Reddit – TUDFE: practical tool to explore Reddit usability in data science and knowledge processing

Jan Sawicki
Warsaw University of Technology
Department of Mathematics and Information Sciences
Warsaw, Poland
j.sawicki@mini.pw.edu.pl

Maria Ganzha
Warsaw University of Technology
Department of Mathematics and Information Sciences
Warsaw, Poland
m.ganzha@mini.pw.edu.pl

Marcin Paprzycki
Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland
marcin.paprzycki@ibspan.waw.pl

ABSTRACT
This contribution argues that Reddit, as a massive, categorized, open-access dataset, can be used to conduct knowledge capture on “almost any topic”. Presented analysis, is based on 180 manually annotated papers related to Reddit and data acquired from top databases of scientific papers. Moreover, an open source tool is introduced, which provides easy access to Reddit resources, and exploratory data analysis of how Reddit covers selected topics.

KEYWORDS
Reddit, online forum, dataset, text mining, knowledge source

ACM Reference Format:
Jan Sawicki, Maria Ganzha, and Marcin Paprzycki. 2021. Reddit – TUDFE: practical tool to explore Reddit usability in data science and knowledge processing. In Proceedings of The Eleventh International Conference on Knowledge Capture (K-CAP 2021). ACM, New York, NY, USA, 4 pages. https://doi.org/TBD

1 INTRODUCTION
Recently, social networks became popular repositories of data used for knowledge capture. The aim of this work is to explore usability of Reddit as a knowledge exploration data source. In this context, we present a review of literature about Reddit. Moreover, a specialized tool (Reddit-TUDFE) is introduced, which allows fast check of Reddit topic’s coverage. The key contributions of this work are answers to the following research question (RQs):

RQ1: What is the most popular method to acquire Reddit data?
RQ2: What are the most researched problems using Reddit as a dataset (do they include knowledge capture/management)?
RQ3: How does Reddit usage in science change over time?
RQ4: Are there topics that are not substantially covered on Reddit?
RQ5: Is Reddit used as a single dataset or with other platforms?

2 WHAT IS REDDIT?
Reddit is a web content rating and discussion website [14]. It was created in 2005 and is ranked as the 17th most visited website in the world, with over 430 million monthly active users1. Reddit is divided into thematic subfora (subreddits) dynamically created by its users. Therefore, the topic structure is evolving, in response to user needs. Each subreddit has its moderators who may supervise submissions and comments. Comment are linked to submissions, or earlier comments, forming a tree-like structure.

Most of the subreddits are public (for registered and non-registered users) with some exceptions (based on karma points, comments, gold, moderator status, time on Reddit, username and others). The tool (introduced in section 6 is based on publicly available data.

2.1 Accessibility – Reddit API vs. Pushshift API
Not only is the data on Reddit publicly accessible, but it is also distributed via the official Reddit API. However, it was found that most researchers do not actually use it. Instead, they choose Pushshift API [1]. None of the analysed papers state the explicit reason for this choice (very few mention how the dataset has been retrieved). However, testing the capabilities of Reddit API and Pushshift API shows that the key factor is that the Reddit API does not allow easy retrieval of historical data, while Pushshift does.

3 DATA ACQUISITION AND PROCESSING
To explore Reddit, as seen by scientists, a dataset of 180 papers was assembled. All of them were related to Reddit and submitted to arXiv between 01-01-2019 and 01-03-2021 (retrieved on 30-03-20212). This dataset has been processed manually and automatically. First, collected papers have been manually annotated with four attribute sets: topic (a general area of research), methods (theoretical approach, e.g. neural network, text embedding), dataset and technologies (practical software, e.g. BERT [4]). Next, obtained results were merged with publicly available data, i.e. the content (title and raw text) and the bibliometric metadata. This allowed extraction of information presented in Section 4.

