r/worldnews data analysis

Riccardo Cantoni

April 3, 2018

Todo list

Contents

Abstract

Reddit is a widely used platform for online discussion and debate. ”r/worldnews” is a specific part of Reddit focusing only on news of international relevance. In a moment in History where the impact of online news on thoughts, beliefs and ideas of the general public is in constant growth, it is of primary importance (and equal interest) to study such a platform.

In this work, data are collected for a short period of time and analised statistically to try and observe trends and information about the demographics, interests, common traits, information sources and the overall quality of the debate hosted by the website.

1 Reddit and r/worldnews

Reddit, which describes itself as ”The front page of the internet”, is a well known mass discussion website [7]. It is structured as a collection of user-created pages, ”subreddits”, that act as boards where users submit contents pertaining to a specific subject. Posts can be commented, as can the comments themselves, producing a potentially infinite comment tree. Each post and each comment can be up or down-voted by users, determining a scoring system.

Contents can be viewed and ordered in a number of ways, but posts with higher score will always tend to float towards the top of the subreddit, achieving higher visibility, while posts with lower or negative rating will be relegated to secondary pages. Users can also subscribe to one or more subreddits and receive the corresponding activity feed on their Reddit home page.

”r/worldnews” is a subreddit devoted to discussions about and around news of international relevance . It is one of the more active subreddits, with more than 18 million subscriptions, and hundreds of posts per day [11]. It is regulated by a set of rules that can be summarised as follows:

- Only links to news websites/blogs are allowed: no images/text/videos, no links to social networks.
- The articles linked must be in English language.
- The news linked must be of international relevance, and no more than one week old.
- No misleading titles, and no editorials or opinions
- Comments containing bigotry, offensive content, personal attacks, images are disallowed.

These rules are enforced by a moderation team. Some information about the moderators was gathered by directly contacting them:

- There are 26 active moderators currently active on the subreddit
- The participation is on an entirely voluntary basis
- The community is more or less self-regulating, so the amount of moderation required is relatively small. Still, occasional interventions are needed.
- Some topics are cause for a higher need for control on part of the moderation team (i.e. terrorist attacks in western countries)

Violation of these rules will cause the immediate removal of the submission without warning, while repeated or extreme violations will result in a subreddit level permanent ban.

2 Data workflow

2.1 Data gathering

Reddit offers a comprehensive set of APIs that allow the gathering of very specfic sets of data. However, given the scope of this research, a cURL[3] GET proved to be sufficient to retrieve a .json containing the data.

The transparency of the offered data can be rated as follows:

- **Tim Berners-Lee's Deployment scheme [15], 3 stars**: the information is available in machine readable, non proprietary format (.json), but neither RDF or SPARQL standards are used.
- **Tim Davies's 5-Stars of Open Data Engagement[1], 2 stars**: the information is demand driven and in-context, but there is no support for discussion about it.

The URL from which the .json was retrieved is:
https://www.reddit.com/r/worldnews/top.json?sort=top&t=day&limit=50 The parameters specify different aspects of the required data:

- **top, sort=top**: the selection and subsequent ordering of the posts extracted is based on their score as $upvotes - downvotes$
- **t=day**: the selection is limited to posts submitted within the last 24 hours
- **limit=50**: the selection is limited to the first 50 results

The file was retrieved every 24 hours at 8PM local time (GMT+1), starting on the 25th of January until the 6th of February, for a total of 12 days. It contained a list of objects describing different posts from the subreddit. Each of them had a large number of attributes, from which a subset was selected as relevant:

- **title**: the title of the post, as given by the user. "r/worldnews" guidelines state that it "must fairly represent the article"
- **url**: URL of the linked article
- **score**: the score given to the post
- **permalink**: link to the reddit page of the post
- **created_utc**: timestamp of the submission
- **num_comments**: the number of comments to the post
- **domain:** the domain from the URL linked

The extraction of the relevant information from the original file and its subsequent processing were performed by some "ad-hoc" python scripts. An other dataset was acquired on the 6th of February, representing the 100 most voted posts of the previous month, thus starting from the 6th of January. The data were subjected to the same processing workflow.

2.2 Automated data processing

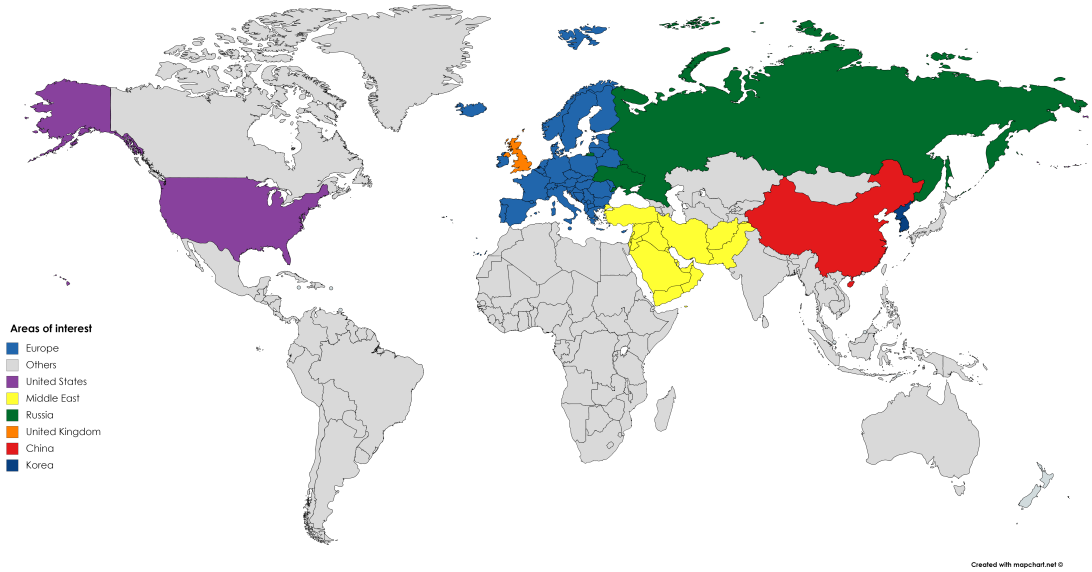
In order to attach more information about the news agency/website from which the article came from, an additional .json data structure was used to store data about the most common news agencies: this was done by hand, picking those sources that appeared multiple times, and building a table of triples in the form $< regex, agency, country >$, where $regex$ is a regular expression necessary to recognise the source from the URL domain, $agency$ is the complete formatted name of the source, and $country$ is a geographical tag (see section 2.2). Sources that don't have a recognisable nationality (such as blogs and small websites) and sources that only appear sporadically (less than 3 times) have not been added to the table.

The script then uses this table to add data to the posts whose source is stored: every domain is matched against the regular expressions, and if a match occurs, the corresponding agency name and country are added.

Then, the URLs are added to a set-like structured stored in an additional .json file. By ensuring that no duplicates occur, this table will be used to detect when the same page has been posted multiple times. URLs stored this way are passed through a minimal normalisation functions that cuts away querying strings, discarding everything after the first "?" character.

2.3 By-hand classification

During the process of classification and analysis, geographical localisation was used to characterise the source of the news, as well as their subject. Countries were distributed among 7 groups, plus a class "Others" for those countries that appeared in the data sporadically or not at all. The definition of these groups was based on, rather than actual geography or political views, a certain homogeneity of the most debated topics emerging from the news pertaining to them. For instance Ukraine belongs to the group "Russia", since the news about Ukraine are usually about its relationship with the Russian Federation. These areas are: United States, Russia, Europe, United Kingdom, China, Korea and Middle East. The latter features the most pronounced difference between its geographical definition and classification: in fact it encompasses the actual Middle East, as well as Turkey, the Arab Peninsula, Iraq, Iran, Afghanistan and Pakistan. Sources that are clearly international in intent ("Climate Change News", for instance) or whose nationality is unclear (personal blogs, mostly), were left untagged.



