Link Topics from Q&A Platforms using Wikidata: A Tool for Cross-platform Hierarchical Classification

Alyssa Shuang Sha
Australian National University
Canberra, ACT, Australia
alyssa.sha@anu.edu.au

Bernardo Pereira Nunes
Australian National University
Canberra, ACT, Australia
Bernardo.Nunes@anu.edu.au

Armin Haller
Australian National University
Canberra, ACT, Australia
armin.haller@anu.edu.au

ABSTRACT

This paper proposes a novel rule-based topic classification tool for questions on Q&A platforms mediated by the Wikidata ontology – an open and accessible multilingual ontology curated by a large community of online users. Q&A platforms are important sources of information on the Web and often appear as part of Web search results. By adopting Wikidata taxonomic relations as references, our tool can categories the Web content from different platforms in a unified coarse-to-fine mode based on their domain coverage. To validate and demonstrate the potential applicability of our tool, a set of use cases and experiments are carried out on two popular Q&A platforms – Zhihu and Quora, where the impact of topic categories on question lifecycles is explored. Furthermore, we compare our results with the output generated by GPT-3 classifier. This tool sheds light on how structured knowledge bases can enable data interoperability and serve as a filtering functionality to mitigate classification bias of OpenAI.

CCS CONCEPTS

• **Applied computing** → **Enterprise ontologies, taxonomies and vocabularies**; **Information integration and interoperability**.

KEYWORDS

Topic classification, Wikidata ontology, Entity Linking, Q&A platforms

ACM Reference Format:
Alyssa Shuang Sha, Bernardo Pereira Nunes, and Armin Haller. 2023. Link Topics from Q&A Platforms using Wikidata: A Tool for Cross-platform Hierarchical Classification. In *15th ACM Web Science Conference 2023 (WebSci '23), April 30–May 01, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3578503.3583625

1 INTRODUCTION

Question and Answering platforms (Q&As) often use a set of crowd-sourced tags to help organise and describe their questions. Zhihu[1] and Quora[2] are popular examples of Q&As that rely on folksonomies to classify their questions. Although folksonomies play an important role in helping users to browse and find relevant information on the Web, not all folksonomies are structured, hindering the automatic classification and identification of topic-related questions. This paper proposes a rule-based topic classification tool using the Wikidata ontology as a reference to enable seamless interoperability among multiple platforms which can improve user experience and information searchability.

Wikidata[3] is a large multipurpose Knowledge Graph (KG) containing billions of resources and built on the top of a community-curated ontology composed of items (concrete or abstract entities and classes of entities) and properties (relationships linking entity pairs). Items and properties are identified by Uniform Resource Identifiers (URIs) and represented by alphanumeric codes. Figure 1 depicts part of the Wikidata top-level ontology supported by the *subclass of* (P279) relationship where *entity* Q35120, as the ontology root, is the superclass of all items in Wikidata. Depending on the ontological distance [27] to entity Q35120, the top-level items can be classified as first-layer entities, second-layer entities, third-layer entities, etc. The Wikidata ontology consists of a number of nodes, where further the leaf nodes are from the root, the more specific meaning they represent.

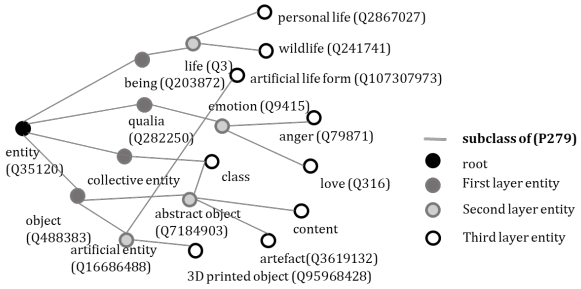

Figure 1: Example of the top-level ontology in Wikidata

Besides items and properties, Wikidata provides external identifiers used in other systems (databases, authority control files, online encyclopedias, etc.) and displays them as links in Wikidata items when the formatter URL (P1630) is defined. Therefore, information from other platforms can be linked to Wikidata entities and categorised following taxonomic relationships in structured KG using external identifiers.

