POWERED BY

ISD | Institute for Strategic Dialogue

CASM technology

Identifying Sock-Puppets on Wikipedia:

A Semantic Clustering Approach

Nestor Prieto Chavana, Chris Inskip, Carl Miller & David Weir

EXECUTIVE SUMMARY

The idea that information manipulation, even 'information warfare' poses a rising, even existential, threat to democracies is one of the most important of our age. Civic societal coalitions have sprung up to fight it. Global summits have been convened. New government task-forces have been created, as have academic centres, regulatory groups, and a fledgling industry of private-sector start-ups. Perhaps not since 9/11 have we seen such a fast and decisive pivot of so many different actors towards a single threat.

Malign activity has targeted a number of information environments, including every major social media platform: Twitter, Facebook, YouTube, Instagram, TikTok, standalone websites and many others. This paper, however, is dedicated to possible platform manipulation on a venue that tends to be much less researched than mainstream social media: Wikipedia.

This report presents work that set out to create, trial and evaluate a method to try to detect covert and organised manipulation of Wikipedia at scale. It aims to explore whether this method worked, and whether it might be useful if deployed more expansively.

The paper tests a technique called semantic clustering. Made possible only by the release of transformer-based pre-trained language models, it allows us to cluster editors on the basis of the meaning (the 'semantic content') of the edits that they make. The ambition was to see whether this might be useful to identify clusters of editors who had made suspicious edits.

The material focus of this paper is the English-language Wikipedia entry for the Russo-Ukrainian war, and 48 other pages about Ukraine that link directly to it. Our collection contains 1,988 editors making over 13,483 revisions to these pages. We collected both known malign editing behaviour and legitimate editing behaviour, to test whether they sometimes fell into different clusters.

Overall, once the process was complete, 176 clusters were identified. Most of the clusters, whether legitimate or not, really represent edits conducted on a single page or a small number of pages with a common topic. We found this technique to be a useful way of grouping together different accounts that make very similar edits on Wikipedia. However, it did not necessarily split editors by whether these edits were legitimate or illegitimate. Vandalism on one page can look very different from vandalism on another, most semantic clusters in our analysis likely relate to the themes and topics the editors discuss, rather than to whether the editing activity is within Wikipedia's policies or not.

We did identify some clusters that contained a number of single-controller 'sock-puppet' accounts, those using false identities for deceptive purposes. This typically happened where

an editor intended to make a consistent series of changes, and either used multiple accounts to do so to circumvent a ban, or to camouflage their activity. For instance:

- 'Cluster 1' has ten editors and five sock-puppets, three known to be linked to a single user. They focus on the introductory paragraphs of the 'Ukraine' article struggling over whether it is located in Central or Eastern Europe.

- 'Cluster 12' has eight editors. All four banned for sock-puppetry are linked to a single user. All the edits in the cluster are to the 'Casualties of the Russo-Ukrainian War' article, and are engaged in an edit war over specific casualty details including the personal details of affected individuals.

- 'Cluster 78' is composed of ten editors, eight of which are blocked for a variety of reasons such as vandalism, sock-puppetry and IP blocks.[1] Most edits are engaged in an edit war over the naming of Ukraine ("Ukraine" vs "The Ukraine").

- 'Cluster 100' is composed of nine editors and four sock-puppets, with the two most prolific linked to a single user. All edits from the sock-puppets are to the 'Russia–Ukraine relations' page and consist of updating statistics on the country comparison table in this page, including army sizes on each side.

Moving forwards, a form of detection using semantic clustering may be useful, where different accounts whose edits have extremely high semantic similarity are distinguished from all other editors for closer examination.

Automated analysis has important limitations, even on a platform as open as Wikipedia. Automatically analysing accounts is difficult due to the unstructured nature in which the data is produced and stored. To link sock-puppet accounts to users, this lack of structured data creates a need for manual navigation through users' pages and sock-puppet investigation archives in order to accurately identify these principal accounts.

It will be vital to link any automated analysis ultimately to the human-led investigations and sanctions regime that Wikipedia has already built. This includes Wikipedia Administrators, such as CheckUsers, who are empowered to conduct sock-puppet investigations on the platform. As the tradecraft of platform manipulation continues to develop on other platforms, it will be important to ensure that Wikipedia's largely volunteer community of defenders have the best tooling possible to protect the integrity of the platform.

[1] IP blocks are enforcement activities which prevent edits from either users with either a single or a number of IP addresses.

Background

Wikipedia has perhaps more claim than anything else to being the first digital wonder of the world. Over the two decades since its foundation, it has grown into the largest collection of human knowledge ever assembled, available to everyone, for free.

Over these two decades, however, we have also seen the rise not just of new sources of information, but new threats to information too. Platform manipulation, inorganic amplification, the exploitation of false identity, malicious automation, and coordinated inauthentic behaviour are all now observed and tracked across multiple information spaces for multiple reasons. They vary greatly in how they are conducted and by whom.

Different organisations have both defined and responded to information threats in different ways. The term 'fake news' began to rise in visibility during the US 2016 presidential election to describe spoofed news websites sporting fantastical headlines to hoover up American clicks.[2] As foreign state involvement became clearer, a more militaristic vocabulary arose, one of information operations, information warfare and of 'sub-threshold' or 'grey zone' conflict. Researchers sometimes also used terminology that links digital manipulation to longer-standing traditions, such as propaganda, active measures, or psychological warfare. The social media platforms themselves tried to avoid the minefield of content-based definitions entirely by focusing on behaviour, with their own designations of 'inauthentic' and 'coordinated inauthentic behaviour'. Joining all of that was yet another dictionary, this time the language of regulation, one of 'violative content' and 'online harms'.

Wikipedians too have arrived at their own way of defining information threats. As an open-source project, it is predicated on five pillars, the first being that Wikipedia is an encyclopaedia, 'not a soapbox, an advertising platform, a vanity press…nor an indiscriminate collection of information.[3] Some malign behaviour is therefore described as 'not here to build a Wikipedia'. This refers to behaviour that does not intend to contribute to Wikipedia in good faith, such as 'treating editing as a battleground', 'dishonest and gaming behaviours' (such as the use of sock-puppets), no interest in collaborative work, and having a long-term agenda inconsistent with building an encyclopaedia.

It is likely that Wikipedia, much like other platforms, has seen information threats become increasingly sophisticated. A number of Wikipedians with experience in responding to information threats were interviewed for this research but have been kept anonymous at their request.

"For the first ten years of Wikipedia, we mainly called it vandalism", said one discussant. However, as Wikipedia became larger, more visible and more important, new threats became

[2] https://www.bbc.com/news/technology-46136513
[3] https://en.wikipedia.org/wiki/Wikipedia:Five_pillars

clear. Sometimes individuals, groups or entire communities would clash over Wikipedia content for ideological reasons. These clashes often produced edit wars, where groups would constantly attempt to conform Wikipedia's content to their own point of view. This could be accompanied by threats, abuse and attacks against other Wikipedians seen to stand in their way.

In some cases, Wikipedias could be subject to 'community capture', where a certain political or nationalistic group could reshape a Wikipedia page according to their own ideological worldview, unopposed. The Croatian-language version of Wikipedia, for instance, was for over a decade edited to align its content with the Croatian far right, whitewashing crimes committed by Croatia's Nazi-allied Ustashe regime during the Second World War.[4]

Interviewees most often pointed to another sort of threat on Wikipedia however, and that was 'undisclosed paid editing'. This is commercially-motivated activity to change Wikipedia which, because undisclosed, sits outside of Wikipedia's guidelines. Interviewees stressed how undisclosed editing had increased in sophistication over time. From individual freelancers, larger-scale organisations have reportedly emerged to sell undisclosed paid editing services. "Undeclared [editing] has expanded over the years. I now see conversations about large numbers of money. I don't know if that's true…. if you can get say 10k for an article, you can get a team of people building up reputations, you can have peer to peer cooperations.. You could have VPNs. You could get around [defences]" said an interviewee.

> *"We're in a different world than where we were in 2007/8. There are botnets. There are people considering how to put certain information out into web in certain ways - there are information wars. And you're absolutely right, Wikipedia is the place to go. It's right up there as one of the biggest sites in the world."*
>
> *Respondent*

Undisclosed paid editing is conducted for a variety of reasons. "You can split disinformation into two camps" said one administrator with a long experience of confronting undisclosed paid editing. "There's commercial. That's PR, whitewashing, and pushing allegations onto rivals." "Then", he added, "there's the political. Editing to affect elections, go after campaigns, target an ethnicity. These are ideologically-driven bad actors".