The classification process, except for the automated analysis of domains and the subsequent attachment of data relative to the source of the article (as explained in 2.1), was done by hand: the .json file was converted to .csv format using an online tool [4] for easier reading and easier data insertion. 3 additional attributes were added.

A geographical tag referring to the subject of the article, following the geographical classification system explained above. This step proved to be particularly tricky, as in many cases articles don't really have a physical subject that can be assigned to a country or group of countries. The definition of a specific policy proved to be necessary to associate tags in a meaningful manner:

- If the article is about a physical event, it is tagged with the area where the event took place.
- If the article is about a person, it is tagged with the area where this person lives/works/operates.
- If the article is about a statement given by someone, or a communication between multiple parts, the tag should reflect the geographical localization of the topic of the statement/communication, rather than that of the parts involved, or the place where the statement/communication took place. For instance, the article with the headline "Slovenia recognizes Palestine as an independent state" was classified as "Middle East", rather than Europe. "Russia Will Not Recognize US Sanctions Against North Korea", while involving 3 different areas, is classified under "Korea", and so on.

Two tags are used to classify the subject of the article. One macro-category tag was used to divide posts based on macroscopic fields: national news, international politics, economics, religion, science, environment and technology. Sub-categories are meant to pinpoint specific ”hot topics”: events or topics that are currently of particular interest: the ”Russiagate” scandal, ”Brexit”, political tensions in the Middle East, North Korea and its foreign affairs, Donald Trump (news that focus specifically on his person or his role in an event), articles where the topic is Vladimir Putin, Islam, the migration crisis.

It is important to note that the whole classification process was done purely on the basis of post titles. The amount of material that needed to be treated made any reading of the actual articles unfeasible. As a consequence, unclear or misleading titles might cause incorrect labeling.

2.4 Final data format

After the classification step, files were converted back to .json format, using a custom-made python script. This proved to be necessary in order to filter out and discard unprintable characters appearing during the earlier steps. The resulting files were then fused together, and re-analysed to search for new recurring sources. If found, these were added to the source table.
The final produce, both for the collection of daily news and the monthly dataset, is a list of .json objects, each describing a post with the following attributes:

- **date**: the day in which the post was retrieved from r/worldnews.
- **title**: the title of the post, as given by the user who posted it.
- **url**: the URL of the original article.
- **permalink**: the URL of the post.
- **domain**: the domain from the original article's URL.
- **score**: the $upvotes - downvotes$ score associated to the post.
- **num_comments**: the number of comments to the post.
- **created_utc**: the timestamp of the submission.
- **source_agency**: the name of the news agency / website that wrote the original article.
- **source_area**: the geographical area where the source_agency is based.
- **subject_area**: the geographical area where the subject of the article is located.
- **macro_tag**: macroscopic division in broad topics.
- **sub_tag**: only tags special, ”hot” topics.

At the end of the data gathering process, 2 datasets are produced, one containing the top 50 daily (between 25/1/2018 and 5/2/2018) posts, with 50 posts per day for 12 days, for a total of 600 posts. The other containing the 100 most voted posts between 6/1/2018 and 6/2/2018.

3 Daily post analysis

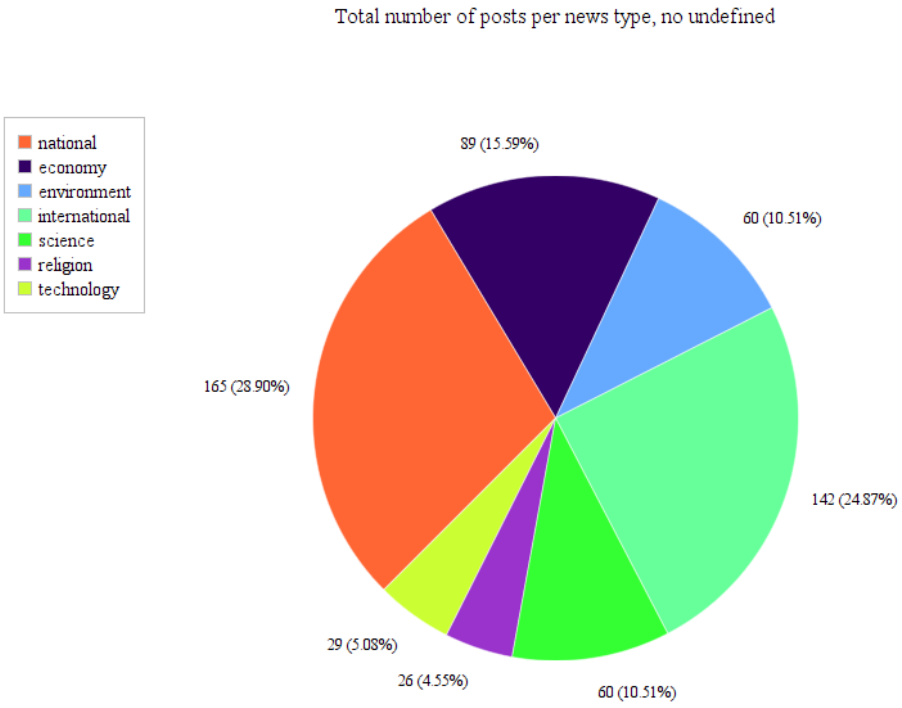
The entire analysis was done by querying the relative .json files using the ”jsonquery” javascript library [5], via its online application. The URL table (see section 2.1) built in order to filter out repeated submissions of the same article, proved to be completely unnecessary, since no such repetitions occurred. This might be consequence of the work of the moderation team (though submitting something already posted by someone else is not forbidden by the rules), or might be caused by the fact that out of multiple posts linking the same page only one gets upvoted enough to end up in the top 50. The part of the APIs that allow access to moderation logs is currently being reworked and not usable.

3.1 Discussion topics

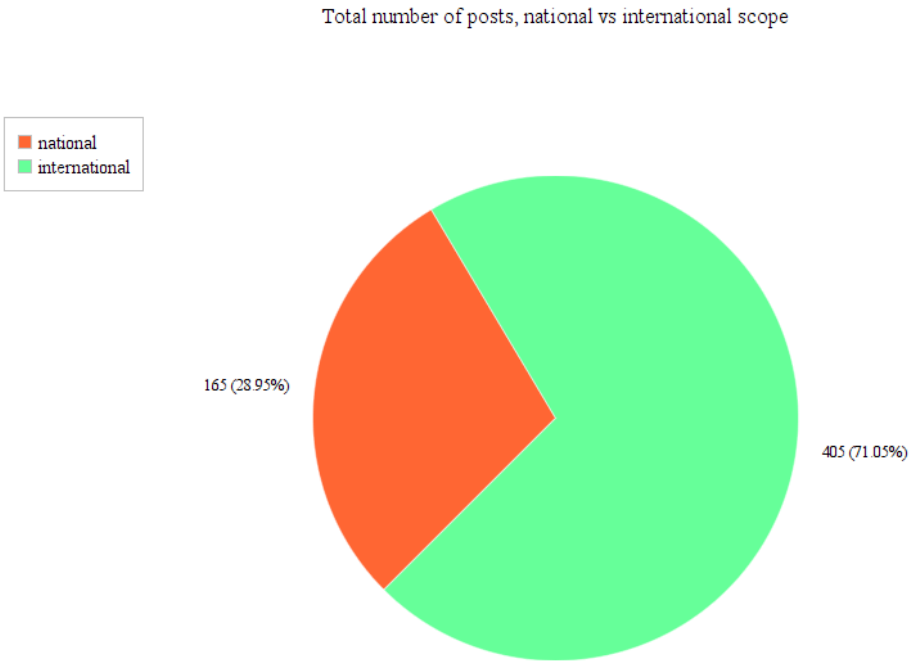
The classification of posts in broad fields of discussion makes it possible to study ”what type of news” are currently of interest to the users of r/worldnews. Excluding posts tagged as ”undefined”, macroscopic tags are distributed as follows:

1. National news, 165 posts (28.9%): news with a national point of view, albeit of international relevance.
2. International news, 142 posts (24.87%): news about international affairs.
3. News about international economics, 89 posts (15,59%).
4. Scientific news, 60 posts (10.51%), about scientific discoveries and advancements.
5. News about the environment, 60 posts (10.51%).
6. News about technology and its interactions with human life and society, 60 posts (10.51%).
7. Articles strongly related to religion, or in which the focus is on the religious aspects of an event, 26 posts (4.55%)

It is important to note that the same news could be represented multiple times through multiple articles or posts.