The traffic growth of most questions on Q&As follows a two-stage lifecycle which includes a rapid-growing stage followed by a saturation phase [1, 2, 31]. This study refers to the intersection of the two phases as the "knee" point [30]. To demonstrate how the

[1] http://www.zhihu.com
[2] https://www.quora.com
[3] https://www.wikidata.org

WebSci '23, April 30–May 01, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0089-7/23/04.
https://doi.org/10.1145/3578503.3583625

proposed tool works, we use Zhihu as a data source to explore the relationship between question lifecycles and their topics. Zhihu, the most popular Chinese Q&A platform for users to create, acquire and share knowledge (similar to Quora), relies on labels attached to questions to achieve topic classification. As presented in Figure 2, the Zhihu question "Why Trump became President of the United States?" with a "Donald J. Trump" label attached can be grouped by using the label's identifier (Topic ID: 20023724) generated from Zhihu, which can be linked to the Wikidata entity "Donald Trump" (Q22686). Likewise, the posts under the same topic (Donald Trump) from other platforms (Twitter, Quora, ABC News, etc.) can be connected through Wikidata ontology, which enables the automatic interlinkage with existing datasets. In addition to "Donald J. Trump", "US president", "Presidential Election", "Politics of US", "White House", and "The United States" are attached to the same question in Figure 2. The fact that many labels contain distinct semantic meanings being attached to a single post makes it difficult to categorise the information from various platforms to the topics that have similar semantic coverage. Thus, measuring and unifying the semantic coverage for classification results become pressing. In our tool, Wikidata which is language-independent and enables cross-platform entity linking is adopted as a structured database to achieve data interoperability and hierarchical topic classification among multiple platforms.

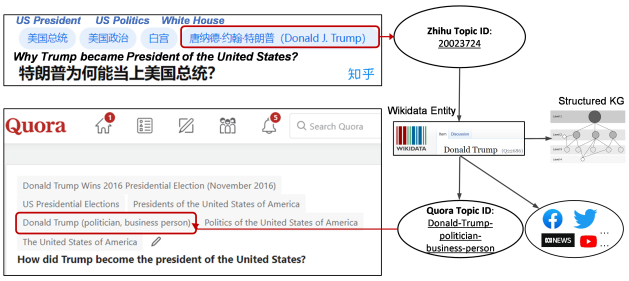

Figure 2: An example of linking questions from Zhihu and Quora using Wikidata entities

The main contributions in this work are three-fold:

(1) Our proposed tool can assist in optimising the existing KG in Q&As by improving the structure and automaticity of topic classification.
(2) The tool enables the creation of a global space with a selected hierarchy for end users, allowing data compatibility and interoperability among multiple platforms.
(3) The hierarchical classification results can serve as a filtering function to optimise the classification results of GPT-3.

2 RESEARCH PROBLEM

Data integration and data interoperability are two major focusing areas for organisations that tend to implement advancements in their workflow. **Data Inconsistency** exists when various and conflicting stories of the same data appear in different places, which has been pointed out as a significant challenge for big data integration and interoperation [13, 14]. More specifically, the data from heterogeneous sources could lead to inconsistent information levels, therefore requiring additional resources to optimise unstructured data.

Finding the standard operational methodologies between two systems for integration and implementing the query operations and algorithms can help meet the challenges of large data entities. This research aims to propose a rule-based solution to manage data inconsistency in cross-platform topic classification by using the structured database.

3 RELATED WORK

Several studies on topic classification focus on extracting relevant information by using tags to predict the popularity or information flow, or to develop clustering methods of online content [4, 12, 21, 28, 33]. Other research looks at tags themselves, trying to analyse their dynamics, popularity, semantics and engagement [5, 15, 17, 32, 37]. However, the tags from many platforms are unstructured, expressed from a single language and semantically duplicated with each other, making the topics on those platforms confusingly categorised without having a uniform structure. Thus, bridging the gap between unstructured and structured data is crucial.

Recent approaches that automatically link unstructured data to structured knowledge bases (e.g., Wikidata) revolve around Entity linking (EL) [7, 11, 16, 34]. Entity linking has been widely applied in NLP [10, 18, 25, 26] on domains including news, biographical text, and movie/show plots. Most of them applied Machine Learning (ML)-based approaches which appear to be costly and time-consuming [34, 35, 38]. A few previous studies have applied EL in the domain of Q&A services focusing on mapping the answers and evaluating the labelling accuracy [8, 23, 24, 29], however, the problem with unifying the semantic coverage of classification results has not been tackled in this field. There are tools proposed by previous research that adopted structured knowledge graphs to achieve web content classification [7, 9, 11, 20, 22, 36], however, they could neither be applied to questions nor took the factor of semantic coverage into account. Without involving a training process, our proposed tool takes advantage of the built-in external identifiers from Wikidata to achieve hierarchical topic classification and enable data interoperability.