The 'Orange Moody' case in 2015 was, said another respondent, "when the community woke up and smelled the coffee. This was the discovery of 381 sock-puppet accounts operating an undisclosed paid editing ring.[5] "This wasn't just random individuals" the discussant continued, "it was organised, almost like a criminal enterprise - they use spammy tactics to

[4] https://upload.wikimedia.org/wikipedia/commons/1/14/Croatian_WP_Disinformation_Assessment_-_Final_Report_EN.pdf
[5] See https://en.wikipedia.org/wiki/Operation_Orangemoody

promote their services, they often don't deliver what they're offering." "Some of the organisations", another continued, "are actually criminal - arms of larger criminal enterprises. One based out of Pakistan -that also did fake degrees. Online crime. The organisation itself was deemed to be criminally subverted. Fake credentials. It's a fairly nice and easy thing to try."[6]

Wikipedia's response to these threats is also unlike any social media platform. Wikipedia is protected by a volunteer community who work within a 'user rights' regime to hold and use various powers. CheckUsers are the only people who can see the IP address and User Agent who is associated with each edit, browser information, etc; they are basically the best attempt at capturing our fingerprint online. Overseeing all of this is the Arbitration Committee, or ArbCom, the final binding body - who will look at behaviours within an area, investigate the 'bad actors', and then give the administrators within the area extra tools to deal with it - i.e. entire topic-level blocks for regions of Wikipedia. Processes already exist on Wikipedia to investigate suspicious activity on a case-by-case basis; these are known as 'sock-puppet investigations' and are often conducted by users with a privileged capacity to view technical information about any editor, known as a 'Check User' power.

Respondents did not know the extent to which undisclosed paid editing activity interacted with either state actors or geopolitics. All agreed that it was possible, either through a commercial arrangement between state actors and undisclosed paid editing providers, or through the adoption of the same methods by in-house actors.

[6] https://en.wikipedia.org/wiki/Operation_Orangemoody

This Report

This report presents work that set out to create, trial and evaluate a method for detecting covert and organised manipulation of Wikipedia at scale, that targets the Wikipedia page for the Russo-Ukrainian war and a series of other pages about Ukraine linked to it.

The objective of the project was to apply a detection capability that ISD and CASM have increasingly deployed across other platforms to try to identify suspicious behaviour on Wikipedia. This a process made possible only by the latest generation of Natural Language Processing technology called 'semantic mapping'. Explained in more detail below, semantic mapping allows researchers to map Wikipedia edits on the basis of their meaning, or semantic content, and then editors, through averaging their edits, on the same basis. Clusters of closely located editors would, according to this method, have made semantically similar contributions to Wikipedia, whereas those far apart would have contributed in different ways.

Researchers use this method to represent the text in a semantic space, where texts with similar meanings are close to each other, and those with different meanings are situated further away. This allows for a visual representation of editors already banned by Wikipedia, and those that are not, which provides an opportunity to interrogate the differences in their behaviour. If it can elucidate these differences, we hope it might then open up possibilities for some similar approaches to be integrated into the investigatory and enforcement activities that Wikipedians conduct.

As with research of platform manipulation across other platforms, we recognise it will often be impossible to truly discern the motivations of online actors. Some activity may be commercially motivated, others conducted by a state, and some conducted by ideologically-motivated actors who genuinely believe what they are doing is right.

This report is not concerned with discerning the objectives of these actors, rather it examines editing behaviour across these pages that is defined by Wikipedians as 'not here to build a Wikipedia'. The method employed to isolate this type of activity is detailed below.

Step One: **Collect Data**

The focus of the paper is on editing behaviour within 48 pages related that (a) were about Ukraine, or (b) were directly linked to the English-language Wikipedia page on the Russo-Ukrainian War. The focus of this analysis was on two sets of data related to these pages. First, 855 editors were identified who had been blocked by Wikipedia's enforcement

processes, making a total of 4,076 edits to one of these 48 in-scope pages.[7] The most common reason given was for sock-puppetry, followed by IP blocks (which are also commonly imposed to prevent a single user making multiple edits without an account).

Block Reason	Count
Sock-puppet	322
IP Block	191
Other	161
Vandalism	106
Disruptive	106
Not Here	53
Edit Warring	20
Harassment	15

Table 1: Count of editors per block reason

The second dataset collected was a broadly comparable control sample. This was a set of 1,133 unblocked editors who made a total of 9,407 edits to the same set of pages.

Step Two: **Semantically Represent Edits from Blocked and Unblocked Editors**

The next step was to create what is called a semantic representation of all the edits. This serves to translate the raw language of the text in the edit into a representation that captures the deeper meaning of it. It enables the use of natural language processing techniques to further analyse and compare the edits made by different editors.

We create these representations using the latest generation of natural language models based on the transformer architecture. The base model has been pre-trained on vast amounts of online text and then fine-tuned to specialise it at creating meaningful representations of sentences.[8] Transformer-based pre-trained models are able to create contextual semantic representations that not only look at individual words, but also their surrounding context in order to better identify their meaning. This results in richer and more nuanced representations of the source text.

As discussed in greater length in the annex to this report, Wikipedia data poses unique challenges to this process. Rather than messages or posts, Wikipedia edits are 'diffs', which can be comprised of any number of both additions and subtractions from a Wikipedia page. The actual meaning of any edit is not contained just within the diff itself, but the surrounding page that is being changed. For example, changing a single word in a sentence can alter the

[7] N.B editors may be tagged with multiple block reason.
[8] https://huggingface.co/sentence-transformers/all-distilroberta-v1

meaning significantly if it shifts the tone or intent of a paragraph. To account for this, we incorporate as much of the context around a change as the Wikipedia API returns in its responses.

Step Three: **Cluster Editors**

Next, we clustered editors together based on the semantic similarity of their edits. To do this, we first create a representation for each editor by averaging all of their contributions. The intuition behind averaging all contribution representations to serve as the editor representation is akin to finding the "centre of mass" of the information present in the edits. The resultant representation aims to capture the net contribution of the editor to the set of pages under analysis. Therefore, for two editors to be considered similar and thus be grouped together, they will have made a similar net contribution.

The clustering is then performed using a pipeline of commonly used tools in the realm of data analysis and machine learning.[9] We perform an additional process known as 'parameter tuning' in order to prioritise the creation of clusters that are homogeneous in regard to the block reasons of the included editors, and set a relatively small minimum size of five. With this approach, 176 clusters were created overall, as presented below.

Figure 1: Plot of editor clusters (with clusters marked by colour)

[9] https://huggingface.co/sentence-transformers/all-distilroberta-v1

To support cluster investigations, we built an interactive dashboard to display and manipulate the data. This dashboard provides functionality to explore the clusters generated by the clustering process and ranking clusters based on blocked proportions and specific block reasons. It makes it possible to select groups of editors and review their metadata, see the metadata for revisions by these editors, and it shows the textual changes they have made in a readable format. The dashboard also provides links to the corresponding Wikipedia user and revision pages, in case more context is required for analysis. Screenshots of the dashboard can be found in the Appendix.

Step Four: **Characterise Clusters**

In order to understand the behaviours captured through this clustering process, we qualitatively characterised a set of clusters. Given the large number of clusters, we prioritise this work by ranking clusters according to the proportion of blocked users for different reasons and focus on the top ten. We first perform this analysis by ranking clusters based on the proportion of users blocked for any reason. We then repeat the process for the top ten clusters ranked by the proportion of users blocked for sock-puppetry, and for vandalism.

In our detailed cluster analysis, we pursued two objectives. Our first aim was to identify the overarching topic or theme that defined each cluster. The second, related goal was discerning the underlying reasons for the cluster formation, finding any shared characteristics among accounts or their edits that led to their grouping.

First, we performed an examination of the editors within each cluster, including the number of editors in the cluster and their block status. For those facing blocks, we looked into the reasons for their blocking, searching for any commonalities. For instance, a shared administrator blocking multiple editors simultaneously might indicate a coordinated effort or the discovery of a sock-puppet network. We also looked at the diversity of block reasons in the cluster. Having a high proportion of a single block reason compared to others would suggest that the characteristics that led to the formation of the cluster may be representative of that block reason. For editors that were blocked for sock-puppetry in particular, we analysed their user pages, block logs, and any related sock-puppet investigations in order to determine whether the same user was behind them.