Given how classes were defined, it is possible to group together all those articles that are of international scope or inherently supranational. Scientific, religious, technological, environmental and economic news as a whole, represent the vast majority of the material submitted (71%), while those of national scope are the minority, in line with the general guidelines of the subreddit.



Analysing these categories together with geographical tags makes it possible to observe, to a point, what types of events happen in different areas of the world.

	usa	russia	eu	china	uk	me	kor	others	world
national	19	20	33	7	26	12	2	48	165
economics	11	10	13	3	12	3	6	19	89
environment	4		5	1	5	1	1	16	60
international	17	11	18	5	12	35	12	14	142
science	2		5	6	4	2		8	60
religion			9	1	1	8		5	26
technology	4		1	1	2		2	5	29
undefined	1	1	9	2	1			7	29

	usa	russia	eu	china	uk	me	kor	others	world
national	0.33	0.48	0.35	0.27	0.41	0.2	0.09	0.39	0.28
economics	0.19	0.24	0.14	0.12	0.19	0.05	0.26	0.16	0.15
environment	0.07	0	0.05	0.04	0.08	0.02	0.04	0.13	0.1
international	0.29	0.26	0.19	0.19	0.19	0.57	0.52	0.11	0.24
science	0.03	0	0.05	0.23	0.06	0.03	0	0.07	0.1
religion	0	0	0.1	0.04	0.02	0.13	0	0.04	0.04
technology	0.07	0	0.01	0.04	0.03	0	0.09	0.04	0.05
undefined	0.02	0.02	0.1	0.08	0.02	0	0	0.06	0.05

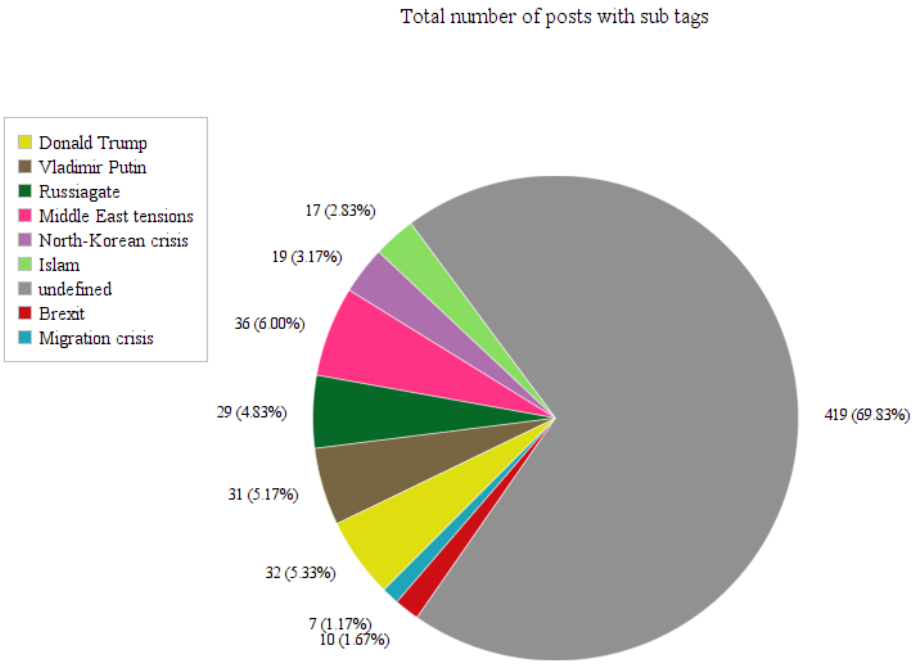
The results are shown above in plain numbers and in ”probabilistic” form, normalised on the number of posts with the same geographical tag. Interestingly, albeit not surprisingly, different areas of the world give origin to very different news.

There are no religious, technology-related, scientific or environmental news about Russia. News about Russia have a 48% probability of being about Russia's internal affairs, possibly due to its recurrent political troubles. Recent events include demonstrations organised by the opposition to Vladimir Putin, the arrest and subsequent release of Alexei Navalny, and the murder of a minor member of the opposition. News about

the Middle East and Korea have more than a 50% chance of being of the ”international affairs” type, twice as much as the global probability. The Middle East also sees a higher incidence of posts about religion, probably discussing around the Islamic faith China leads in scientific news: Chinese scientists have cloned primates, started working on lasers of unprecedented power and successfully implanted lab-grown ears on human children. The UK also shows a higher probability of ”generating” posts about its internal affairs. Notable events include the ongoing proceedings for Brexit and the widespread discontent for new NHS cuts. These connections can be analysed further by incorporating ”sub_tags” that pinpoint specific events to explore the impact of these events on the total numbers:

- Articles about Vladimir Putin make up for 55% of the news that fall under ”Russia” and ”National Affairs”.
- Half of the remaining international news about Russia have to do with the developments of ”Russia-gate”: 25% of the total.
- Brexit doesn’t seem to have such a strong impact on UK news: it makes up for 20% of it’s national affairs and 30% of its economic news.
- The ongoing North-Korean crisis makes up for the vast majority of the articles about the Korean area: 73% of all news and 70% of international news.
- Article that focus on the person of Donald Trump are 45% of ”national” news about the US.

”Sub tags” can also be used to measure the frequency of the specific subjects they are attached to.



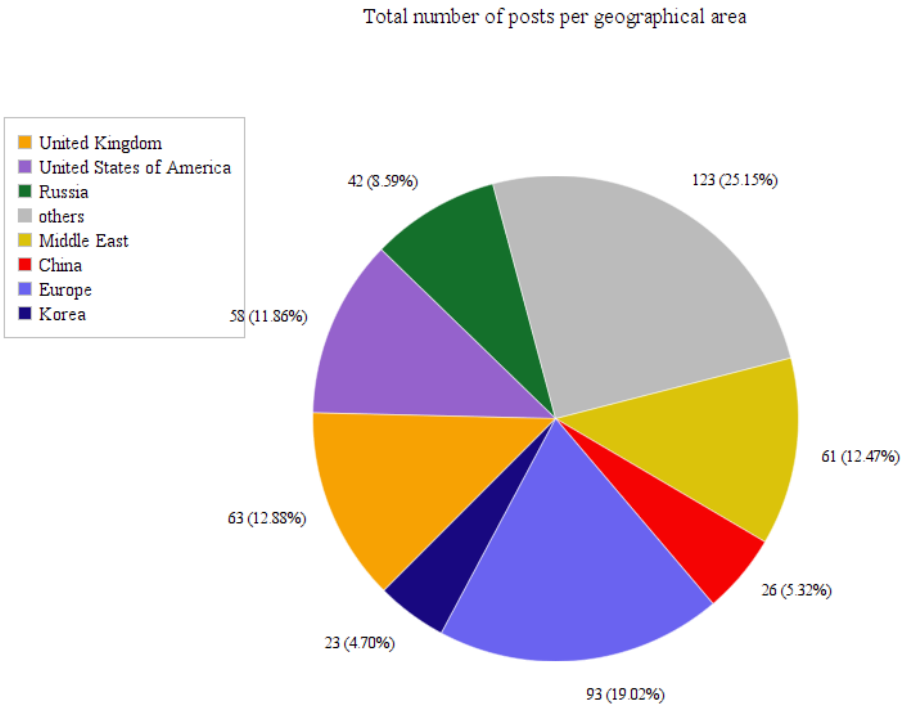
It is evident that the impact of these topics is not great. 70% of the total is in fact without a sub_tag. However 4 of them achieve notable numbers:

1. Tensions in the Middle East: 6% of the total.
2. Donald Trump: 5.33%

3. Vladimir Putin: 5.17%

4. Russiagate: 4.83%

Looking at geographical tags alone, it is possible to obtain a measure of how many articles talk about the different regions:



The results show that, after the ”Rest of the world” group, Europe leads with 19.02%, followed by the UK, the Middle East area and only then the US, with 12.47%. This would appear to be in contrast with the fact that 40% of Reddit users are from the United States, as evidenced by a 2017 survey [8].