4 APPROACH

As shown in Figure 3, our approach consists of three components: A) Top-down entity extraction, B) Bottom-up topic tracing, and C) Applications. We first extract items that have the external identifiers of our target platforms (Zhihu and Quora) from Wikidata and organise them into a hierarchical structure according to their taxonomic relations. Secondly, a bottom-up topic tracing procedure is conducted with a target layer n. After the rule-based mapping, the topics from different platforms are uniformly classified into categories with similar semantic coverage to assist with data interoperability and model fine-tuning.

We use Quora and Zhihu questions as data sources and Wikidata as the reference ontology to demonstrate our approach. The reasons for choosing these two platforms to demonstrate our approach are: 1) they have large user bases, 2) the content features on them are retrievable, 3) the identifiers of their labels exist in Wikidata, and

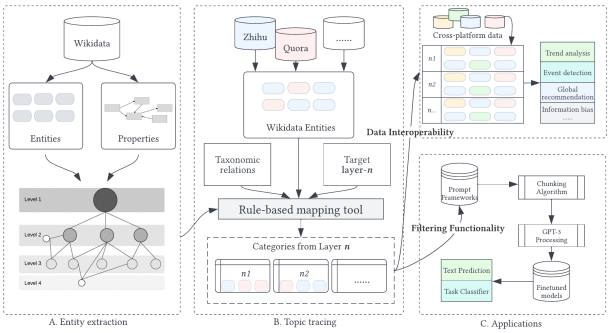

Figure 3: An overview of the data-processing pipeline

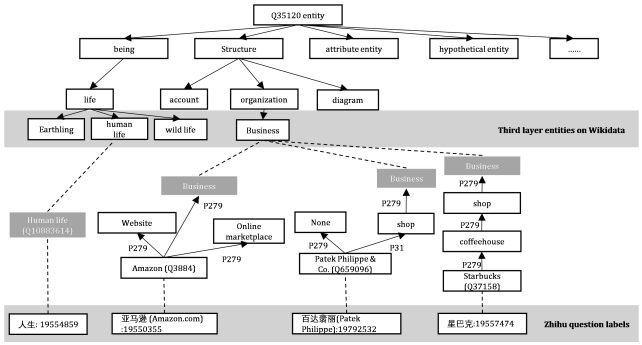

Figure 4: Zhihu labels & Wikidata entities mapping process

4) the current topic structure on Zhihu and Quora can be enhanced by utilising a structured database.

4.1 Data Collection

All the entities from Wikidata that have Zhihu label identifiers were extracted using the Wikidata Query Service (WQS)[4]. Properties P3553 and P3417 were applied respectively for entities have Zhihu or Quora identifiers. In total, we extract 14, 837 Wikidata entities that contain Zhihu topic identifiers and 190, 857 Wikidata entities that contain Quora topic identifiers. Using the Zhihu API and the Python zhihu-oauth[5], 1575 labels (˜5% of the existing Zhihu labels) were randomly collected from Zhihu in April 2022.

Besides the Wikidata entities and Zhihu labels, we also collect the time-series data of questions under the extracted labels to support the use case demonstration in Section 4.2. Zhihu questions were retrieved using their IDs (QIDs) which are always eight to nine digits long and monotonically increase over time. Hence, we can capture newly posted questions and record their traffic (view numbers) by selecting QIDs. The question traffic data were traced 1560 times for one week period, which will be plotted as question lifecycles for us to demonstrate with different topic categories. In order to compare whether the classification tool works for different platforms, 90,940 questions under 101 featured topics are collected from Quora.

4.2 Running Example

The purpose of this running example is to link topics from Q&A platforms to Wikidata entities from a selected layer as an application of our proposed tool. Specifically, the proposed rule-based topic classification approach can be split into two steps: top-down entity extraction and bottom-up topic tracing (shown in Figure 4).

4.2.1 Top-down entity extraction. As introduced in Figure 1, following the *subclass of* property, the tool will traverse from the "root" entity (Q35120) to locate the "leaf" entities at different layers to build a hierarchical entity structure. For example, in Figure 4, entity *being* is placed at the first layer, *life* is at the second, *human life* and *wild life* are located at the third layer.