To understand the nature of the edits in a cluster, we surveyed the pages that attracted edits from the clustered accounts. We looked at the diversity in the edited pages, to determine the degree to which a cluster was focused on a specific topic. When relevant we also looked at editors' revision history outside of the sample pages, to identify whether any patterns in their behaviour in the collected data were also present outside.

Finally, we analysed a sample of revision diffs made by the editors in each cluster. We attempted to determine the characteristics of the changes they were making, while searching for topical or thematic commonalities between the edits.

RESULTS

At the highest level, there is no clear visual distinction between blocked and unblocked users. The semantic position of an editor, we find, is more likely to represent the overall themes that they contribute towards rather than whether those contributions are legitimate or not. This means that the method cannot be used to simply create an overall mapping that clearly identifies violative behaviour.

However, through a deeper analysis of the individual clusters, we find six of special interest. These are dense, small clusters of sock-puppet accounts where multiple accounts are known to belong to a single user. Four of these clusters contained named accounts, and an additional two contained groups of IP accounts that all originated from the same range and were making similar edits, suggesting sock-puppetry.

Figure 2: Plot of editor clusters, highlighting sock-puppet clusters

It should be noted that these clusters (in bold below) were amongst the highest in terms of their proportion of blocked editors overall, and also had some of the highest proportions of known sock-puppet accounts.

Top ranked cluster of users blocked for any reason		**Top ranked cluster of users blocked for sock-puppetry**		**Top ranked cluster of users blocked for vandalism**	
Cluster	% blocked	Cluster	% blocked for reason	Cluster	% blocked for reason
0	**100.0**	62	60.0	130	36.4
130	90.9	148	60.0	59	33.3
39	87.5	6	54.6	61	33.3
27	83.3	**1**	**50.0**	45	30.8
61	83.3	**12**	**50.0**	28	20.0
25	**80.0**	104	50.0	36	20.0
62	80.0	76	50.0	48	20.0
78	**80.0**	152	44.4	62	20.0
144	80.0	**100**	**44.4**	78	20.0
148	80.0	171	42.9	82	20.0
156	80.0	110	42.9	94	20.0
				119	20.0
				163	20.0
				172	20.0

Table 2: Analysed editor clusters

Given that the grouping of editors is based entirely on the content of their edits, it is interesting that these sock-puppet accounts are clustered together and we chose to focus on these clusters for further analysis to understand the characteristics that led to their grouping.

Cluster 1. This cluster has ten editors, five of which are blocked for sock-puppetry, three controlled by the same user. All sock-puppet accounts of this user that were collected were in this cluster. The edits in this cluster centre on the introductory paragraphs of the '*Ukraine*' article. The edits include users changing the stated location of Ukraine from Eastern to Central Europe, as well as an edit war to remove references to the CIA World Fact-book. The figure below shows examples of edits introduced by the sock-puppet accounts, all centred around changing the location of Ukraine from Eastern to Central Europe.

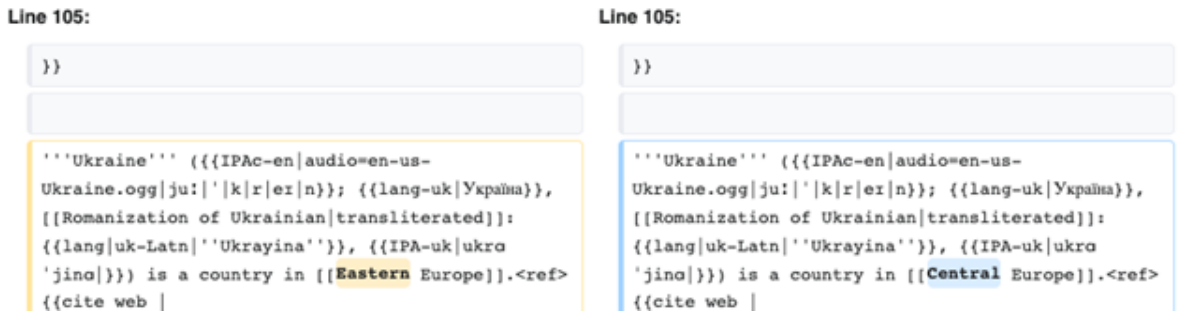

Figure 3: Sample revision by editor in Cluster 1, changing the location of Ukraine

Cluster 12. This cluster is formed of eight editors. All four of the accounts banned for sock-puppetry belong to the same user. Edits in this cluster are all to the '*Casualties of the Russo-Ukrainian War*' page and editors appear engaged in an edit war over specific casualty details including the personal details of affected individuals. The figure below shows examples of such edits.

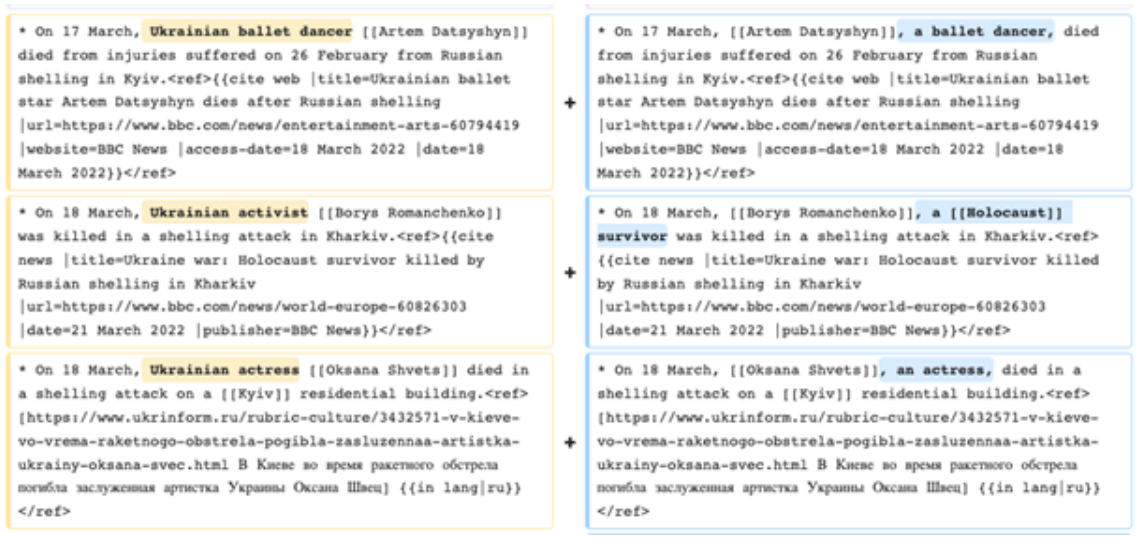

Figure 4: Sample of revision by editor in Cluster 12, modifying personal details of casualties

Cluster 78. This cluster is composed of ten editors, eight of which are blocked for a variety of reasons such as vandalism, sock-puppetry and IP blocks. Most edits seem engaged in an edit war over the naming of Ukraine ("Ukraine" vs "The Ukraine"), as shown in the figures below. Though not confirmed to be sock-puppets through account analysis, it is also possible to find IP accounts in this cluster making similar edits.

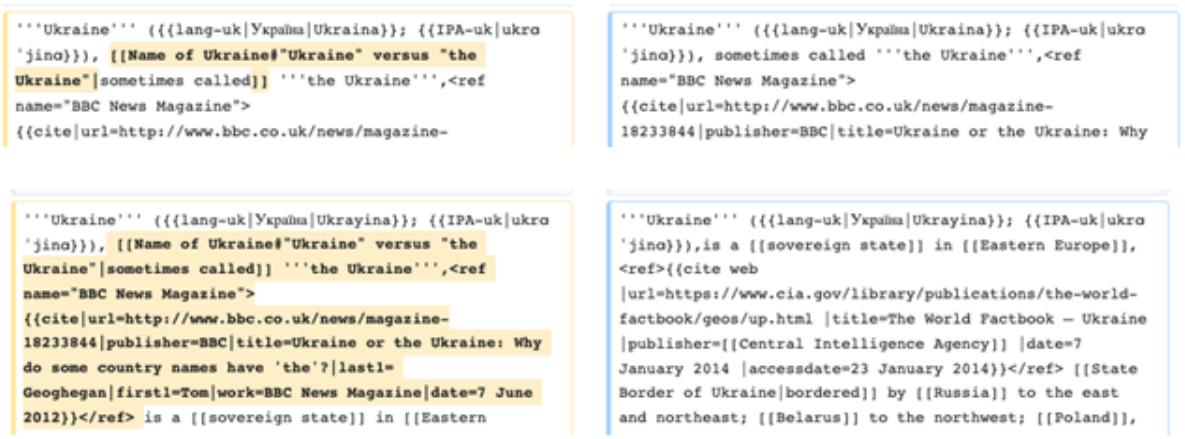

Figure 5: Sample of revisions by editor in cluster 78, editing the naming of Ukraine

Cluster 100. This cluster is formed of nine editors, four of them being blocked for sock-puppetry. A large proportion of the edits come from two editors who are sock-puppets of the same user. Most of the edits by accounts in this cluster are updating country statistics such as GDP, population and geographical extension, as well as data relevant to the war such as number of combatants and armament on each side. Most of these edits are to the '*Russia–Ukraine relations*' and '*Armed Forces of Ukraine*' articles.