This discrepancy could be explained in multiple ways: r/worldnews users could show a different distribution to the general one from Reddit as a whole. There are other subreddits which could draw American users away from r/worldnews: r/USnews [10], specifically about US national-level news, and r/news[9], for generic news. The first only has 10500 subscribers and a handful of posts per day, and is therefore too small to account for the discrepancy. The latter has 15.6 million users, enough to shift the balance. Proving the influence of r/news on the demographics of r/worldnews would require the gathering of data from both subreddits in parallel, and this was not possible due to time restrictions, especially since the geographical labeling of articles is extremely time consuming.

An other possible explanation would be that users from the US are present in the subreddit, but they prefer posting and discussing news that are about other areas of the world, this proves impossible to demonstrate. In the next subsection this will be discussed further using the geographical characterisation of sources, rather than that of article subjects.

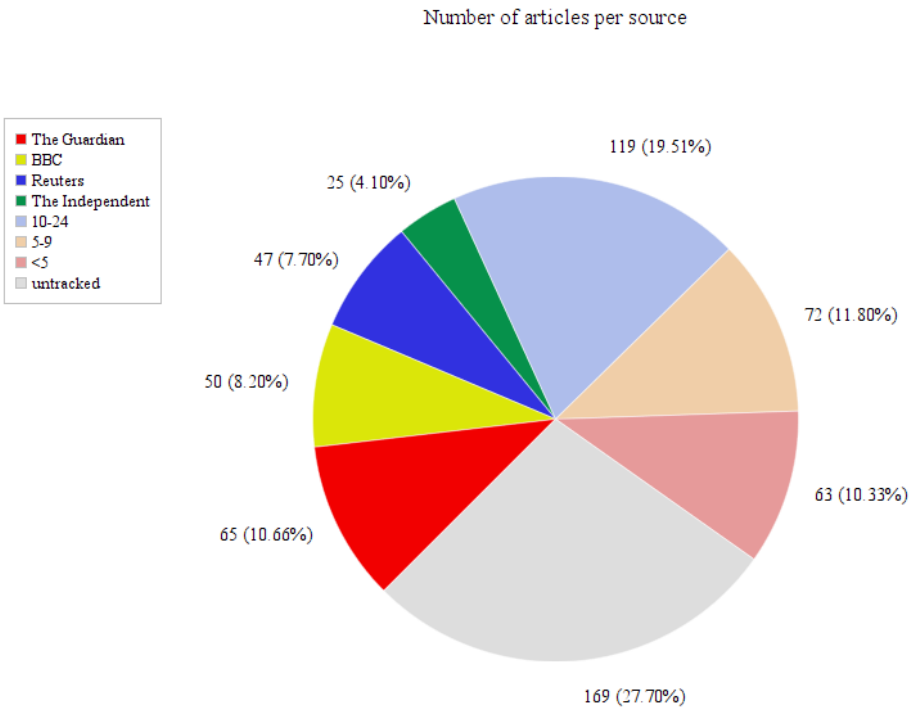
In any case it is currently not possible to gather data about the nationality of users without prolongued surveys and cooperation with administrators to keep the surveys visible to all users. There is no ”nationality” field in one’s user profile, and there is no automatic way to gather profile data via APIs.

3.2 Sources

In this section the sources of the submitted articles will be the subject of the analysis. The relevant data are the ”source_agency”, representing the name of the source, and ”source_area”, describing its geographical location.

Out of 600 articles submitted, 169 come from non-tracked sources that appear less than 3 times, or that were not recognised by the regular expression matching system. The distribution of the remaining 431 shows:

- The Guardian, 65 articles. The guardian is a daily newspaper from the UK, with a 200 year old history. It is widely regarded as leaning towards UK’s mainstream left [12].
- BBC, UK’s public service broadcasting corporation, 50 articles.
- Reuters, 47. Reuters is a major international news agency based in London, UK.
- The Indepentent, British online newspaper, 25 articles.
- 10 sources have between 24 and 10 submissions, for a total of 119 articles.
- 11 sources have between 9 and 5, for a total of 72.
- 63 articles come from all the remaining tracked sources, with less than 3 to 4 submissions.



A fact emerges, that a small set of sources are by far more represented than all the others: the 4 most represented sources, in fact, constitute 30.66% of the total. This raises the issue of a noticeable lack of plurality in news providers. Moreover, 50% of the total comes from only 14 voices.

Plurality in the media, and in the news especially, is widely recognised to play a fundamental role in the democratic debate[17], by providing diverse opinions and points of view.

A free and diverse media are an indispensable part of the democratic process. They provide the multiplicity of voices and opinions that informs the public, influences opinion, and engenders political debate.[20]

A healthy democracy depends on a culture of dissent and argument, which would inevitably be diminished if there were only a limited number of providers of news"

The problem appears to be even more extreme when the geographical origin of the articles is taken into consideration:

origin	UK	USA	Europe	Middle East	Russia	China	others	undefined
total	226	146	36	12	4	4	46	126
percentage	0.38	0.24	0.06	0.02	0.01	0.01	0.08	0.21

Nearly 40% of the articles come from a single country. Nearly 75% come from only 2. These values are even higher if we consider that the "undefined" group is made of articles from sources that have no possible national localisation. If we exclude these, UK's contribution nears 50% if the total.

One cause of this is certainly the fact that the users of the subreddit are a population with specific traits: they are necessarily English speakers, with easy access to the internet and the know-how necessary to use Reddit, which is not extremely user friendly, especially if compared to other mainstream platforms and social networks. A report from 2013[2] showed that most Reddit users are young, educated males.

This would suggest a predominance of news coming from English-speaking, developed countries, which is confirmed by the numbers. Still it would once again seem natural to see predominantly US-made news, given the fact that most "redditors" are from the US[8]. These numbers, together with those shown above about the subject of news not being particularly concentrated around the US, could suggest that r/worldnews had a population which is in fact differently distributed to that of Reddit in general: with a smaller percentage of US users and a larger population from Europe and the UK.

