Content Modeling in Multi-Platform Multilingual Social Media Data

Arman Setser, Libby Lange, Kyle Weiss, and Vladimir Barash

Abstract. An increase in the use of social media as the primary news source for the general population has created an ecosystem in which organic conversation commingles with inorganically seeded and amplified narratives, which can include public relations and marketing activity but also covert and malign influence operations. An efficient and easily understandable analysis of such data is important, as it allows relevant stakeholders to protect online communities and free discussion while better identifying activity and content that may violate social media platform terms of service. To accomplish this, we propose a method of large-scale social media data analysis, which allows for multilingual conversations to be analyzed in depth across any number of social media platforms simultaneously. Our method uses a text embedding model, i.e., a natural language processing model that holds semantic and contextual understandings of language. The model uses an "understanding" of language to represent posts as coordinates in a high-dimensional space, such that posts with similar meanings are assigned coordinates close together. We then cluster and analyze the posts to identify online topics of conversation existing across multiple social media platforms. We explicitly show how our method can be applied to four different datasets, three consisting of Chinese social media posts related to the Belt and Road Initiative and one relating to the Russia-Ukraine war, and we find politically-influenced conversations that contain misleading information relating to the Chinese government and the Russia-Ukraine war.

1 Introduction

Social media is one of the largest sources of news and information in the developed world. However, coordinated and inauthentic manipulation and malign influence operations have flourished in this new media ecosystem (Howard et al. 2017). Analysis of large-scale social media data is important, as it allows one to distinguish between natural conversation and the coordinated spread of harmful or misleading narratives. Common forms of analysis usually consist of considering simple text-based features (e.g., words and hashtags) and/or network features such as user-user follower relationships. These features and relationships are analyzed on a single platform, e.g., X (formerly Twitter),

10.54501/jots.v2i2.136

to identify groups of users spreading similar content, or to identify key content spread by users (Shahsavari et al. 2020; Shu et al. 2017; Benigni, Joseph, and Carley 2017; Jachim, Sharevski, and Pieroni 2020). Such methods are powerful, as they allow cohesive communities of online discussion to be identified, which can then be further manually evaluated by human experts. The downside of this approach is that it is dependent on data specific to the social media network(s) in question. For example, if one were analyzing data from multiple social media platforms simultaneously, there would obviously be no direct links between users on different platforms, e.g, a user on X, would not directly be a friend or follower of a user on Reddit. Therefore, links between platforms would be found only through intermediary pieces of content, e.g., hashtags and URLs, which appear on multiple platforms. This can make capturing communities across multiple platforms difficult, as such intermediary links typically lack important semantic and contextual information, e.g., whether they were used positively or negatively. With this approach, it can also be difficult to capture multilingual communities across platforms, since the intermediary platform-linking pieces of content can typically only link a language on one platform to the same language on another platform.

We propose an alternative method that consists of using a text embedding-based model to analyze social media posts. A text embedding model is a natural language processing model that represents texts as coordinates in a high-dimensional space. The model is pretrained to understand the context and semantics of texts, such that texts with similar meanings are assigned coordinates close together (Kalyan, Rajasekharan, and Sangeetha 2021; Acheampong, Nunoo-Mensah, and Chen 2021). We use a multilingual model pretrained on all common languages we encounter in our datasets of interest, e.g., English, Russian, Ukrainian, and Chinese, which allows the model to perform accurately across all posts we consider (Reimers and Gurevych 2019). Since similar-meaning posts are assigned close coordinates, we then apply a clustering algorithm to identify groups of closely packed points, which will correspond to texts focused on a specific topic of conversation. Analysis of these groups will allow one to understand the individual topics being discussed in a conversation landscape, as well as their relation to each other.