Figure 6: Sample of revision by editor in Cluster 100, changing country statistics in article table

Cluster 0. Cluster 0 is formed of seven editors making only a few edits each. All editors in this cluster are IP accounts originating from the same range, and all of them have been blocked through an IP range block. All the edits in this cluster have been part of individual revisions. All edits are to the '*Russia–Ukraine relations*' page, making almost the exact same change, which consists of removing a link to the Wikipedia article about '*Malaysia Airlines*

Flight 17'. This article describes the shooting down of a passenger flight by Russian-controlled forces, which happened on 17 July 2014.

Users are repeatedly removing the same span of text, which indicates there are other users adding this text, suggesting an edit war going on in this section of the article. These IP users were likely used to get around restrictions meant to prevent edit wars, such as the three-revert rule. Most of the accounts in this cluster only have a handful of edits to pages outside of the Ukraine sample we collected. A review of a sample of these edits shows they consist of very similar changes, removing references to this same incident from a variety of other articles. On this basis, we can speculate that these accounts are engaged in coordinated editing intended to remove references to this event from Wikipedia.

Figure 7: Sample of revisions by editor in Cluster 0, removing link to Malaysia Airlines Flight 17 article

Cluster 25. This cluster is formed of five editors, four of which are blocked. All blocked editors are IP accounts blocked through an IP range block for disruptive editing. In the context of the seed network, edits in this cluster are to the '*International sanctions_during the 2022 Russian invasion of Ukraine*' article. The contributions are additions of Russia-Switzerland relations, particularly on the position of Switzerland as an important destination for rich Russians to manage their wealth.

This IP range is also noteworthy for being blocked for adding links to videos that admins have determined to "contain antisemitic and conspiracist content" as indicated in its block log. Figure 8 below show edit summaries added by IPs in this same range, though not the ones in this cluster.

Revision as of 19:11, 24 April 2022 (edit) (undo)

172.58.236.74 (talk)

(...this happened once in 1998 when they refused to settle with Jewish groups regarding Nazi Gold...(the TREATH that is (only..))...so the immediately accepted ALL the terms of the "settlements" (they are indeed all SUBSERVIENT slaves to Jewish- Cabalistic-Talmudic interests at the end (& despite of all APPEARANCES) ... the rest is SHOWMANSHIP only!:-))

(Tags: Reverted, Mobile edit, Mobile web edit, review edit)

Next edit →

Revision as of 03:41, 9 May 2022 (edit) (undo)

172.58.239.101 (talk)

(→Connection to illegal activities: Swiss people will bear the burden promoted by corrupt Swiss legislators (e.g. & today UBS is mostly FOREIGN owned. ..check! & HOW did this happen?)... https://m.youtube.com/watch?v=xDB2rG4rTKw)

(Tags: Mobile edit, Mobile web edit, review edit)

Next edit →

Figure 8: Samples of revision logs by editors in cluster 25, adding links to antisemitic videos.

LIMITATIONS AND CAVEATS

There are a number of important limitations to this project that are essential to be recognised alongside the results that this report presents. These are presented below.
The report only focused on text. While Wikipedia is primarily a text-based platform, images are present. Manipulation may occur through the use of misleading images, as well as altering captions and descriptions of existing images. The method used in the report is purely text-based, however, image captions are included within the text-based 'diffs'.

Reliance on known bad-actors. The report, of course, focused on known-bad actors and compared them with unblocked editors. This methodology will not recognise forms of information manipulation or malign activity that cannot be detected by Wikipedia's current enforcement processes and procedures, which may be systematically different from known violative behaviour.

Blocked editors are not necessarily information manipulators. We assume throughout the report that that blocked editors provide indicators of manipulation. However, not all blocked users will be information manipulation actors. For example, while rare, some users are voluntarily blocked to take a break away from the platform. Others may be banned for conduct that is not connected to editing behaviour.

'Diffs' pose a challenge to textual processing. Diffs are a form of data that is not often encountered by a language model during training, and it is uncertain how well a model can appropriately analyse them. In addition, some diffs can be very long. The language model used for this analysis can incorporate up to 400 English words. Longer sequences of text are truncated by the model, and thus not used for identifying similar content. Therefore, an assumption made here is that there is enough context in that starting span to provide useful groupings. In addition, this inclusion of surrounding context can itself result in the specific change being overshadowed by the context in which it was made. An empirical analysis of the performance of different models fell outside the scope of this investigation.

The report focuses only on main article content. We have focused on editing behaviour that can be observed in the text of articles in Wikipedia. However, it is also possible that information manipulation can occur in other forms on Wikipedia, such as on user and article talk pages, revision descriptions, and on pages dedicated to Wikipedia's policies, investigation, arbitration, and administrator elections. These areas of Wikipedia are extremely important but fell outside of the scope of this specific project.

The analysis is not predictive: This analysis is performed on a static snapshot of the data, and so does not aim to be a predictive approach. The process aims to use blocked editors,

known to be unwanted on the Wikipedia platform, to facilitate in describing the types of edits and users editing a set of pages.

TECHNICAL ANNEX

Terminology

- **Diff** - Short for "difference," refers to the changes between two versions of a page.
- **Editor** - A user who contributes content or modifications to a page.
- **IP editor** - An editor who makes contributions without logging in, identified by their IP (Internet Protocol) address.
- **Named editor** - An editor who has registered and logs in to a platform with a unique username. Their contributions are associated with this username rather than their IP address.
- **Revision** - A revision, also called version, is a single entry in the history list for a page.
- **Contribution -** In this work we refer to a contribution as a single unit of deletion, addition, or change (addition and deletion) made to generate a new revision. A revision is composed of one or more contributions.
- **API** - An Application Programming Interface (API) provides a set of endpoints for programmatically communicating with the platform (i.e., Wikipedia), e.g. for querying the platform for data relating to specific users or pages.
- **Embedding** - A numerical representation of a data point.

Data Collection Methodology

We selected the Wikipedia article on the Russo-Ukrainian war as our seed page. From there, we expanded to 48 linked pages primarily centered on Ukraine, identified by the presence of 'ukraine' or 'ukrainian' in their titles. We gathered the revision history and associated metadata for these pages. Subsequently, we obtained metadata for every editor responsible for these revisions. For blocked editors, we collected the diffs of their edits. Lastly, we sampled several non-blocked editors from this group and gathered the diffs of their respective edits.

It should be noted that this Ukraine-focused sample was derived from a larger set of collected data based on all pages linked from the seed page. This larger set was stratified to contain an equal proportion of blocked and non-blocked editors before filtering to a Ukraine-focused set. Two effects of this collection process to note are (1) there is a higher proportion of blocked users in this dataset than would be expected if collecting all editors of a page, and (2) while still reasonably balanced, the proportion of blocked editors is not exactly 50%.

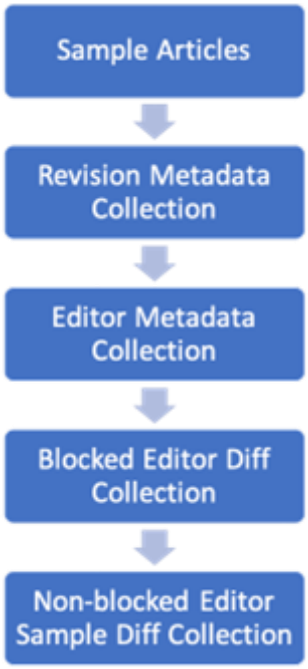

Figure 9: Data collection process flowchart

Edit Representations

Representing the edits made by users poses unique challenges compared to other many other platforms. These challenges are primarily related to the way in which data is structured, which stems from the collaborative nature of Wikipedia. Specifically:

1. The edits are "diffs", which represents the changes made to the article. These diffs not only contain the pure content changes but are also intermingled with an extensive amount of markup used to format the Wikipedia pages.
2. Understanding the significance of a change often necessitates surrounding context. For example, a simple change to a sentence can alter the meaning significantly if it shifts the tone or intent of a paragraph.