Each collected paper has been converted to a raw text file, using PDF Miner [19]. Next, the key features of titles and texts have been cleaned and mined using the NLTK framework [12] (for sentiment and subjectivity) and TF-IDF [17] for vectorization (both from the sklearn library [15]).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

1https://www.statista.com/topics/5672/reddit/#dossierSummary
2https://arxiv.org/search/advanced?&terms-0-term=reddit&classification=computer_science=y&date-from_date=2019-01-01&date-to_date=2021-03-01
4 ANALYSIS AND FINDINGS
As a result of data processing we were able to formulate a number of observations. Let us summarize the most important ones.

4.1 Metadata and bibliometrics
Firstly, let us consider a few noticeable bibliometric and authorship statistics gathered using Semantic Scholar 3:
- There were two major growths in published articles count: one after March 2020 (correlated with the outbreak of the COVID-19 pandemic) and second in October 2020 (correlated with notification dates for many scientific conferences [21]).
- The majority of papers (over 65%) were written by 2-4 authors, with one having 26 authors [5].
- The most prolific authors (with over 10 publications) of Reddit-related papers are Savvas Zannettou (Max-Planck-Institute), Jeremy Blackburn (Binghamton University).

4.2 Analysis of topic, methods and technology
Topics and methods, used in studies, are summarized in subfigures of Figure 1. Top subfigure of Figure 1 shows that the most popular research topics is conversation, which matches the fact that Reddit is a discussion forum. Due to the timing of our work (overlapping with COVID-19 pandemic), the second most common topic is COVID.

Since Reddit consists of text discussions, it is not surprising that the two most common methods in Reddit-related research are text embeddings, used in text processing, and networks, used for social network analysis. Note that “network” (graph) and “neural network” are separate terms. Regarding technologies, over 45% of studies used Pushshift [1] for Reddit data extraction, and over 35% applied BERT[4] embedding (and its variations) for NLP.

Finally, topics and methods have been combined in a correlation heatmap (Figure 2).

Here, few significant correlations have been established.
- Papers related to drugs typically use word embeddings (which can be, however, related to the overall popularity of word embedding(s) (see citation count for [4]).
- Networks are typically applied in analysis of trends.
- Articles dealing with sarcasm often use LSTM networks.
- Research devoted to conversation generation typically applies BLEU metric.

Interestingly, only two papers (1%) are related to knowledge processing (specifically, knowledge graphs [2, 22]). Expanding arXiv search to all articles including “knowledge” and “Reddit”, resulted in 4 records, none related to knowledge capture. Therefore, top knowledge-related conferences were searched, but only one paper [7], about knowledge and Reddit, has been found (published by K-CAP 2011 4). This points to Reddit as an underexplored resource.

The previous remark may be counter-intuitive to the following remark. Moving to RQ5, it was discovered that among papers that use Reddit, over 30% also use Twitter, which is a data source mainly used for sentiment analysis [10]; RQ5). Other datasets utilized together with Reddit are: Facebook, 4Chan, YouTube, Gab. Each of them appears in less than 10% of papers which used Reddit.

4.3 Linguistic analysis
Linguistic analysis was performed using NLTK framework [12]. The sentiment of papers’ texts is mostly neutral (Figure 3).

Interestingly, a number of texts had high subjectivity (see, Figure 4), which should not occur in scientific publications [11]. However, closer scrutiny revealed that use of “subjective” words (e.g. “controversial”, “bias”) results in higher subjectivity scores, e.g.: “However, this openness formed a platform for the polarization of opinions and controversial discussions” [9] (score: 0.95).

5 REDDIT – THE ULTIMATE DATASET FOR EVERYTHING
Let us now address RQ3 and RQ4. Even though, they cannot be unequivocally answered, they can be experimentally evaluated.

5.1 Reddit in scholarly research
As shown in Figure 1, within arXiv, Reddit has been used for a study of a variety of research topics. However, Reddit dataset is present also in other sources. To show how the number of scholarly papers (related to Reddit) changed between 2010 and 2021, 10 databases have been analysed. As shown in Figure 5 the number of articles each year rises year to year dynamically (RQ3).