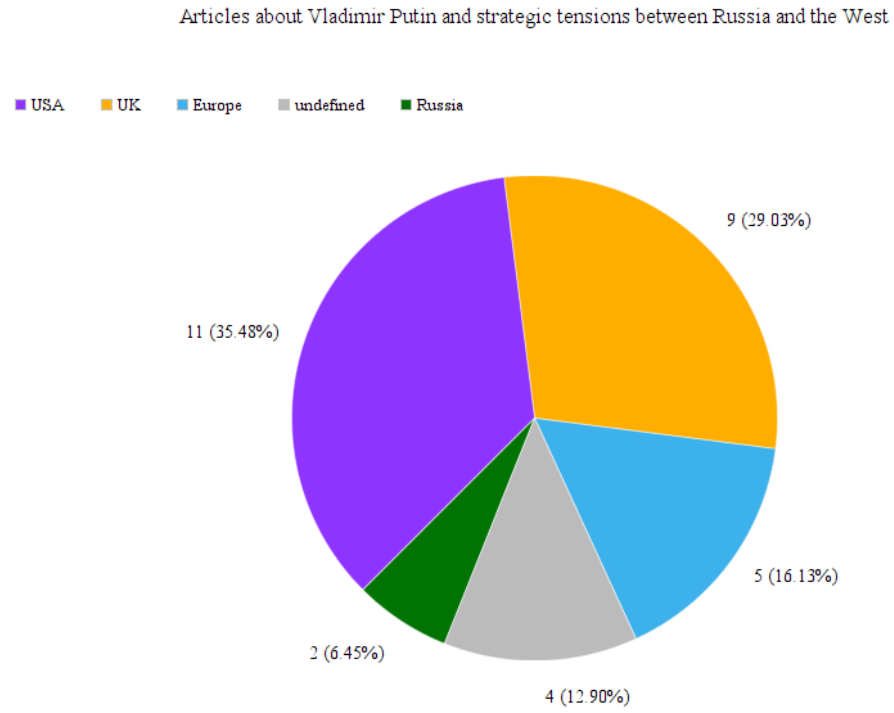
Cross-referencing the origin of the news with the geographical classification of its subject makes it possible to show if and how the geographical distribution of topics varies between different areas of origin. In plain terms it answers the the question "Who writes about who?". In this analysis only United Kingdom, USA and Europe are analysed singularly, while every other area (China, Russia, Korea, Middle East, Others) is grouped together in the "Others*" group. This is because the number of articles from sources in China, Russia, Korea and Middle East is too small to be of any statistical relevance. The numbers are normalised on the total number of articles written by sources from that group, the values therefore describe how likely an article written in a certain area (the row) is of being about a certain other (the column).

	UK	Europe	USA	China	M. E.	Russia	Korea	others	u.d
UK	0.22	0.19	0.07	0.02	0.08	0.05	0.05	0.16	0.16
Europe	0.06	0.44	0.03	0.00	0.19	0.11	0.00	0.11	0.06
USA	0.04	0.10	0.21	0.05	0.11	0.10	0.02	0.19	0.18
others*	0.03	0.06	0.05	0.08	0.14	0.08	0.05	0.39	0.14
u.d	0.03	0.13	0.08	0.08	0.10	0.05	0.04	0.21	0.29
world	0.08	0.18	0.09	0.04	0.12	0.08	0.03	0.21	0.17

Unsurprisingly, it appears that every geographical area tends to "write about itself" the most. However the numbers differ: articles written by European sources are twice as likely likely to be about itself than both the UK and the US (44% vs 22% and 21%, respectively). Europe also shows values close to 0 towards all areas except for Middle East and Russia, where they are twice as much as the world average.

It is impossible to verify if these values are caused "at the source" by newspapers and news agencies focusing differently on different areas of the world, or by the filter imposed by submissions on r/worldnews. Studying this would require a comparative study at the source. In any case this highlights well how certain areas are "talked about", but almost not represented as active parts in the discussion.

The problem of plurailty in news sources is even more extreme when we concentrate on specific, controversial topics. In these cases having a fair representation of the points of view of all the parts involved would be absolutely necessary to allow for a balanced debate. A clear example of this is the increase in tensions between the Russian Federation and Western countries, above all the United States, as Vladimir Putin's Russia tries to regain its former place as a strategical superpower. Clashing spheres of influence, "Economic War", political maneuvers and "Cold War-like" espionage appeared multiple times as headlines in the dataset. The relative sub_tag makes it possible to trace where those headlines were written.



Even though such a specific inquiry only involves a very small number of articles, and therefore doesn't give statistically relevant results, it effectively gives the idea of how extremely imbalanced the debate is. In fact, even the only two Russian articles are by the online version of "The Moscow Times", a dutch-owned weekly newspaper primarily aimed towards foreigners living in or visiting Moscow.

It is clear how, in this case, a highly controversial topic sees one of the parts dramatically underrepresented or not represented at all. Russian news in general are nearly non existent in the subreddit (only 4 out of 600 articles). ITAR-TASS, the main news agency from the Russian Federation, never appears once. The others (Reuters , Associated Press and Agence France-Press), are fairly well represented with a total of 62 articles, more than 1 in 10. What emerges is that, whether it be due to the nationality of users, or to a linguistic barrier, or else, r/worldnews appears to be extremely Western-centric.

While the internet in general is known to dramatically increase the plurality of available sources[13], it appears that this does not reflect on r/worldnews. The small number of voices is likely to cause a "Filter Bubble" effect, causing polarisation of the debate. While tentative, specific answers to this problem have been developed[21], attempting to use keyword extraction algorithms to automatically present multiple aspects of the same news, the problem remains largely unsolved and still affects many online platforms and everyday's political debate [18].

On the other hand, an empirical analysis of the dataset revealed no clearly untrustworthy sources, while the most recurrent ones are among the most trusted worldwide. Therefore, while the fairness and plurality of the debate is certainly not ensured, the emergence of "fake news" seems to have spared the subreddit. This appears to not be caused by intensive moderation: when asked, moderators replied that they "very rarely" have to cancel or hide posts, and when they do it is mostly because they are linking a website that requires payment to show its contents fully. The community seems to be "self regulating", which would be an interesting result for a group of this size: 18 million subscribed users. However, naturally, only a few do actually post articles.

This does not hold for every part of Reddit: there are cases in which the platform has been instrumental in the spreading of blatantly fake news, or has even originated them[19]. The only practical, working approach to trying to solve this seems to be moderation[6]. The results in r/worldnews would seem to indicate that such approach can be effective.

Other than observing their geographical distribution, it is interesting to study if different newspapers and agencies specialise differently in different fields. This can be done by cross-referencing sources and macro_tags:

	National	Int.	Eco.	Env.	Tech.	Science	Rel.	und.
The Guardian	21	13	6	13	6	4		2
BBC	19	10	6	4	2	5	1	3
Reuters	15	14	10	4	2	1	1	
The Independent	8	3	5	4	1	1	3	
CBC	6	1	1	1	1	2	2	2
ABC	4	1	4	4	1	2	1	1
others	92	32	100	16	57	45	18	21
all sources	165	60	142	29	89	60	26	29

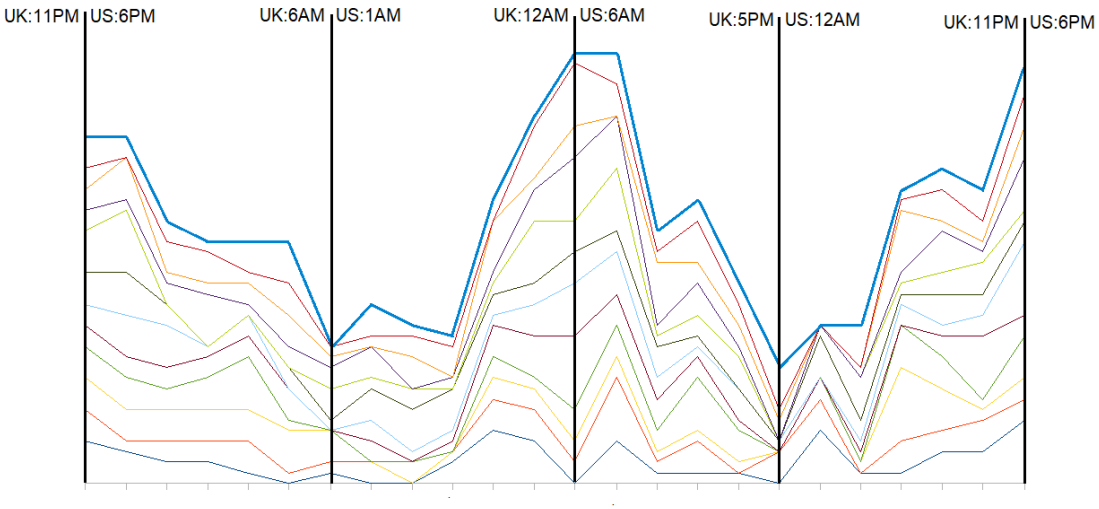
	National	Int.	Eco.	Env.	Tech.	Science	Rel.	und.
The Guardian	0.32	0.2	0.09	0.2	0.09	0.06	0	0.03
BBC	0.38	0.2	0.12	0.08	0.04	0.1	0.02	0.06
Reuters	0.32	0.3	0.21	0.09	0.04	0.02	0.02	0
The Independent	0.32	0.12	0.2	0.16	0.04	0.04	0.12	0
CBC	0.38	0.06	0.06	0.06	0.06	0.13	0.13	0.13
ABC	0.22	0.06	0.22	0.22	0.06	0.11	0.06	0.06
average	0.32	0.16	0.15	0.14	0.06	0.08	0.06	0.05
others	0.24	0.08	0.26	0.04	0.15	0.12	0.05	0.06
all sources	0.28	0.1	0.24	0.05	0.15	0.1	0.04	0.05

The second table shows the number of articles per source and category, normalized on the total of articles from that source. It expresses what percentage of the articles from a given source (or group of sources) is devoted to which topic. Only the top 6 sources per number of articles are analised individually.