Previous work has shown success in analyzing social media data using specific language features (Sawhney et al. 2020; Mozafari, Farahbakhsh, and Crespi 2019; Kumar Kaliyar, Goswami, and Narang 2021; Xu et al. 2020; Rogers, Kovaleva, and Rumshisky 2019; Zhang and Pan 2019). However, such works are more limited in scope, either classifying data into a set of predetermined categories, requiring a large amount of human intervention to assist in the data processing, or only considering data with specific features/outcomes such as calls to action, offline mobilization, event detection, and disease tracking. Our proposed approach is more general than previous works, capturing the many topics that can occur in the general conversation of a social media landscape, without requiring a human to predefine the categories or assist in their classification. As a result, features such as calls to action can be derived from our method through human analysis of the relevant conversation landscape. The method application is network agnostic, in the sense that it is completely independent of relationships between users, allowing connections between platforms to be found more easily. More subtly, the method results can depend on the platforms at hand, since conversations can be biased toward different topics on different platforms.

Our approach uncovers distinct groups of conversation in a collection of posts/documents, similar to a topic model (Vayansky and Kumar 2020). A standard topic modeling approach uses statistical techniques to find a limited set of topics for a set of unstructured texts, whereas the method we propose aims to produce a structure for online conversation that can then be analyzed to find associated topics. In our approach, posts with similar meaning will be clustered together, which also implies that clusters with similar

meanings will be located near each other. Then, by analyzing a snapshot of clustered post embeddings, one can see not only the topics represented as clusters, but also the relationships between those topics, represented by clusters in close proximity. Although similar to a topic model, our method offers more detail through providing not only the topics of conversation but also the relationships between those topics.

We demonstrate our method by applying it to data relating to the Chinese Belt and Road Initiative (BRI) collected from Chinese social media platforms Weibo (a Chinese microblogging website) and WeChat (a Chinese messaging and social media app). The BRI is of particular interest, as it is a large-scale Chinese government infrastructure project aimed at increasing China's standing and influence among neighboring countries. Thus, an understanding of the internal Chinese-language BRI-related conversation taking place on Chinese social media platforms is key in understanding the tactics used and messages spread by the Chinese government, including potentially misleading messages intended to influence target audiences. Weibo and WeChat were chosen because of their high use in China, as well as the ease of access to data on those platforms for a non-Chinese audience, relative to other, smaller China-focused social media platforms.

Data relating to the Russia-Ukraine war was also collected and analyzed to demonstrate the effectiveness of the model on additional languages. This data was collected in English, Russian, and Ukrainian languages from X, Reddit, Facebook, Instagram, YouTube, VKontakte (a Russian-based social media platform), and Telegram (a messaging app). In addition to being multilingual, this dataset provides another opportunity to analyze a dataset rife with disputed and misleading claims concerning the war, e.g., casualties, battle victories, and military aid. These seven platforms were chosen because of their ability to capture a wide range of conversations around the war. The application of our methods to this dataset will show the ability of the model to capture multiple viewpoints surrounding a topic, and show how those viewpoints relate in a larger conversation space.

Using the above datasets, we show that the contextually rich multilingual clusters created using our method provide a complete and insightful mapping of online conversation across various social media platforms. The multi-platform nature of the datasets also allows us to show how conversations vary on different social media platforms. It is important to note that our method is not itself a form of automatic information classification, e.g., what some would call malicious information, misinformation, or disinformation. Rather, it is a tool that should be used by a human expert with relevant domain knowledge to break down an online conversation landscape into its individual topics, which can then be manually classified. Our proposed method allows a human expert to make judgements on these individual topics, and investigate further, rather than having to parse thousands of unstructured texts at once.

2 Data

As a relevant test of our process, we considered three separate BRI-related datasets from Chinese social media platforms WeChat and Weibo, which were obtained through Meltwater, a data provider (Meltwater, n.d.). The posts were gathered by matching them to a query created by a subject matter expert, which we present in Appendix A. Note that the queries used to pull the data were in Chinese, although we present English translations as well. All posts were created within the two months prior to when the data was pulled, i.e., October and November 2022. As mentioned in the introduction, this is a particularly insightful use case, as internal Chinese social media data is well known to be biased toward China-positive content (King, Pan, and Roberts 2013). Therefore,

an analysis of Chinese government-related topics on Chinese social media can produce insights into how the Chinese government spreads positive-biased information targeted toward its own citizens. Our subject matter experts have previously observed that WeChat is typically used by government-related organizations to disseminate information, while Weibo tends to be marketed as a "breaking news" platform, suggesting that it is more likely to be used for developing events and stories (Guo and Zhang 2020).

