing Social Movements*, arXiv.org. Accessed April 8, 2022. https://arxiv.org/abs/1808.01742
Rißler-Pipka, Nanette. 2019. "In Search of a New Language: Measuring Style of Góngora and Picasso." *Romanische Studien*: 117–150. Accessed April 8, 2022. http://www.romanischestudien.de/index.php/rst/article/view/639
Rosen, Margit. 2004. "The Algorithmic Revolution - On the History of Interactive Art." ZKM Karlsruhe. Exhibition text.
Saraf, Parang and Naren Ramakrishnan. 2016. "EMBERS AutoGSR: Automated Coding of Civil Unrest Events." *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Accessed April 8, 2022. https://www.kdd.org/kdd2016/papers/files/autogsr_kdd16.pdf
Sassenhagen, Jona and Christian J. Fiebach. 2020. *Traces of Meaning Itself: Encoding Distributional Word Vectors in Brain Activity, in Neurobiology of Language*. Accessed April 8, 2022. https://www.mitpressjournals.org/doi/full/10.1162/nol_a_00003
Schiewe, Jürgen. 1998. *Die Macht der Sprache: eine Geschichte der Sprachkritik von der Antike bis zur Gegenwart*. Munich: Beck.
Schmidt, Arno. 1970. *Zettel's Traum*. Stuttgart: Stahlberg Verlag.
Schneier, Bruce. 2020. "Hacking Society." RSA Conference, February 27, 2020. Accessed April 8, 2022. https://www.youtube.com/watch?v=NVGtfVj_9Y0
Searle, John R. 1998. *Mind, Language and Society: Philosophy in the Real World. New York*: Basic Books.
Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. *Intriguing properties of neural networks*. Accessed April 8, 2022. https://arxiv.org/abs/1312.6199
Schwitters, Kurt. (1922) 1974. *Das Literarische Werk: Band 2: Prosa 1918–1930*. Cologne: DuMont Schauberg.
Stein, Gertrude. 1988. *Lectures in America*. London:Virago Press.
Stein, Leo and Albert Barnes. 1950. *Reise ins Selbst: Die Briefe, Papiere und Tagebücher von Leo Stein*. Edited by Edmund Fuller. New York: Crown.

Trogemann, Georg. 2010. “Code und Maschine.” In *Code – Zwischen Operation und Narration*, edited by Andrea Gleininger and Georg Vrachliotis, 41–54. Zurich: Birkhäuser.
Trogemann, Georg. 2016. “Von poietischen Prozessen und poetischen Maschinen.” Lecture at Kunstverein Kassel during the exhibition “The Paradox of Knowing Universals” by Ralf Baecker. Accessed April 8, 2022. http://www.georgtrogemann.de/ueber-das-machen/
Trogemann, Georg. 2019. *DAS 18. KAMEL und Die Habitate des Denkens – Über das Paradox, Technik in der Kunst zu lehren*. Accompanying text to Ars Electronica Panel discussion “100 Jahre Bauhaus: We are not alone,” September 5, 2019, Linz.
Tufekci, Zeynep. 2014. *What Happens to #Ferguson Affects Ferguson: Net Neutrality, Algorithmic Filtering and Ferguson*. Accessed April 8, 2022. https://medium.com/message/ferguson-is-also-a-net-neutrality-issue-6d2f3db51eb0#.4qidti8yi
Walter, William Grey. 1953. *The Living Brain*. London: Duckworth.
Wang, Yafei, David Reitter, and John Yen. 2017. “How emotional support and informational support relate to linguistic alignment.” *Social, Cultural, and Behavioral Modeling - 10th International Conference*, SBP-BRiMS, proceedings, Washington, 25–34.
Wittgenstein, Ludwig. 1984. *Werkausgabe in 8 Bänden - Band 4: Philosophische Grammatik*. Frankfurt am Main: Suhrkamp.
Woznicki, Krystian. 2020. Interview with the author. March 29, 2020.
Zhang, Xiang, Junbo Zhao and Yann LeCun. 2016. *Character-level Convolutional Networks for Text Classification*. Accessed April 8, 2022. https://arxiv.org/abs/1509.01626
Zimmering, Raina. 2020. “Digitale Rebellen: Die Zapatisten in Mexiko nutzen Wissenschaft und digitale Medien für ihren Widerstand. Dies zeigt sich auch in der Coronakrise.” *Junge Welt*, April 20, 2020: 12. Accessed April 8, 2022. https://www.jungewelt.de/artikel/376773.soziale-bewegung-in-mexiko-digitale-rebellen.html

Experimental approach for a literal untargeted adversarial black box attack

Christian Heck

In this approach, the goal will be to literally trick an NLP model so that the subjective information in an expression reverses its classification. In short, an opinion, emotion, or attitude about a topic or person that is normally interpreted by the machine as negative or bad will be "misinterpreted" as positive.