In order to provide a unified view of data, entities from one specific layer need to serve as the ideal categories for the topic classification task. The layer selection is set as a parameter in the proposed tool that can be adjusted. The entities beyond the third layer are too numerous and specific (e.g., *Starbucks*, *Donald J. Trump*), by using which users might end up with too many categories thus losing the point of classifying them. The first and second layers of labels are again too abstract (e.g., *being*, *object*, *structure*), which makes it difficult for users to have a clear understanding of their semantic meanings. In view of the above considerations, third-layer entities are set as our classification entries in this running example to show the audience how to classify questions into categories which cover a similar range of meanings.

4.2.2 Bottom-up topic tracing. A bottom-up approach is applied to map Zhihu topic labels to the third layer Wikidata entities by using taxonomic properties forwardly. After data collection procedures, a set of tuples containing the Zhihu topic identifier and its corresponding Wikidata URI are generated using WQS.

By running the URI in SPARQL query under the *subclass of* and *instance of* relations, the classification tool can track the entity's 'parents', 'grandparents', and so on until it reaches the entity at previously selected layer. The mapping steps are listed in Algorithm 1. The inputs are the graph from Wikidata with *subclass of* (P279) and *instance of* (P31) properties specified; dataset Z that contains Zhihu topic IDs that we crawled randomly; dataset W that has all Zhihu identifiers and their corresponding entity URIs from Wikidata; L_1, L_2, L_3 that contain the 1st, 2nd and 3rd layer entities.

ALGORITHM 1: MAPPING ZHIHU TOPICS TO THE 3RD LAYER WIKIDATA ENTITIES

Input: entities E, graph $G = (E, R, S)$, $P279 \in R$, $P31 \in R$, Z (a set that contains zhihu *topic IDs*), $W = \{(i, f(i)),\ i \in N\}$, L_1, L_2, L_3
Output: $\Gamma = \{(e_i, z_i),\ i \in N\}$

```
for z in Z do
    if z ∈ w_i then
        e ← w_f(z)
        if e ∈ L_1 or e ∈ L_2 then
            Γ_z ← (none, z)
        else
            while e ∉ L_3 do
                while G(e, P279, s) = ∅ do
                    for é ∈ E do
                        if é ∈ G(e, P31, s) then
                            e ← é
                for é ∈ E do
                    if é ∈ G(e, P279, s) then
                        e ← é
            Γ_z ← (e, z)
    else
        Γ_z ← ∅
return Γ
```

[4] https://query.wikidata.org
[5] https://pypi.org/project/zhihu-oauth/

Figure 4 depicts several special occasions that were managed by our algorithm individually. Zhihu labels that do not have Wikidata identifiers are assigned with an empty value for their output tuple. Some labels (e.g., *Human life*) can be directly mapped to Wikidata third-layer entities. Some of the labels (e.g., *Amazon*) might have multiple 'parents', the one fall under the third layer Wikidata entities is selected as the final category. If none of them belongs to the third layer entities, the tool then continues to track their 'grandparents' until one of them does. For the bottom entities that do not have an upper class in Wikidata (e.g., *Patek Philippe &Co.*), P31 property is applied to find their instance and then traverse the instance's upper class to continue the mapping. Finally, the output for this algorithm is a set of tuples where each tuple includes the Zhihu Topic ID and the successfully mapped third layer Wikidata entity.

5 APPLICATIONS

This section presents how to use the proposed tool to classify questions into topic categories that have consistent semantic coverage and visualise the lifecycle patterns of their traffic growth. We also compare our results with GPT-3, a popular language model developed by OpenAI, and demonstrate how to use the classification result to improve the performance of GPT-3.

5.1 Cross-platform Topic Classification Results

Previous research has tried to analyse the time-sensitivity of Zhihu topics by clustering them according to the recurring frequencies of those topic labels [31]. Without a standard topic classification scheme, it is difficult to relate the recurring features of topics to their semantic meanings. Our proposed tool provides a way to classify labels with an inconsistent range of meanings to categories that have the same semantic coverage.

After completing the mapping procedures, 581 Zhihu labels and 98 Quora labels are successfully classified as Wikidata third-layer entities. Table 1 shows some samples from our classified data. Not limited to the third layer entities, our proposed method enables users to choose needed topic classification depths by adjusting the parameters from the running example. The Chinese labels from Zhihu are automatically translated into English in the topic tracing process.