It emerges that all sources follow a similar distribution: roughly 50% of the articles are about the economy or about national news, with a slight majority of the latter. Reuters differs slightly from the others by showing to be twice as likely as the average to be the source of articles about economics, while The Guardian only half as much. However, articles from The Guardian appear to be 5 times as likely to talk about environmental issues as the global average. Other minor discrepancies have not been considered relevant since the size of the dataset is too small (see section 7).

3.3 User activity

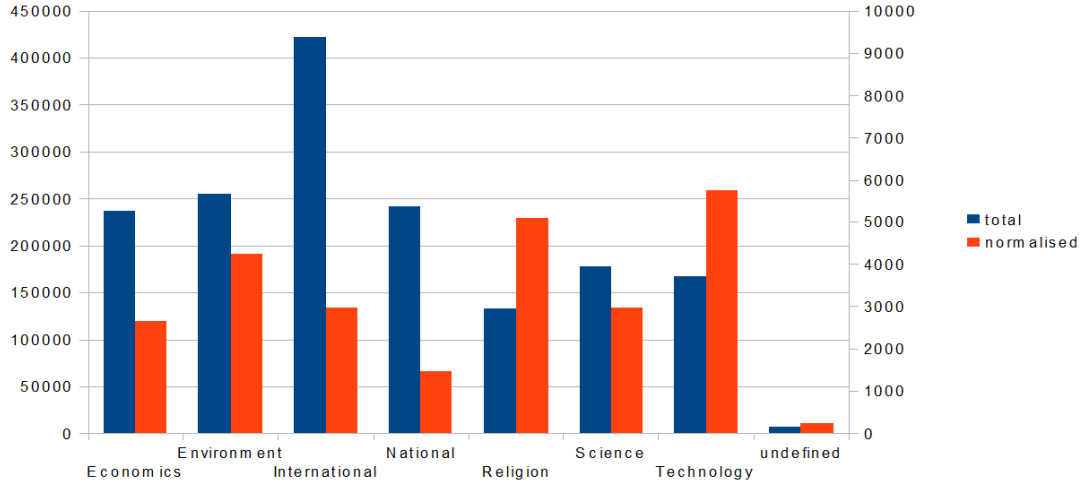
User activity can be analysed to obtain additional information about the subreddit and its demographics. The first intuitive step is to plot the number of posts per hour of the day, to see if any trend arises.



In the chart above each series represents a day. There are variations among the curves, possibly due to the size of the dataset. However, stacking series toghether to show the total sum of posts per hour of the day highlights a clear trend: there are two peak hours at 1PM and 12PM (Rome time), that are consistent with a large portion of users coming from Europe and the UK rather than one with many Norther Americans. This would be further evidence that the demographic distribution of users of r/worldnews is quite different from the one of the whole of Reddit.

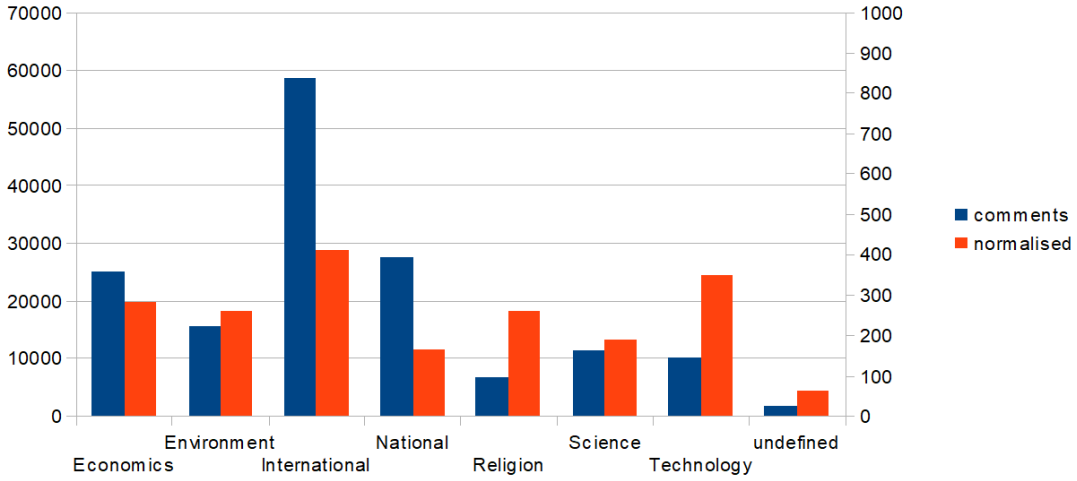
In order to explore this further, it could prove useful to observe when comments are posted or votes are issued. However, there is no timestamp associated to comments or votes in the dataset (see section 6).

The number of comments, and the score associated to a post, can be used to measure what topics attract more commenters or receives more upvotes, as a simple way of measuring the user's interest.



The chart shows in blue the absolute score assigned to each category, and in orange the value normalised on the number of posts falling in that category. Notably those categories that are more commonly posted are not the same that receive the most votes: National news, Economics and International affairs make up

for 70% of the total submissions, but the corresponding posts receive on average only half of the votes given to posts about Technology, Religion, Environment. These posts are posted less frequently, but appear to be more interesting to the average user. This is not caused by the different category frequency in the subreddit making different posts more or less visible, since categories that have similar numbers can still have different "votes per post" values: for instance National and International news have similar amounts of posts (165 to 142), but the latter has twice as many average votes per post (1461.90 to 2973.79). A similar discrepancy, albeit less pronounced, is present between science articles and environmental news.
Observing comments, a similar trend emerges:



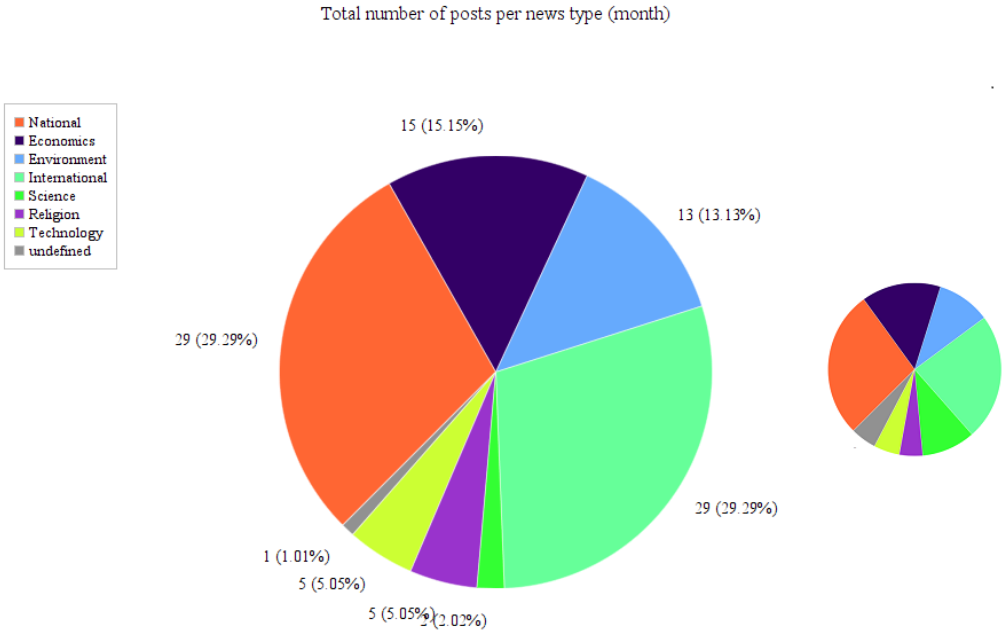
There are similarities and differences: again no strong correlation between the frequency of a category and the average number of comments emerges, again national news, despite being the most posted group, are the group that shows the least amount of user activity. However International news, that showed received low average scores, show by far more comments than any other category. Technology and religion show again to attract user activity despite the small number of posts.