The reason for outlaying results of Google Scholar may be one of its primary criticisms [8, 8, 20], i.e. incorrect bibliometrics (due to use of automated algorithms) [6, 13] and its “inability or unwillingness to elaborate on what documents its system crawls” [6]. Moreover, Google Scholar declares inconsistently the number of query results and the actual number of returned results (e.g. a query returns 1000 results and declares 58 6005).

5.2 Google Trends
The next experiment explored presence of popular trends in Reddit. For all Global Google Trends 20206 their Reddit presence has been measured (see table 1). Overall, 79% of top Google Trends have a dedicated subreddit, while all of them are widely discussed. Table 1 illustrates top 1 in each Google Trend category.

6 IS THIS ON REDDIT? – REDDIT-TUDFE
To further explore whether Reddit supports capturing “knowledge about any area”, a tool for easy exploratory data analysis (EDA [3]) was designed. Reddit-TUDFE allows quick search of any topic on Reddit, checking if/how it is represented, and how it is discussed. Reddit-TUDFE delivers the following functions:

(1) Uses Reddit API to search for best matching subreddit.
(2) Downloads newest N posts from the subreddit, using Pushshift API and a combination of PRAW8 and PSAW9.
(3) Performs basic text cleaning (tokenization with NLTK [12], removal of stopwords, punctuation, numbers).
(4) Generates and displays post titles and content wordclouds10.

---

1https://api.semanticscholar.org/v1/paper/
2https://github.com/praw-dev/praw
3https://trends.google.com/trends/yis/2020/GLOBAL/
4https://github.com/amueller/word_cloud
5https://scholar.google.com/scholar?start=990&q=reddit&hl=en&as_sdt=0,5&as_ylo=2020&as_yhi=2020 accessed on 11-09-2021
6https://trends.google.com/trends/yis/2020/GLOBAL/
7https://github.com/praw-dev/praw
8https://github.com/dmarx/psaw
9https://github.com/amueller/word_cloud
10https://scholar.google.com/scholar?start=990&q=reddit&hl=en&as_sdt=0,5&as_ylo=2020&as_yhi=2020 accessed on 11-09-2021

The code follows state-of-the-art solutions for code sharing ([16]) and is publicly available on GitHub\footnote{\url{github.com/JanSawicki/reddit-tudfe/}} as a Jupyter Notebook [18].

To present capabilities of the application, let us present a few examples, as subfigures of Figure 6

- Left subfigure shows result for the phrase “music”, a generic term, which is certainly discussed on Reddit. One may see particular genres: rock, pop, rap, relaxing, electronic, etc.
- Middle subfigure displays results for phrase “rock”, a bit narrowed topic, but still vague and also present in Reddit, including artists/bands like: Rolling Stones, AC/DC, Led Zeppelin, Queen, Pink etc.
- Right subfigure contains a strictly specific topic, i.e. the band “The Beatles”, which is also widely covered on Reddit. Here one may see individual members: John Lennon, Paul McCartney, Ringo Starr, and George Harrison.

The wordclouds are build from posts related to a subreddit dedicated (or closest) to the searched topic. Reddit-TUDFE allows to...
quickly check if, and how, a particular topic is covered. Note that similar examples can be derived for any other topic, while Reddit shows potential in e.g. building ontologies, or semantic graphs.

7 CONCLUDING REMARKS

This work provides evidence that Reddit is a robust, but underutilized resource for knowledge capture, in almost any field of interest. In this context, the following conclusions can be formulated:

- Reddit offers publicly available data, which can be easily retrieved with Pushshift API.
- Most popular techniques for Reddit knowledge capture are: word embeddings, graph networks, and neural networks.
- Reddit covers majority (79%) of topics that appear in Global Google Trends, sustaining the claim that Reddit is a robust source of knowledge about “everything”.
- Reddit research becomes more popular over time (based on count of published articles).
- Reddit is most commonly used in tandem with Twitter.

This analysis and the Reddit–TUDFE tool provide foundation for future research on Reddit and its potential in fully automatic knowledge extraction and knowledge graph building.