5.2 A structural way to explore question lifecycles

The proposed classification tool can be also used to investigate the characteristics of topics at different semantic levels as well as the relationship between topic meanings and the lifecycle of corresponding questions. In Figure 5, we plot the classified 3rd layer topics into a coordinate system consisting of their popularity and the median knee points (median time taken for their traffic to reach saturation) of their related questions. The statistical result shows no significant relationship between the popularity of the questions and the time it takes for the two-stage questions to reach the saturation point. The time to reach the second lifecycle stage varies greatly between topics with different levels of popularity. Besides, we can observe the questions' lifecycle under the categories with similar semantic coverage. Figure 6 and 7 give examples of the

Table 1: Part of the topic classification results on Zhihu and Quora

2nd layer entity	3rd layer entity	Zhihu Labels	Quora Labels
Organization (Q43229)	business (Q4830453)	百达翡丽-Patek Philippe & Co. 中国联通-China Unicom 网易游戏 NetEase Games 阿里巴巴集团 Alibaba Group…	Google Small-business Startups Amazon ……
Abstract object (Q7184903)	Information (Q11028)	作文-composition 数值分析-numerical analysis 战略管理-strategic management…	Computer-Programming Journalism Recipes History-of-the-United-States…
System (Q58778)	science (Q336)	分析化学-analytical chemistry 心理学-psychology 物理学-physics 几何学-geometry…	Computer-Science Economics International-Relations Neuroscience…
Service (Q7406919)	Education (Q8434)	上海教育-education in Shanghai 化学教育-chemical education 成人教育-adult education…	Health-Promotion Higher-Education Language-Education…
Goods (Q28877)	Entertainment (Q173799)	游戏-game 音乐-music…	Music Rock-Music TV-Series…

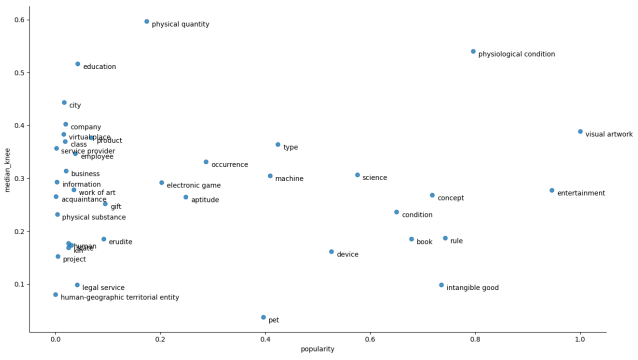

Figure 5: Knee point and popularity relationships for topics mapped to Wikidata 3rd layer entities

lifecycle of questions under *city* (Q515) and *occurrence* (Q1190554) categories. The normalised median knee point of the *city* questions is 0.4437, which is above the median knee (0.3231) of *occurrence* questions. The saturation point arrival time is relatively late for questions with substantial 2-stage lifecycles. For example, the life pattern of the Shanghai City topic reached a plateau roughly 100 hours after it was posted. On the other hand, the occurrence topic is relatively more time-sensitive, as the majority of the questions in our sample group reached their knee points within 40 hours of being posted.

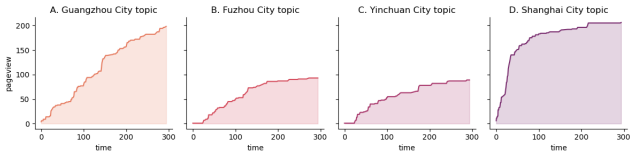

Figure 6: Lifecycle examples of the questions under city (Q515) category

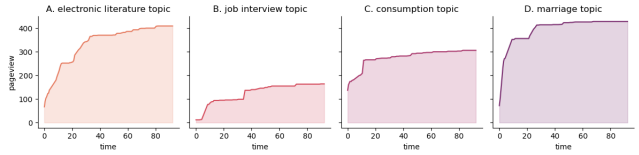

Figure 7: Lifecycle examples of the questions under occurrence (Q1190554) category

5.3 Compare with and fine-tune GPT-3

The Generative Pre-trained Transformer 3 (GPT-3) is an autoregressive language model that has shown powerful in-context few-shot learning abilities [3]. Developed by OpenAI, GPT-3 can also classify words and phrases accurately into pre-determined categories [6]. Following the classification example provided by OpenAI, we adopt a base engine (text-davinci-002) using the endpoint API to conduct unsupervised classification for the labels of Quora and Zhihu. "Temperature" is set to 0 and "Top p" is set to 1. Our purpose for involving this component is to investigate how the proposed tool performs differently from GPT-3 and how we can contribute to existing models to achieve more sophisticated classification (refer to Figure 3-C.Applications).