Figure 10: Example of diff collected from Wikipedia API

We used Wikipedia's Compare API[10] to collect the differences in the text of an article before and after a revision. The API returns this data in the form of a table where one column contains the text before the revision, and another after the revisions, according to Wikipedia's detection of what has been changed. This often includes some context before and after the actual text that was changed, for example including the full sentence when only a word was changed. The figure below shows an example of a revision diff returned by the API.

In order to get a complete view of the meaning of an edit, we incorporate the full context provided by the Wikipedia API. This allows the model to create a richer representation of the text but comes at the cost of including content that may have not been produced by the editor. For example, in the figure above, the editor has changed only a single word in a sentence. In this case, we use the full sentence to create a representation.

Each diff table collected contained multiple rows, corresponding to the multiple changes made in a single revision. The revisions were split into individual contributions based on these rows. Exploratory analysis found that in some cases, the tables contained records where no change had been made. It was also found that, while markup existed to indicate the changes made to a text in the form of deletions and additions, it was not consistently present. It was also difficult to line up changes, since the number of additions to a text did not necessarily match the number of deletions. To be able to use the collected diffs for creating embeddings of the individual contributions, some cleaning and pre-processing was required.

In order to find the exact changes made in each contribution, they were first parsed and tokenized using the Python Spacy[11] library. We then applied Python's difflib[12] package, a

[10] https://www.mediawiki.org/wiki/API:Compare
[11] https://spacy.io/
[12] https://docs.python.org/3/library/difflib.html

built-in library used for comparing sequences. This produced the exact words or characters that were removed and/or added in each contribution. We discarded contributions where the changes included only punctuation or whitespace. Contributions were labelled based on whether they were an 'addition', a 'deletion', or a 'change', which includes both an addition and a deletion.

We use the "all-distilroberta-v1"[13] language model from the SentenceTransformers Python library to create a numerical representation, or embedding, of the state of the text before and after a given change. The model used for this has been trained at semantic similarity tasks such as paraphrase detection, which is the task of identifying sentences that express the same or similar meanings, despite using different wording. The effect of this training is that texts that have a similar meaning receive similar representations. The main benefit of this is that when applying clustering algorithms, similar texts are more likely to end up being grouped together. After this step, we concatenate the representation of the before and after stats, creating a representation for the change as a whole.

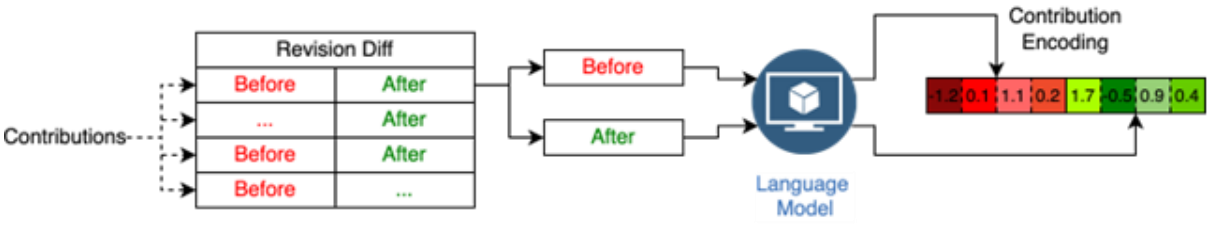

Figure 11: Contribution encoding process

Classifying Block Reasons

Editors can be blocked for many different reasons. We first summarise a number of the most relevant reasons for why a user may be blocked from contributing to the platform. We then condense this set of block reasons into higher-level labels to simplify analysis.

To obtain the reason why a user has been blocked, we can make a call to the Wikipedia Users[14] API, requesting the user's details. Along with indicating whether the user is blocked and the block duration, this includes a free-text field indicating the block reason. This text often includes a codified label in the form of a Wikipedia tag, such as "[WP:SOCK]" to indicate the user is a sock-puppet. These tags are shortcuts to administrative pages in Wikipedia which describe editing policies. However, including this information is not required, and we find this is omitted in many cases.

[13] https://huggingface.co/sentence-transformers/all-distilroberta-v1
[14] ttps://www.mediawiki.org/wiki/API:Users

When Wikipedia tags are included, many are synonymous or otherwise highly related. For example, [WP:SOC], [WP:SOCK] and [WP:SOCKPUPPET] all reference sock-puppetry, and [WP:VD], [WP:VAND], [WP:VANDAL], [WP:VANDALIZE], and [WP:VNDL] all refer to vandalism. We use regular expressions to search for Wikipedia tags in block reason texts. We extract all the unique tags and manually code them into the appropriate block reason based on the text in the tag and/or the administrative pages the shortcuts link to.

Clustering Editors

For each editor, we take the average of the representations of all their respective contributions. This results in a single 1,536-dimension representation for each editor. Our intuition is that the resultant editor representation will capture the editor's net contribution to the set of pages under analysis. Therefore, for two editors to be considered similar and thus be clustered together, they will have made a similar net contribution. We expect that many nuances or unique characteristics of individual edits will be lost in favour of capturing the editor's broader net contribution.

Clustering pipeline: We cluster these editor representations using a pipeline of UMAP and HDBSCAN. HDBSCAN is a widely applicable density-based clustering approach, though, as with many density-based approaches, it tends not to work well with dimensionality beyond around 50 (due to the 'curse of dimensionality', where the notion of density vanishes as the number of dimensions increases). To mitigate this, dimensionality reduction techniques such as UMAP are often applied to the representations prior to applying HDBSCAN.

Hyperparameter selection: In order to select parameters for UMAP and HDBSCAN, we performed a grid search over several combinations and selected the configuration that produced the highest homogeneity score. This metric is higher when clusters contain only members of a single class, with the different classes being the editor's block reason. This encourages the model to produce clusters containing users blocked for mostly the same reason.

Cluster outliers: The HDBSCAN clustering algorithm is not required to assign every data point (editor) to a cluster, and often assigns many points as 'outliers'. Outliers are typically points located in regions of low density and deemed by the algorithm to be not representative enough of any one identified cluster. We exclude outliers from analysis, though we note that outliers may include both (i) editors that are very different from the other editors (i.e., their edits are sufficiently unique), as well as (ii) editors that are situated 'between' clusters, not having enough similar editors to form a distinct cluster.

Analysis Dashboard Screenshots

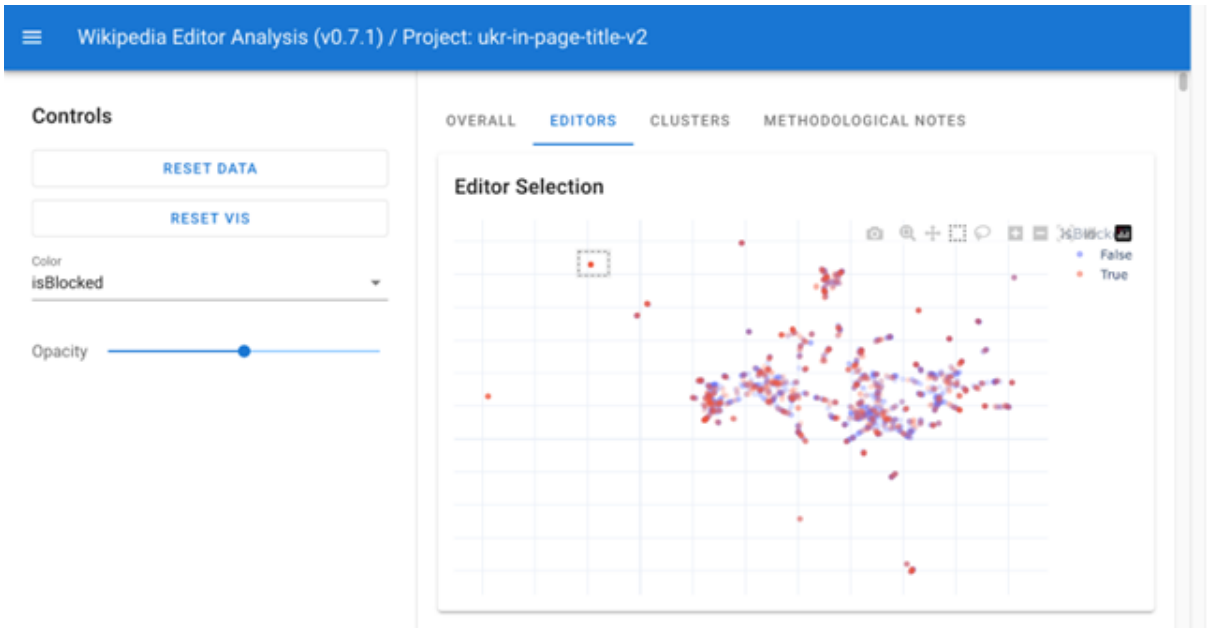

Figure 12: Dashboard main screen, showing semantic map of editors coloured by the block status.