4 Monthly data

The second dataset contained the 100 top scoring posts from the period 6/1/2018 - 6/2/2018. The two sets are biased in two different ways: the first draws posts from a short period of time and is therefore strongly deformed by the news that happened in that period, user preferences are relatively marginal, since picking the daily top 50 basically means picking all the posts that received ”some” upvotes (the lowest has 27). On the other hand the second dataset is less dependent from global events, since the timeframe is longer, but is strongly tied to user preferences, since it contains the 100 most voted posts of the month. In order to obtain a monthly dataset independent from user activity it would have been necessary to collect many more posts (1500 to keep the same ”top 50 daily” range used in the first set). This was not possible due to the time costs of manually tagging such a large amount of data.

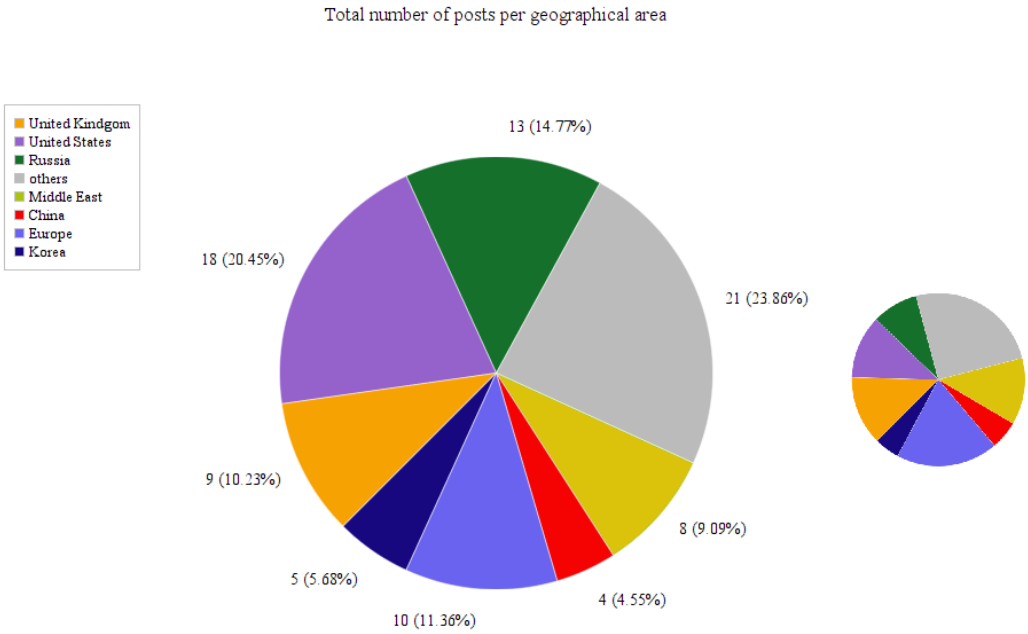
This second set can therefore be used to study if and how the different dependencies affect the final results, by comparing biased and unbiased (or less biased) data.

The relative frequency of posts belonging to different macroscopic categories is intuitively biased by what happens during the time frame under consideration. Comparing results from two different periods of time provides insight on the strength of this influence.

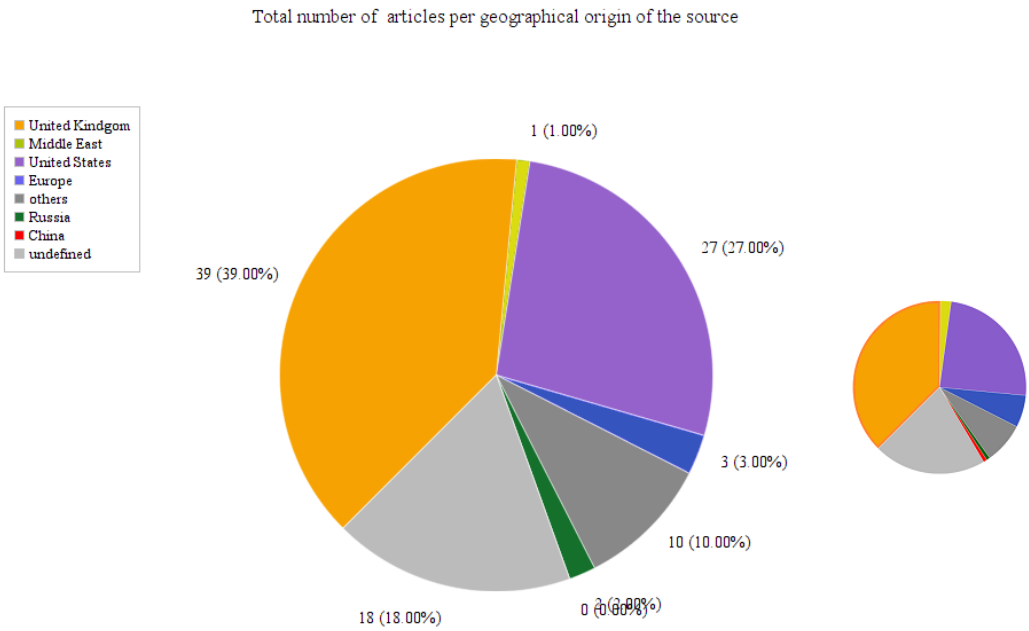


The chart reports results from the monthly set, and from the main set (in small, on the right). It shows that the proportions are reasonably similar, especially taking into account the increase in variance in the monthly dataset, caused by the smaller size. Grouping together all ”international” news we obtain a ”national to international” ratio that is practically identical in both sets: 29.95% to 71.05% in the main set, and 29.00% to 71.00% in the monthly one.

Moving on to the geographical distribution of the subjects we obtain again similar results:



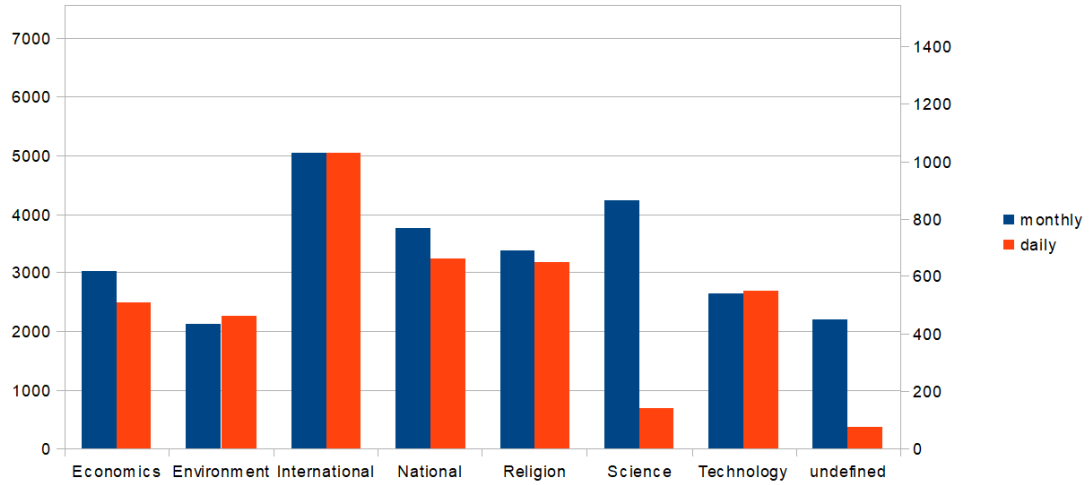
The proportions are quite similar, the only minor discrepancy is an increase in the percentage of news about the US and Russia, at the expense of Europe and Middle East. However, again, the variance in the monthly dataset could very well account for this, given the fact that the total shift of balance amounts to the equivalent of circa 15 articles over four different groups. The same holds for the are of origin of news sources:



The fact that sets with a different bias show almost identical results would seem to indicate (not prove) that the dependency is not affecting the final results in any noticeable way. Therefore the results regarding

the distribution of discussion topics, the geographical distribution of article subjects, and the geographical distribution of sources are at least confirmed, if not proved correct.