We pick out a small portion of the final classification results shown in Figure 8 to empirically illustrate some issues with GPT-3 classifier. Firstly, although the base model can correctly classify labels from different platforms into the relevant categories before being fine-tuned, it **lacks consistency in the semantic coverage for the final categories** (e.g., same-level cities and universities are classified into different categories). GPT-3 classifier does not cope with a hierarchical structure as it is built based on relatedness rather than semantic subordination and affiliation. This flat and discrete categorization structure lacks the strengths of generalization, which causes difficulty in bringing the data to the same level to analyze. GPT-3 classifier does not cope with a hierarchical structure as it is built based on relatedness rather than semantic subordination and affiliation. In order to inform the model how specific (or general) we want the final categories to be, we use 20% of our previously classified data as the guidance (prompt input) to fine-tune the base model to generate the results with our desired semantic coverage. After fine-tuning, the consistency of classification results has been improved for items 1 to 5, and the final categories appear to be broader (item 6 and 7) following the few-shots guidance.

Compared to our results, the other issue that appears in the classification results from GPT-3 is **representation bias**. For instance, "Same-sex marriage" is classified as a "social issue", and "confession" is classified as a "religious term". Similar issues for text generation have been pointed out by previous research on GPT-3 [19]. Because of the sample bias in the database used to train the language model, it can parrot or even amplify social biases. Item 8 shows that adding a few prompt inputs does not seem to help GPT-3 avoid this issue. Besides, "Project 985" and "Project 211" as well-known Chinese projects for higher education, failed to be recognized by GPT-3. However, a similar label, "Ivy League", can be classified usefully. One possible reason could be that the corpus of different languages and cultures used to train the model is unbalanced, which leads to models' limited understanding and perspectives for certain concepts.

		itemLabel	Wikidata_layer3_entity	GPT3_beforeFT	GPT3_afterFT
Inconsistent categories	1	Guangzhou	city	Geography	location
	2	Shenzhen	city	city	location
	3	Yinchuan	city	place	location
	4	Civil Aviation University of China	service provider	Education	institution
	5	Tsinghua University	service provider	School	institution
Lacks generalization	6	Starbucks	business	Coffee	company
	7	Cristiano Ronaldo	human	Athlete	person
Bias	8	same-sex marriage	occurrence	social issues	relationship
	9	homosexuality	social group	lifestyle	social issue
	10	confession	information	religious term	genre
Limited understanding	11	Project 985	project	NA	NA
	12	Project 211	project	NA	NA
	13	Ivy League	tournament system	school	education
	...	.....	.....	.....	.....

Figure 8: Part of the comparison of results with GPT-3 (before and after the fine-tuning)

6 CONCLUSION

This research proposed a novel rule-based approach to classify unstructured topics from Q&As into Wikidata entities with a selected level of domain coverage [6]. Our method builds a bridge between unstructured information and structured data by demonstrating how Wikidata can help to optimise the topic structure from other platforms. By addressing the semantic inconsistency problem existing in question labels, this study can assist in classifying and analysing Web content with different languages and formats under a unified scheme. Relying on the external identifiers from Wikidata, our proposed tool could be easily applied to other social media, News platforms, or KGs to create a global space for information retrieval. Furthermore, our approach can serve as a benchmark to evaluate the performance of similar hierarchical classification tasks and provide filtering functionality to fine-tune existing classifiers based on the customized semantic depth. Future work involves enhancing this work by adopting LDA-driven tools with Wikifier to allow direct correspondence from text to Wikidata entities. Interestingly, we point out some issues in the model of OpenAI according to the comparison of the classification results. It is worthwhile to explore further how to re-balance the training data to reduce the bias of the language models in future studies.