The first dataset consisted of broad BRI-related terms, which resulted in over 400,000 total posts before processing. The second dataset consisted of Southeast Asia BRI railway projects, yielding approximately 10,000 posts before processing, and the third dataset focused on the China–Pakistan economic corridor (CPEC), yielding approximately 10,000 posts before processing. Counts of obtained posts for each query term are also shown in Appendix A. In all three cases, the Weibo data is no more than 3% of the size of the WeChat data after processing and clustering. This is likely because BRI-related news stories are typically not "breaking" or "urgent," which leads to relatively low discussion on Weibo due to its "breaking news" nature, as mentioned above. Due to the WeChat dominance, the majority of the clusters we found for the three BRI datasets were predominantly WeChat-centric.

The large imbalance in the size of the three datasets is reasonable, since it reflects the fact that the general BRI query is much more broad than the other queries. In our analysis, we show how the specificity of the CPEC and Southeast Asia queries leads to highly specific clusters focused on detailed events and conversation, whereas the general BRI query provides a high-level overview of BRI-related conversation. Due to the large amount of data collected in the general BRI query, we randomly selected a sample of 50% of the data and used that sample in our work.

We also considered an additional dataset related to the Ukrainian counteroffensive in the Russia-Ukraine war with data taken from X, Reddit, Facebook, Instagram, Telegram, VKontakte, and YouTube in English, Russian, and Ukrainian. Instagram and YouTube text data was taken from comments and images/video descriptions made on the platforms. The data was also gathered using Meltwater, with a date range of September 15, 2023–September 22, 2023. This resulted in just under 100,000 posts; the queries are shown in Appendix A. Analysis of this dataset will demonstrate the capability of the model to simultaneously analyze texts in multiple languages. Using data from many different platforms at once will also provide a dataset with both Russia and Ukraine-biased posts, which we show leads to distinct discoverable topics using our method.

3 Methods

3.1 Data Processing

We consider a set of users' posts across the various social media platforms we seek to analyze. We first perform a data cleaning step to ensure that all data we use in the model is standardized. This consists of removing emojis, removing URLs, removing line breaks/tabs/extra spaces, enforcing a maximum length for the text, removing appearances of "@username," removing any appearances of "RT" and "QT" that appear at the beginning of texts, and then dropping duplicate texts.

Although there are some cases when one user mentions another prominent user, many cases of one user mentioning another are irrelevant, e.g., a friend mentioning a friend, thus leading to the removal of user mentions. Removing "RT" and "QT" at the beginning of texts is done for data on all platforms, although it is motivated by X data, which includes these tags when messages are retweets or quote tweets of another. Although URLs can

contain valuable text data, such as within the title or text of a linked article, we choose to simply remove all links for standardization purposes, since not all links are guaranteed to contain useful text information. The maximum text length is typically determined by the model; we explain this choice in the next section.

3.2 Embedding and Clustering

After standardizing the posts, we then use a text embedding model to embed the set of posts. The text embedding model is a pretrained natural language processing model that represents each text in a high-dimensional coordinate space (Lin et al. 2022). Ideally, the contextual understanding of the model will ensure that coordinates belonging to contextually similar posts will be close to each other in the space, whereas noncontextually similar posts will have more separated coordinates. The specific text embedding model we use is "text-embedding-ada-002," the most powerful text embedding model available through the OpenAI platform for our use case, which is capable of simultaneously handling all languages we consider in this work (English, Russian, Ukrainian, and Chinese) (OpenAI 2022). We enforce a maximum tokenized text length of 8,191, which is the maximum allowed by the model, where the tokenization is handled by the model-associated tokenizer, "tiktoken." For each input text, the model produces a 1,536-dimensional coordinate point, which we refer to as the embedding for that text.