REFERENCES

[1] Jason Baumgartner, Sarvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In Proceedings of the International AAAI Conference on Web and Social Media. Vol. 14. 830–839.
[2] Lei Cao, Hujuan Zhang, and Ling Feng. 2020. Building and using personal knowledge graph to improve suicidal ideation detection on social media. IEEE Transactions on Multimedia (2020).
[3] Victoria Cox. 2017. Exploratory data analysis. In Translating Statistics to Make Decisions. Springer, 47–74.
[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
[5] Ivan Garibay, Tolstam A Ogbar, Nilofar Yousefi, Ece C. Math, Madeline Schiappa, Steven Scheinert, Georgios C Anagnostopoulos, Christina Bouwens, Stephen M Fiore, Alexander Mantzaris, et al. 2020. Deep agent: Studying the dynamics of information spread and evolution in social networks. arXiv preprint arXiv:2003.11611 (2020).
[6] Jerry E Gray, Michelle C Hamilton, Alexandra Hauser, Margaret M Janz, Justin P Peters, and Fiona Taggart. 2012. Scholarly: Google Scholar and its value to the sciences. Issues in Science and Technology Librarianship 70, Summer (2012).
[7] Janna Hastings, Oliver Kutz, and Till Mossakowski. 2011. How to model the shapes of molecules? Combining topology and ontology using heterogeneous specifications. In Proc. of the Deep Knowledge Representation Challenge Workshop (DKR-11), K-CAP-11. Citeseer.
[8] Péter Jacsó. 2005. Google Scholar: the pros and the cons. Online information review (2005).
[9] Jasser Jasser, Ivan Garibay, Steve Scheinert, and Alexander V Mantzaris. 2020. Controversial information spreads faster and further in Reddit. arXiv preprint arXiv:2006.13991 (2020).
[10] Vishal Kharde, Prof Sonawane, et al. 2016. Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971 (2016).
[11] Heidi M Levitt, Francisco I Surace, Max B Wu, Brad Chapin, Jacqueline G Har- grove, Cara Herbitter, Ethan C Lu, Meredith R Maroney, and Alissa L Hochman. 2020. The meaning of scientific objectivity and subjectivity: From the perspective of methodologists. Psychological methods (2020).
[12] Edward Loep and Steven Bird. 2002. Nltk: The natural language toolkit. arXiv preprint cs/0205028 (2002).
[13] Emilio Delgado-López-Cózar, Enrique Orduña-Malea, Alberto Martín-Martin, and Juan M Aylón. 2017. Google Scholar: the big data bibliographic tool. Research analytics: boosting university productivity and competitiveness through scientometrics (2017), 59.
[14] Alexey N Medvedev, Renaud Lambotte, and Jean-Charles Delvenne. 2017. The anatomy of Reddit: An overview of academic research. In Dynamics on and of Complex Networks. Springer, 183–204.
[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (2011), 2825–2830.
[16] Jeffrey M Perkel. 2018. Why Jupyter is data scientists’ computational notebook of choice. Nature 563, 7732 (2018), 145–147.
[17] Anand Rajaraman and Jeffrey David Ullman. 2011. Data Mining. Cambridge University Press, 1–17. https://doi.org/10.1017/CBO9781139054502.002.
[18] Bernadette M Randles, Irene V Pasquetto, Milena S Golshan, and Christine L Borgman. 2017. Using the Jupyter notebook as a tool for open science: An empirical study. In 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, 1–2.
[19] Yunde Shinya. 2015. Pdfminer: Python pdf parser and analyzer. Retrieved on 2015 (2015).
[20] Mary Shultz. 2007. Comparing test searches in PubMed and Google Scholar. Journal of the Medical Library Association: JMLA 95, 4 (2007), 442.
[21] Giuliana Viglione. 2020. How scientific conferences will survive the coronavirus shock. Nature 582, 7811 (2020), 166–168.
[22] Houyu Zhang, Zhenghao Liu, Chenyuan Xiong, and Zhuyuan Liu. 2019. Grounded conversation generation as guided traverses in commonsense knowledge graphs. arXiv preprint arXiv:1911.02707 (2019).