A framework called TextAttack is used for this purpose (Morris et. al. 2020). TextAttack is an open-source Python toolkit for adversarial attacks, adversarial training, and data augmentation in NLP. It was developed by the Qdata lab at the University of Virginia to allow both researchers and developers to test and investigate the weaknesses of their NLP models. To this end, code from more than 15 papers in the literature on NLP adversarial attacks has been implemented in the framework. Developers can use these as so-called "recipes" to generate a specific type of adversarial example: adversarial perturbations. Here, TextAttack iterates through a dataset (list of inputs to a model) and looks for an adversarial perturbation for each correctly predicted example.

As described in the chapter "Thought vectors," the following experimental approach uses the Kuleshov recipe to generate perturbations that are grammatically valid and semantically similar to the original input (Kuleshov et. al. 2018). The "input" is a scraped tweet from the CMSI dataset used by researcher

Rezvaneh Rezapour for her research (Rezapour 2018): “Do you think we WANT to protest? Nope. We wanna live. We protest because we are being slaughtered. #ferguson.”
According to her research, people connect to movements and events by expressing their feelings and changing their language, as described in the chapter “Black Lives Matter.”
The NLP model targeted in this experiment is an LSTM model (Hochreiter and Schmidhuber 1997), trained on the Yelp Review Dataset. The Yelp Review is a binary sentiment classification dataset containing 1,569,264 samples from the 2015 Yelp Dataset Challenge (Zhang, Zhao, and LeCun 2016).
The following code example shows how a sentence that (according to sentiment analysis) has %87 negative connotations can be turned into one with %64 positive connotation after an attack is performed on it. This can be done by exchanging only one word, when “think” becomes “imagine.”

```
whoami@machine:~$ textattack attack --model lstm-yelp --recipe kuleshov –
interactive
Loading datasets dataset yelp_polarity, split test.
Loading pre-trained TextAttack LSTM: lstm-yelp
Attack(
 (search_method): GreedySearch
 (goal_function): UntargetedClassification
 (transformation): WordSwapEmbedding(
  (max_candidates): 15
  (embedding): WordEmbedding
 )
 (constraints):
  (0): MaxWordsPerturbed(
    (max_percent): 0.5
    (compare_against_original): True
   )
  (1): ThoughtVector(
    (word_embedding): WordEmbedding
    (metric): max_euclidean
    (threshold): 0.2-
    (window_size): inf
    (skip_text_shorter_than_window): False
    (compare_against_original): True
   )
  (2): GPT2(
    (max_log_prob_diff): 2.0
```

```
      (compare_against_original):  True
     )
   (3): RepeatModification
   (4): StopwordModification
  (is_black_box):  True
)

Running in interactive mode
---------------------------

Enter a sentence to attack:

„You think we WANT to protest? Nah. We wanna live.“

Attacking...
%64) 1 <-- (%87) 0)

You think we WANT to protest? Nah. We wanna live.

You imagine we WANT to protest? Nah. We wanna live.


Enter a sentence to attack:

„We protest because we are being slaughtered. #Ferguson“

Attacking...
%73) 1 <-- (%72) 0)

We protest because we are being slaughtered. #Ferguson

We demonstrate because we are being slaughtered. #Ferguson
```

Reference List

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. “Long short-term memory.” Neural Computation 9, issue 1780–1735 :8. Accessed April 2022 ,8. https://www.researchgate.net/publication/13853244_Long_Short-term_Memory

Kuleshov, Volodymyr, Shantanu Thakoor, Tingfung Lau, and Stefano Ermon. 2018. Adversarial Examples for Natural Language Classification Problems. Accessed April 2022 ,8. https://openreview.net/forum?id=r1QZ3zbAZ

Morris, John, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin and Yanjun Qi. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. Accessed April 2022 ,8. https://arxiv.org/abs/2005.05909

Rezapour, Rezvaneh. 2018. Using Linguistic Cues for Analyzing Social Movements. arXiv.org. Accessed April 2022 ,8. https://arxiv.org/abs/1808.01742

Zhang, Xiang, Junbo Zhao and Yann LeCun. 2016. Character-level Convolutional Networks for Text Classification. Accessed April 2022 ,8. https://arxiv.org/abs/1509.01626