REFERENCES

[1] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Discovering value from community activity on focused question answering sites: A case study of stack overflow. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 850–858. https://doi.org/10.1145/2339530.2339665

[2] Vasudev Bhat, Adheesh Gokhale, Ravi Jadhav, Jagat Pudipeddi, and Leman Akoglu. 2015. Effects of tag usage on question response time: Analysis and prediction in StackOverflow. *Social Network Analysis and Mining* 5, 1 (1 2015), 1–13. https://doi.org/10.1007/s13278-015-0263-3

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda

[6] Our code and datasets are available at https://github.com/alyssa-sha/Zhihu-labels-classification-by-using-Wikidata-identifiers

Askell. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 4. https://ojs.aaai.org/index.php/ICWSM/article/view/14033

[5] Riley Crane and Didier Sornette. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences* 105, 41 (10 2008), 15649–15653. https://doi.org/10.1073/pnas.0803685105

[6] Robert Dale. 2021. GPT-3: What's it good for? *Natural Language Engineering* 27, 1 (2021), 113–118. https://doi.org/10.1017/S1351324920000601

[7] Antonin Delpeuch. 2020. OpenTapioca: Lightweight entity linking for Wikidata. In *CEUR Workshop Proceedings*, Vol. 2773.

[8] Davi Faisca Duarte, Sean Wolfgand Matsui Siqueira, and João Luís Tavares da Silva. 2018. Exploring the Correlation of Semantic Entities Between Questions and Answers in Q&A Communities. In *Proceedings of the Euro American Conference on Telematics and Information Systems*. 1–5.

[9] Lars C. Gleim, Rafael Schimassek, Dominik Hüser, Maximilian Peters, Christoph Krämer, Michael Cochez, and Stefan Decker. 2020. SchemaTree: Maximum-Likelihood Property Recommendation for Wikidata. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12123 LNCS (2020), 179–195. https://doi.org/10.1007/978-3-030-49461-2{_}11

[10] Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2681–2690.

[11] Bahareh Harandizadeh and Sameer Singh. 2020. Tweeki: Linking Named Entities on Twitter to a Knowledge Graph. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*. 222–231. https://doi.org/10.18653/v1/2020.wnut-1.29

[12] Liangjie Hong, Ovidiu Dan, and Brian D Davison. 2011. Predicting Popular Messages in Twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web (WWW '11)*. Association for Computing Machinery, New York, NY, USA, 57–58. https://doi.org/10.1145/1963192.1963222

[13] Anirudh Kadadi, Rajeev Agrawal, Christopher Nyamful, and Rahman Atiq. 2014. Challenges of data integration and interoperability in big data. In *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*. IEEE, 38–40. https://doi.org/10.1109/BigData.2014.7004486

[14] Dattatray R Kale and Smita Y Aparadh. 2016. A Study of a Detection and Elimination of Data Inconsistency in Data Integration. *International Journal of Scientific Research in Science, Engineering and Technology IJSRSET* 1, 1 (2016), 532–535.

[15] Janette Lehmann, Bruno Gonçalves, José J Ramasco, and Ciro Cattuto. 2012. Dynamical Classes of Collective Attention in Twitter. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. Association for Computing Machinery, New York, NY, USA, 251–260. https://doi.org/10.1145/2187836.2187871

[16] Weichen Li, Patrick Abels, Zahra Ahmadi, Sophie Burkhardt, Benjamin Schiller, Iryna Gurevych, and Stefan Kramer. 2021. Topic-Guided Knowledge Graph Construction for Argument Mining. In *Proceedings - 12th IEEE International Conference on Big Knowledge, ICBK 2021*. IEEE, 315–322. https://doi.org/10.1109/ICKG52313.2021.00049

[17] Yu-Ru Lin, Drew Margolin, Brian Keegan, Andrea Baronchelli, and David Lazer. 2013. #Bigbirds Never Die: Understanding Social Dynamics of Emergent Hashtags. *Proceedings of the International AAAI Conference on Web and Social Media* 7, 1 SE - Full Papers (6 2013), 370–379. https://ojs.aaai.org/index.php/ICWSM/article/view/14407

[18] Xiao Ling, Sameer Singh, and Daniel S Weld. 2015. Design Challenges for Entity Linking. *Transactions of the Association for Computational Linguistics* 3 (6 2015), 315–328. https://doi.org/10.1162/tacl{_}a{_}00141

[19] Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*. 48–55.

[20] Jerry Fernandes Medeiros, Bernardo Pereira Nunes, Sean Wolfgand Matsui Siqueira, and Luiz André Portes Paes Leme. 2018. TagTheWeb: Using Wikipedia Categories to Automatically Categorize Resources on the Web. In *The Semantic Web: ESWC 2018 Satellite Events*, Aldo Gangemi, Anna Lisa Gentile, Andrea Giovanni Nuzzolese, Sebastian Rudolph, Maria Maleshkova, Heiko Paulheim, Jeff Z Pan, and Mehwish Alam (Eds.). Springer International Publishing, Cham, 153–157.