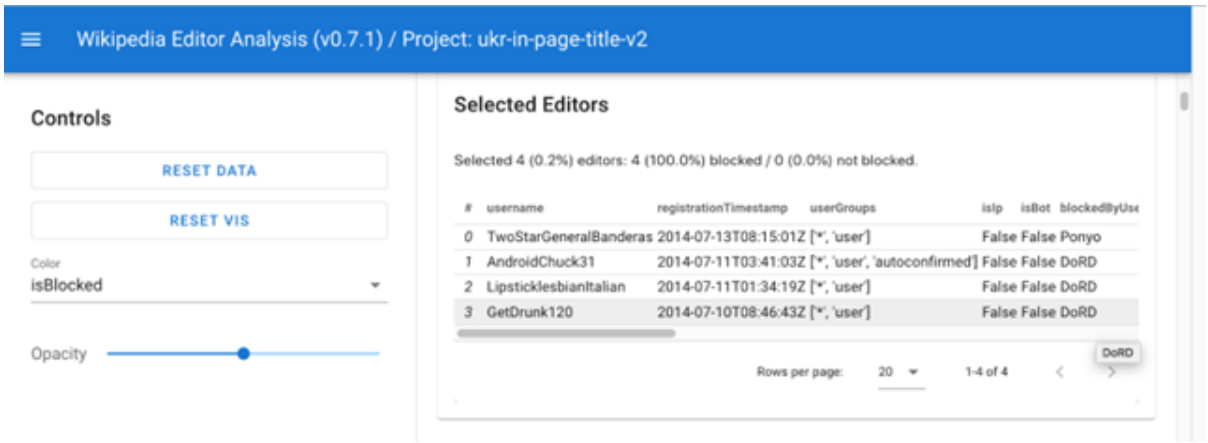

Figure 13: Dashboard table with editor metadata

Figure 14: Hybrid plot showing semantic mapping of editors and their edits

Contribution_ID	Revision_ID	Username	Page	Diff	Topic_ID
617372574_1	617372574	SamA509	Russo-Ukrainian_War	{ {about|Russian military intervention in Crimea and reported interference in eastern Ukraine|the secession crisis in Crimea and subsequent Russian annexation|2014 Crimean crisis|the secession crisis in eastern Ukraine|2014 pro-Russian conflict in Ukraine} }	18
929426658_1	929426658	176.98.70.118	Russo-Ukrainian_War	{ {About| Russian military intervention in post-Euromaidan Ukraine|Russian annexation of Crimea|Annexation of Crimea by the Russian Federation|the unrest in eastern Ukraine|2014 pro-Russian unrest in Ukraine|the ongoing military conflict in Donbass|War in Donbass} }	-1
1048517221_1	1048517221	92.40.171.75	Russo-Ukrainian_War	{ {about|Russoooee-Ukrainian War in post-Euromaidan Ukraine|Russian annexation of Crimea|Annexation of Crimea by the Russian Federation|the unrest in Eastern and Southern Ukraine|2014 pro-Russian unrest in Ukraine|the ongoing military conflict in Donbas|War in Donbas|the 1917–1921 war|Soviet–Ukrainian War} }	18
619362291_1	619362291	SNAAAAKE!!	Ukraine	{ {main|2014 pro-Russian unrest in Ukraine|2014 Crimean crisis|2014 Russian military intervention in Ukraine} }	-1

Figure 15: Sample of edit diffs

Detailed Cluster Analysis

We identified six clusters of interest that contained groups of accounts that were determined to belong to the same user. This determination was made through manual review of the account block logs and details. It's difficult to run this analysis automatically, mainly due to the lack of uniformity in how this information is stored. For users blocked due to sock-puppetry, the block reason text will sometimes contain an indication of who is believed to be the main user behind the account. When it is included, this information can take multiple forms, such as a direct mention of the main account username, but it can also point to a sock-puppet investigation page. The sock-puppet investigation page itself will sometimes include the name of the main account, but sometimes it will be the name of the sock-puppet instead, or to a different sock-puppet of the main account. This is likely an effect of how these investigations are carried out. An account may be believed to be the main account of a user, only to later find out this is actually a sock-puppet for a different account.

All this means is that to find the main account that's already been identified through a sock-puppet investigation, it's necessary to manually access and parse the block reasons, user pages and investigation pages. Manually navigating to the blocked account's user page usually has a reliable indicator of who the main account is. The figure below shows an example of such a notice on a sock-puppet account page.

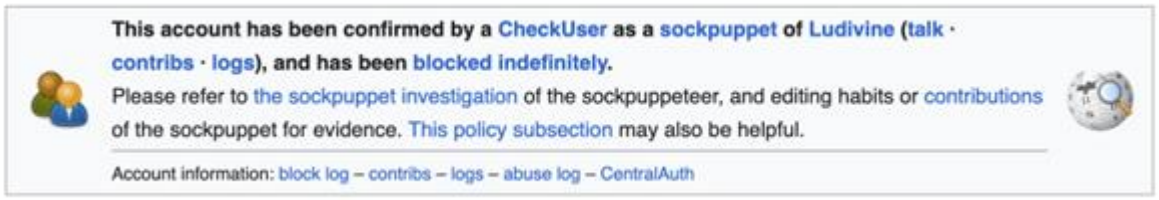

Figure 16: Screenshot of user page with sock-puppet notice

Four of the analysed clusters contained sock-puppet accounts of the same main user. An additional two contained groups of IP accounts that all originated from the same range, and were making similar edits, suggesting sock-puppetry.

Cluster_0
All seven editors in this cluster are IP accounts originating from the same range and blocked through an IP block. All the edits by these accounts are focused on a single page, the '*Russia–Ukraine relations*' article. All edits are making almost the exact same change, removing a link to the Wikipedia article about '*Malaysia Airlines Flight 17*'. The almost identical nature of the text in these edits are likely what led to the formation of this cluster.

Editors	7
Contributions	9
Revisions	9
Pages	1

Table 3: Cluster 0 overall stats

Page	**Contributions**	**Percentage**
Russia–Ukraine_relations	9	100.0

Table 4: Cluster 0 edited pages

Cluster_1
Five of the accounts in this cluster are blocked for sock-puppetry. Three of these are sock-puppets of the same user, while the rest are linked to other main accounts. The edits by these accounts are changing the location of Ukraine from Eastern to central Europe, as well as removing references to the CIA world factbook. Over half of the edits in this cluster were made by this set of sock-puppet accounts.
Edits by other accounts in the cluster make changes to the initial sentences of the article, for example changing it to call Ukraine a "former" country in Eastern Europe. While not all are making the same changes, the context provided for all the edits is the full initial paragraph of the article, which likely is what led to the formation of this cluster.

Editors	10
Contributions	20
Revisions	20
Pages	1

Table 5: Cluster 1 overall stats

Page	Contributions	Percentage
Ukraine	20	100.0

Table 6: Cluster 1 edited pages

Cluster_12

This cluster has eight accounts, four of them blocked, all for sock-puppetry. All of these accounts belong to the same user. Edits in this cluster are all to the '*Casualties of the Russo-Ukrainian War*' article, and editors appear engaged in an edit war in this page. The edits are not focused on a specific area of the page, and include the addition and removal of spans discussing specific casualties and personal details of the affected people. The owner of all the sock-puppets appears to have a long history of this behaviour in Wikipedia, with reports and temporary bans going back to 2010. The sock-puppet accounts in this cluster have a relatively small quantity of edits (70-300). Many of their other edits are to pages of notable people such as politicians from countries like Japan and India. They also have numerous edits to the '*List of COVID-19 deaths in North America*' article, with a similar format.
The majority of the contributions in this cluster are by the sock-puppet accounts. Contributions by the other accounts appear quite similar, also adding reports of casualties that occurred during the war, along with media references. Edits from both the blocked and unblocked accounts frequently take a format that starts with the date of the event, a short description of the same, some personal details of the affected persons, and a reference to a media site. The fact that they are following this similar format is likely to be the reason behind the formation of this cluster.