This method can also be used , to evaluate if and how user activity is deformed by global events happening during the timeframe, by comparing the results from the main set, highly biased, to the ones from the monthly set, less biased. Analysing data about votes is not feasible, since the monthly dataset only contains posts that have very high scores, making any comparison meaningless. However, user activity can still be evaluated by comparing the number of comments.



The chart reports the number of comments per post, for each category of articles. The series are very different in magnitude, due to the fact that in the main set posts have been picked when they were less than a day "old", while in the monthly set they can be as much as 30 times older. Clearly older posts can receive much more votes than younger ones. Once scaled accordingly to take this effect into account, the two sets appear to be remarkably similar in most categories, except for "Science" and "undefined", in which the difference is quite sizeable.

In both cases the number of comments per post in the main set is smaller by an order of magnitude. Deep inspection of the dataset revealed that, given the small numbers involved, just two articles can be the cause of this discrepancy:

- *"Cancer vaccine eliminates tumors in mice - 90 of 90 mice cured of cancers with lymphoma, similar results observed in colon, breast, and melanoma cancers."*[16], published by the Stanford Medicine News Center, has 4772 comments attached, the 54% of all the comments given to scientific posts. The article is about a discovery of extraordinary importance in the field of medical science, and it is the most commented post in the dataset, as well as the most voted. Alone it shifts the balance completely.

- *"Drone saves two Australian swimmers in world first"*[14], linked from BBC.com, is the only comment tagged "undefined" to be part of the top 100 most voted of the month. This is not entirely surprising since the average number of votes given to "undefined" articles is very low relatively to the other groups. Being the only element of the category it completely changes the average.

These data appear to show that the average number of comments per post category is not heavily influenced by the specific news happened during the timeframe. The only discrepancies are in those categories that have the lowest average comment count, and thus appear only sporadically in the top 100. Their extremely small numbers make it so that single outliers shift the values completely.

5 Conclusions

1. Articles about national news of international relevance are the most posted, followed by international affairs and economics. Science and environmental issues together amount for 20% of the total. Religion and technology-related posts are only posted sporadically.

2. Different areas of the world are talked about in different proportions, in each category of news. Some areas show a strong concentration of news of a specific type: Russia and the United Kingdom (internal affairs), Middle East and Korea (international affairs), and China (Science).

3. Tracking different specific topics or events shows how strong their impact can be on the amount of news that fall into a given category and geographic area: Vladimir Putin (Russia, International affairs: 55%), the North-Korean crisis (Korea, International affairs: 70%), Donald Trump (USA, National news: 45%).

4. At the same time the specific tracked topics only constitute a small part of the total set of articles posted: 30%, ensuring a certain degree of variety in the news discussed.

5. The analysis of the geographical distribution of article subjects shows that the most "talked about" areas are Europe, the UK and the Middle East. Given that the most represented country (by number of users) on Reddit is the US, this would be in contrast with the intuitive idea that US users would be interested in US-related news. Further results suggest that r/worldnews does not in fact share the same distribution of nationality as Reddit as a whole. Not only US-related news are only 10% of the article posted, but US sources only constitute 24% of the sources linked. Moreover, user activity, measured by the number of posts per time of the day, is more consistent with an Europe/UK based community, rather than a Northern American one. It would appear that the most represented demographic group is in fact that of users from the UK: 13% of the articles are about the UK, 40% of the articles are from UK-based sources.

6. The analysis of sources and their geographical distribution shows a worrying lack of plurality in the debate: a small number of sources hold a sizeable part of the total of articles, mainly The Guardian, Reuters and BBC. this reflects into an equal lack of plurality in the countries involved, with the UK amounting for 40% of the total. The disparity is even more extreme when specific, controversial topics are observed: the tensions revolving around the clash of strategic interest between Russian Federation and Western countries are the subject of 31 posts, only 2 of which are from Russian sources. This causes a clear and measurable "Filter Bubble" effect, cause by user activity rather than ranking algorithms. At the same time the subreddit shows no clear sign of fake or untrustworthy news.

7. Cross-referencing votes with news category allows to give a measure of the interest of users towards different kinds of posts. "International affairs" and "Technology" are the categories that attract the most votes, while "National news", despite being the most posted, is much less voted.

8. Comparison with a second dataset, differently picked and therefore differently biased, appears to confirm (albeit not prove) the correctness of the previous results. The sole discrepancies can be ascribed to the high variance caused by size restrictions.

6 Possible developments

1. The main limit to this analysis is the size of the dataset: 600 posts over a period of 12 days. As a consequence, the results are of limited statistical relevance. Especially when specific subsets are taken into consideration, such as when specific topics or geographical areas are observed, the numbers are clearly too low. The amount of data gathered was limited by the time allowed: it would have been possible to extract more than 50 posts per day but in doing that even posts with very low score would have been picked, at the risk of encountering unreliable sources, marginal topics and no user activity. Moreover, the process of manually tagging posts with attributes is, naturally, extremely time consuming and next to impossible to automate. It therefore limits the amount of data that can be gathered and processed.
   The development this analysis would benefit the most from is therefore an expansion of the time frame during which the data are collected.

2. The analysis highlighted discrepancies between r/worldnews demographics and that of Reddit in general. It might be interesting to explore this further by collecting information about the nationality of its users, and that of other subreddits. Since there is no personal information to express one's country of origin, the only viable ways to do this would be through surveys (hard to reach enough users, would require cooperation with administrators) or through internet traffic analysis (technically challenging).

3. The effect of moderation on the emergence and spreading of fake news could be explored by doing comparative analysis on other platforms/social networks. It would in any case require the relevant Reddit APIs to be restored, since they are currently not functioning.

References

[1] 5-stars of open data engagement?

[2] 6% of online adults are reddit users.

[3] curl.

[4] Json to csv converter.

[5] jsonquery library.

[6] Moderators of pro-trump reddit group linked to fake news crackdown on posts. *The Guardian*.

[7] Reddit.

[8] Regional distribution of desktop traffic to reddit.com. *statista.com*.

[9] r/news.

[10] r/usnews.

[11] r/worldnews.

[12] Voting intention by newspaper readership quarter 1 2005. *ipsos MORI*, 2005.

[13] Trends in media pluralism, 2017.

[14] Drone saves two australian swimmers in world first, 2018.

[15] Tim Berners-Lee. Linked data. `http://www.w3.org/DesignIssues/LinkedData.html`, June 2009.

[16] Krista Conger. Cancer vaccine eliminates tumors in mice. *Stanford Medicine*, 2018.

[17] Robin Foster. News plurality in a digital world. *Reuters Institute for the Study of Journalism*, 2012.

[18] Kalev Leetaru. Why 2017 was the year of the filter bubble? *Forbes*.

[19] Sapna Maheshwari. How fake news goes viral: A case study. *The New York Times*.

[20] House of Lords Select Committee on Communications. The ownership of the news. 2008.

[21] Souneil Park Seungwoo Kang Sangyoung Chung Junehwa Song. Newscube: Delivering multiple aspects of news to mitigate media bias.