[21] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Bad News Travel Fast: A Content-Based Analysis of Interestingness on Twitter. In *Proceedings of the 3rd International Web Science Conference (WebSci '11)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2527031.2527052

[22] Liqiang Nie, Yongqi Li, Fuli Feng, Xuemeng Song, Meng Wang, and Yinglong Wang. 2020. Large-Scale Question Tagging via Joint Question-Topic Embedding Learning. *ACM Transactions on Information Systems* 38, 2 (2020). https://doi.org/10.1145/3380954

[23] B P Nunes, A Mera, R Kawase, B Fetahu, M A Casanova, and G H B de Campos. 2014. A Topic Extraction Process for Online Forums. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*. 541–543. https://doi.org/10.1109/ICALT.2014.158

[24] Bernardo Pereira Nunes, Ricardo Kawase, Besnik Fetahu, Marco A Casanova, and Gilda Helena B de Campos. 2014. Educational forums at a glance: Topic extraction and selection. In *International Conference on Web Information Systems Engineering*. Springer, 351–364.

[25] Priya Radhakrishnan, Partha Talukdar, and Vasudeva Varma. 2018. Elden: Improved entity linking using densified knowledge graphs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1844–1853.

[26] Jonathan Raiman and Olivier Raiman. 2018. Deeptype: multilingual entity linking by neural type system evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[27] Sylvie Ranwez, Vincent Ranwez, Jean Villerd, and Michel Crampes. 2006. Ontological Distance Measures for Information Visualisation on Conceptual Maps. In *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, Robert Meersman, Zahir Tari, and Pilar Herrero (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1050–1061.

[28] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the Mechanics of Information Diffusion across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. Association for Computing Machinery, New York, NY, USA, 695–704. https://doi.org/10.1145/1963405.1963503

[29] Gaetano Rossiello, Nandana Mihindukulasooriya, Ibrahim Abdelaziz, Mihaela Bornea, Alfio Gliozzo, Tahira Naseem, and Pavan Kapanipathi. 2021. Generative relation linking for question answering over knowledge bases. In *The Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 20*. Springer, 321–337.

[30] Ville Satopää, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *Proceedings - International Conference on Distributed Computing Systems*. 166–171. https://doi.org/10.1109/ICDCSW.2011.20

[31] Alyssa Shuang Sha, Armin Haller, and Yingnan Shi. 2022. Effects of Label Usage on Question Lifecycle in Q&A Community. In *Thirtieth European Conference on Information Systems (ECIS 2022)*.

[32] David A Shamma, Lyndon Kennedy, and Elizabeth F Churchill. 2011. Peaks and Persistence: Modeling the Shape of Microblog Conversations. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW '11)*. Association for Computing Machinery, New York, NY, USA, 355–358. https://doi.org/10.1145/1958824.1958878

[33] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. 2010. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *2010 IEEE Second International Conference on Social Computing*. 177–184. https://doi.org/10.1109/SocialCom.2010.33

[34] Houcemeddine Turki, Dennis Priskorn, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, and Alejandro Piad-Morffis. 2022. *Enhancing multilingual and biomedical named entity recognition using Wikidata semantic relations ACM Reference Format*. Vol. 1. Association for Computing Machinery. https://pypi.org/project/langdetect/

[35] Xu Wang, Chen Yang, and Renchu Guan. 2018. A comparative study for biomedical named entity recognition. *International Journal of Machine Learning and Cybernetics* 9, 3 (2018), 373–382. https://doi.org/10.1007/s13042-015-0426-6

[36] Tianxing Wu, Guilin Qi, Cheng Li, and Meng Wang. 2018. A Survey of Techniques for Constructing Chinese Knowledge Graphs and Their Applications. *Sustainability* 10, 9 (2018). https://doi.org/10.3390/su10093245

[37] Jaewon Yang and Jure Leskovec. 2011. Patterns of Temporal Variation in Online Media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. Association for Computing Machinery, New York, NY, USA, 177–186. https://doi.org/10.1145/1935826.1935863

[38] Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. 2019. CollaboNet: collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinformatics* 20, 10 (2019), 249. https://doi.org/10.1186/s12859-019-2813-6