Editors	8
Contributions	134
Revisions	54
Pages	4

Table 7: Cluster 12 overall stats

Page	Contributions	Percentage
Casualties_of_the_Russo-Ukrainian_War	126	94.0
Belarusian_involvement_in_the_2022_Russian_invasion_of_Ukraine	6	4.5
Russian_information_war_against_Ukraine	1	0.7
Russian–Ukrainian_cyberwarfare	1	0.7

Table 8: Cluster 12 edited pages

Cluster_25

This cluster contains five editors, four of which are blocked for disruptive editing. They are all IP editors originating from the same IP range. All the edits in this cluster are to the '*International sanctions during the 2022 Russian invasion of Ukraine*' article.

Outside of the seed pages, the edits by these accounts focus on Switzerland and the topic of banking in particular. This IP range was blocked for adding links to pages that administrators determined contained antisemitic and conspiracist content as per the block log. They add this in the revision edit summary metadata rather than actual text in the articles and have done so in such articles as the '*World Economic Forum*', '*David René de Rothschild*' and '*David Rockefeller*'. Other than this, IPs in the range appear focused on Swiss banking.

The one non-blocked editor in this cluster also appears to be highly focused on editing articles concerning Switzerland. The account has a number of revisions to pages such as '*Health in Switzerland*', '*Banking in Switzerland*', '*Crime in Switzerland*', etc. The edits don't appear malicious. It's likely that the shared Switzerland focus between this account and those of the blocked range is what led to them being grouped together.

Editors	5
Contributions	27
Revisions	21
Pages	1

Table 9: Cluster 25 overall stats

Page	Contributions	Percentage
International_sanctions_during_the_2022_Russian_invasion_of_Ukraine	27	100.0

Table 10: Cluster 25 edited pages

Cluster_78

This cluster is formed of ten accounts, eight of which are blocked for various reasons. There is a small amount of vandalism in this section, but most edits seem engaged in an edit war over the naming of Ukraine. The majority of these edits remove links to articles explaining the difference between *"Ukraine"* vs "*The Ukraine*", both internal and external to Wikipedia. All of these changes are made within the first sentence of the introductory paragraph of the '*Ukraine*' article. Two of the accounts in the cluster are sock-puppets that belong to the same user and are making this exact same edit. Other accounts in the cluster are also editing this initial sentence, with changes such as "*Ukraine, sometimes erroneously called 'the Ukraine',* or changing the pronunciation of the name. However, other accounts are making unrelated changes in the same area or adding vandalism. It would appear that the high proportion of edits to the section of the page discussing the naming of the country are what led to the formation of this cluster.

Editors	10
Contributions	26
Revisions	24
Pages	1

Table 11: Cluster 78 overall stats

Page	Contributions	Percentage
Ukraine	26	100.0

Table 12: Cluster 78 edited pages

Cluster_100
This cluster contains nine users, five of them blocked in total with four blocked for sock-puppetry. Most edits are to the '*Russia–Ukraine relations'* and '*Armed Forces of Ukraine*' articles, and are all adding or updating country statistics. Edits to the first appear focused on the country comparison table, while changes to the second mostly affect the infobox area. A large proportion of the edits come from two editors who are sock-puppets of the same user, both of whom are mainly editing these statistics. A review of a sample of edits by these accounts in pages not in our collection reveals that they are making very similar edits on articles like '*China-Russia relations'*, '*Japan-Russia relations'*, '*China-Japan relations'* and others.

Edits by the non-blocked accounts in this cluster are very similar. The nature of the changes, which mostly consist of changing numbers in tables and infoboxes with little textual change, is likely the reason behind the formation of this cluster.

Editors	9
Contributions	84
Revisions	18
Pages	3

Table 13: Cluster 100 overall stats

Page	**Contributions**	**Percent age**
Russia–Ukraine_relations	55	65.5
Armed_Forces_of_Ukraine	18	21.4
Ukraine	11	13.1

Table 14: Cluster 100 edited pages

Other Analysed Clusters

Top-ranked clusters by proportions of accounts blocked for any reason

Cluster_130 (91% blocked)

Cluster 130 is formed of eleven editors, ten of which are blocked for a variety of reasons. About half the users are IP editors. Half of the blocked users (all the IP users) are blocked through an IP block. Of the remaining blocked users, most are blocked for vandalism, and one for disruptive editing. The vast majority of the edits in this cluster come from a single editor who has been blocked for vandalism. This editor's changes are all part of a single large revision that removed a lot of the content of the '*Ukraine*' article and replaced it with random looking text: song titles and authors, as well as sections that appear to be scraped from web pages. The rest of the users all appear to engage in vandalism as well, replacing legitimate content with gibberish text, nonsensical sentences such as *"I like pie"*, or insults such as *"Fuck Wikipedia Administrators"*.

Interestingly, the single non-blocked user in this cluster is also engaging in vandalism. The only revision in its history is to the '*Ukraine*' page, and consists in adding the sentence *"ukraine is gay"*. It's unclear why this user has escaped moderation. Since this is the only edit in its history it may not constitute a pattern of behaviour. At the same time, it's a clear example of vandalism which likely would've justified a block by itself.

Cluster_39 (87.5% blocked)

Cluster 39 has eight editors, seven of which are blocked for a variety of reasons, such as sock-puppetry, disruptive editing and vandalism. All the editors in this cluster are named users. The majority of the edits in this cluster are to the '*Ukraine*' article, with a few going to the '*Russo-Ukrainian war*' article. The majority of edits are modifying country statistics such as GDP and population in the infobox. These edits appear to be benign for the most part on the basis of the text itself, aside from some instances of vandalism. The reasons for these users being blocked are likely due to behaviour outside of what they are contributing to an article. Edits from the one non-blocked user in this cluster are quite similar to those of the blocked users.

Cluster_27 (83.3% blocked)

Cluster 27 is composed of six editors, five of which are blocked for a variety of reasons including sock-puppetry, vandalism and ip_blocks. Half of the editors are IP users. All of the edits in this cluster were made to the '*Ukraine*' article. Most of these edits are small single word changes, such as replacing "*Recently deposed*" to "*Former*", general spell-checking and updating of dates. Each editor only has one contribution made to the pages in this sample, however, reviewing their editor history reveals many more edits to a large variety of pages. As with Cluster 39, the edits in this cluster don't appear harmful based on the content of the

text. The reason for these users being blocked likely lies on behaviour found on another set of pages or outside of editing.

Cluster_61 (83.3% blocked)
This cluster is composed of six editors, five of which are blocked for various reasons, such as vandalism, IP block and sock-puppetry. Several of these accounts appear to have engaged in vandalism, even those not blocked for this. The cluster contains 50% IP users, all of which are blocked. The cluster contains only a few edits, and these editors seem active across a variety of pages. The sock-puppet user has been blocked for behaviour outside of these pages. The non-blocked user does not have any evident vandalism but appears to participate in some contentious pages and edit wars regarding India.

Edits for which evidence of Wikipedia rule-breaking exists include vandalism to the '*List of wars involving Ukraine'* page, with inclusions of "*russia is just annyoing"*, "*Putin started it. He is a bad person."*, and "*Vladimir Putin is having a meeting with zelensky on sunday, sept 17th 2022 to discuss pineapple on pizza."*

Cluster_62 (80% blocked)
This cluster has five users, three of which are blocked for sock-puppetry and one for vandalism. Most edits are to the '*Russo-Ukrainian War'* page. Most edits are making small changes to wording, such as changing "*Controlled by Russia and the insurgents*" to "*Controlled by Russia and pro-Russian forces*". Users tend to have a longer edit history outside of the pages in this sample. A high proportion of the edits in this folder are to the main image and caption of the article, which used to be a map of Ukraine showing areas controlled by each side, since replaced with a number of different images.

Cluster_144 (80% blocked)
This cluster is formed of 15 accounts, 12 of them blocked, most of them for sock-puppetry and IP blocks. Sock-puppet accounts don't appear to belong to the same users, at least according to the documented investigations. Edits are mostly to the '*Ukraine*' and '*Russo-Ukrainian war'* articles. The majority of the changes remove/add flag icons and file attachments. Changes don't appear harmful based on the text only.

Cluster_148 (80% blocked)
This cluster contains five editors, with four of them blocked, three for sock-puppetry. They don't appear to be sock-puppets of the same user. All edits are to the '*Russo-Ukrainian war*' page, and they focus on a section discussing casualties. These edits centre around updating casualty numbers, references to news articles, including Russian sources, as well as general formatting.

Cluster_156 (80% blocked)

This cluster contains five accounts, two of them blocked for sock-puppetry and two for disruptive editing. Users have a large number of edits in the sample. The cluster shows a diverse range of edit types distributed across numerous pages, thereby presenting challenges in characterising its overarching theme or pattern.

Top-ranked clusters by proportions of accounts blocked for sock puppetry

Cluster_6 (54.5 % blocked for sock-puppetry)

This cluster contains 11 users, six of them blocked for sock-puppetry. The sock-puppets don't appear to belong to the same user. The edits in this page focus on the introductory paragraphs of the '*Ukraine*' article. The edits appear to be over polarising themes, such as removing and adding back spans discussing the territorial dispute over Crimea, changing the spelling of place names, as well as what looks like vandalism in some sections, eg. changing "*country*" to "*dump*". Efforts to correct these changes from other editors and admins are also apparent.

Cluster_104 (50% blocked for sock-puppetry)

There are ten editors in this cluster, with six of them being blocked. Five of the editors are blocked for sock-puppetry, but they don't appear to belong to the same user. Many of the edits in this cluster are done to section titles and links to other Wikipedia articles. The edits cover a large number of pages in the sample, such as '*Ukraine*', '*Russo-Ukrainian War*', '*Russia–Ukraine relations*', '*Modern history of Ukraine*', among many others. This cluster is likely formed due to the high similarity in the text of the edits, which all affect spans that are quite short and have a large number of references to "Ukraine", "Russia", "Crimea" and "war". A small number of edits affect larger paragraphs, and usually include spell checking and modifying links to other articles.

Cluster_76 (50% blocked for sock-puppetry)

This cluster contains six users. Four of these users are blocked, three of them for sock-puppetry. The edits in this cluster are mostly to the '*Ukraine*' and '*Russians in Ukraine*' articles. They discuss notable people from both Russia and Ukraine. The edits to the '*Ukraine*' page appear to centre around sports and athletes. The majority of the contributions in this cluster come from a single editor, who is blocked for sock-puppetry. They appear to be pushing a pro-Russia POV, however it should be noted that this user has been blocked since 2008, so these additions are quite old.

Cluster_152 (44.4% blocked for sock-puppetry)

This cluster is formed of nine users, four of them blocked for sock-puppetry. This cluster seems concerned with geographical divisions of Ukraine, editing a variety of pages such as '*Ukraine*', '*Central Ukraine*', '*Southern Ukraine*', with some edits to the '*Russians in Ukraine*' page. The edits mostly consist of short spans of text such as modifying file and image attachments, links to other Wikipedia pages, adding templates indicating the need to update

pages, and adding external references to country statistics. Sock-puppet accounts don't appear to belong to the same user. These accounts also have a long edit history outside of the sample pages.

Cluster_171 (43% blocked for sock-puppetry)
There are seven users in this cluster, three of them blocked for sock-puppetry, mostly editing pages such as '*2014 pro-Russian unrest in Ukraine*' and '*Russo- Ukrainian War*'. Edits are mostly modifying references to external websites, including many in the Russian and Ukrainian languages, as well as updating statistics around the number of attendants to the 2014 protests. The sock-puppet accounts don't appear to belong to the same user, and two of them only have a single edit each to the pages in this sample, suggesting most of their activity happens somewhere else in Wikipedia.

Cluster_110 (43% blocked for sock-puppetry)
This cluster contains seven accounts, three blocked for sock-puppetry. Most edits are to the '*International sanctions during the Russo-Ukrainian War*' and '*List of invasions and occupations of Ukraine*' pages. Most edits consist of adding templates for redirects, related and outdated notes. No obvious harmful edits. Sock-puppet accounts don't appear to belong to the same user.

Top-ranked clusters by proportion of accounts blocked for vandalism

Cluster_59 (33% blocked for vandalism)
This cluster is formed of six accounts, two of which are blocked for vandalism. Half of the editors are IP accounts, two of which also appear to engage in vandalism, though they are not blocked for this reason. This means the majority of the accounts in this cluster have engaged in vandalism, even if not blocked for it. Edits by accounts in this cluster centre around the *Early History of Ukraine* section of the '*Ukraine*' article. They have likely been grouped together due to editing the same section of the article. The non-blocked accounts in this cluster appear to be making legitimate contributions to this same section.

Cluster_45 (31% blocked for vandalism)
There are 13 accounts in this cluster, four blocked for vandalism and six through IP blocks. The accounts blocked through IP blocks also appear to be engaging in vandalism, though it's not spelled out in their block reason. Again, most of the accounts in this cluster engage in vandalism. Edits in this cluster all centre on facts included in the infobox of the '*Ukraine*' article, especially in the *common_name*, *native_name* and *conventional_long_name* items. The similarity between these changes is likely the reason behind this cluster being grouped together.

Cluster_28 (20% blocked for vandalism)

This cluster has 10 accounts, two blocked for vandalism, two for sock-puppetry and two IP blocks. Edits in this cluster also centre on the *History of Ukraine* section of the '*Ukraine*' article, but one account also edited the *common_name* item of the page infobox, as in cluster 45. The two IP blocked accounts also appear to have engaged in vandalism in the '*Ukraine*' article. The sock-puppet accounts don't appear to belong to the same user, and their edits in this page appear legitimate.

Cluster_36 (20% blocked for vandalism)
There are five accounts in this cluster, with three of them blocked, only one for vandalism. One account blocked through an IP block also engaged in vandalism. Both made very similar edits to the '*Ukraine*' article, changing "*President of Ukraine*" to "*Dictator of Ukraine*". The reason for the other blocked account is uploading copyrighted material. Most of the edits in this cluster are to the '*Ukraine*' and '*Russo-ukrainian war*' articles. A high proportion of the edits are modifying information in the infobox, specifically the *leader_title* items, which is likely what has led to the formation of this cluster.

Cluster_48 (20% blocked for vandalism)
This cluster is formed of 15 editors, nine of them blocked, three for vandalism. Most edits in this cluster are deleting content from the '*Ukraine*' page. This is due to four of the editors completely blanking out the page and replacing it with random text. Three of these users are the ones blocked for vandalism, and the remaining one is blocked through an IP block. The reason for this cluster forming is likely a side effect of our handling of revisions, where they are split into individual contributions. This makes it so that a single revision that consists of removing everything in a page makes up hundreds of individual removal contributions. So, the overall representation of the account is made up of many small removals. The rest of the accounts in this cluster also have contributions that are mostly removing content, though they are not blanking the whole page at once. In that sense, these accounts are found to be similar by the language model.

Cluster_82 (20% blocked for vandalism)
This cluster contains ten editors, with six blocked, two of them for vandalism. Many edits in this section correspond to changing words between Russian and Ukrainian spelling, such as *Odessa* to *Odesa* and *Kharkov* to *Kharkiv*. A larger proportion of edits centre on updating/modifying dates of establishment of different states, all of which appear on the infobox of the '*Ukraine*' article. The large proportion of date changes are likely what led to the creation of this cluster.

Cluster_94 (20% blocked for vandalism)
There are five accounts in this cluster, with only one blocked for vandalism, which is an IP account. The account has few edits in the sample, all to the '*Ukraine*' page. These edits appear innocuous, so it's likely that vandalism happened in other pages in Wikipedia, or it was performed by other addresses in the IP range.

Cluster_119 (20% blocked for vandalism)
There are five accounts in this cluster, with only one blocked for vandalism. The blocked account is an IP account, in the same range as that from Cluster_94. As with that account, the edits don't seem obviously harmful. Most edits are to the '*Russo-Ukrainian war'* page and focus on the infobox portion discussing combatants and commanders.

Cluster_163 (20% blocked for vandalism)
This cluster is formed of five accounts, only one blocked for vandalism. Edits are to a variety of pages such as '*2010 Ukrainian presidential election'*, '*2014 Ukrainian presidential election'*, *'Russians in Ukraine'*, '*Ukraine'* and *'Russo-Ukrainian War'*. Several sub-clusters of edits are apparent, such as adding/removing content from tables on different pages or editing images and captions, resulting in many very small contributions. The one contribution by the account blocked for vandalism rather looks like a low-effort edit. They have a few contributions in other pages which are more evidently vandalism.

Cluster_172 (20% blocked for vandalism)
There are five accounts in this cluster, one blocked for vandalism and two through IP blocks. One of the IP accounts has also engaged in vandalism in the relevant pages. Most edits are to the '*Russia-Ukraine relations'* article, and they appear to be mostly small spelling corrections to the paragraphs describing the state of relations between the countries and reporting on disputes.