We use the HDBSCAN density-based clustering algorithm to cluster the set of embeddings, in order to pick out groups of points located closely in the embedding space, which will uncover clusters of contextually similar posts (McInnes, Healy, and Astels 2017). In order to make the clustering process more tractable, we first apply UMAP, a dimensionality reduction algorithm, to the embedding set to reduce the dimension of the space from 1,536 to 10 (McInnes, Healy, and Melville 2018). This step is necessary to reduce the data size, such that the clustering algorithm can be run in a reasonable amount of time (typically less than 10 minutes). The UMAP algorithm contains two parameters that affect the results of the dimensionality reduction process: (1) the minimum allowed distance between points after dimensionality reduction, and (2) the choice of *n_neighbors*, which determines the balance between local and global structure of the data during the reduction process. We choose a minimum allowed distance between points of 0, as this impacts the results only visually, and an *n_neighbors* value of 100, as we found this to produce an appropriate—as qualitatively assessed by our human experts—separation of texts into unique clusters, while placing similar-meaning clusters close together in the low-dimension space.

The clustering output is highly dependent on the choice of parameters used in HDBSCAN. In order to account for this, we consider three separate metrics that can be used to measure the quality of a clustering output: (1) the percent of objects clustered, (2) the average silhouette score of the clusters, which measures how distinguished the clusters are, and (3) the number of clusters. The adjustable parameters we consider in HDBSCAN are the minimum allowable size of a cluster (*min_cluster_size*), the maximum distance that points in the embedding space can be separated by within a cluster (*cluster_selection_epsilon*), the minimum number of samples required for a point to be considered a "core point" in the HDBSCAN algorithm (*min_samples*), and the clustering type (*cluster_type*), either leaf or Excess of Mass. We optimize over these parameters using a grid search method, searching multiple orders of magnitude for each numeric parameter and searching each category for the clustering type, so that the final clustering output has as least 70% of the texts clustered, an average silhouette score of at least 0.3, and a number of clusters between 50 and 100. These choices have been shown by our human experts to produce distinct and insightful clusters, as demonstrated in our results. For all datasets, the searched values of the parameters were [0.001, 0.002,

0.003, 0.004, 0.005] * *N* for *min_cluster_size*, [0.25, 0.5, 0.75, 1] * *min_cluster_size* for *min_samples*, and [0.001, 0.01, 0.1] for *cluster_selection_epsilon*, where *N* is the number of data points. After clustering the data in 10 dimensions, we use UMAP to further reduce the dimensionality to 2 for ease of visualization.

3.3 Labeling

After the clustering algorithm was applied, the resulting clusters were labeled by gathering several representative texts from each cluster and using the gpt-3.5-turbo model, available through the OpenAI API, to provide a representative label for the cluster (OpenAI, n.d.). The gpt-3.5-turbo model is the most powerful chat model currently available through the OpenAI API, which is capable of proving high-quality, accurate, and knowledgeable natural language responses to a provided prompt. The provided prompt was "These are a set of texts for a narrative: {representative texts}. Find a single label in less than five words that encompasses the narrative, being as specific as possible. Present only a single label." The representative texts were taken to be the five central-most texts within each cluster, which is reasonable since those texts will be at the densest region of the cluster, and the clusters were created based on a density-based clustering algorithm. These labels are not intended to be a perfectly accurate representation of the content within each cluster. Rather, they are intended to be a high-level overview of the content within each cluster, that can then be used in conjunction with human analysis. In our work, we present these AI-generated labels. However, we note that several of the authors who are subject matter experts on the dataset topics we analyzed, i.e., the Chinese government and the Russia-Ukraine war, have also manually analyzed the clusters of posts to verify their quality and relevance, as well as to extract useful analytic insights. This human-extracted information was then compared to the AI-generated labels, and the generated labels were determined to be accurate, high-quality representations of the clusters in all cases.

4 Results

4.1 General BRI Projects

The posts for this dataset were embedded and clustered and the resulting clusters were labeled, all following our methods above. The resulting unlabeled clusters are shown for the general BRI dataset in Figure 1; the labels are shown in Table 8 (p. 22). The optimal HDBSCAN parameters are shown in Table 1.

Table 1: Optimal HDBSCAN parameters for each of the analyzed datasets, determined via gridsearch.

Parameter	General BRI Projects	China–Pakistan Economic Corridor	Southeast Asia Projects	Russia-Ukraine War
cluster_type	excess mass	excess mass	excess mass	excess mass
cluster_selection_epsilon	0.001	0.1	0.1	0.1
min_cluster_size	188	29	12	18
min_samples	47	7	3	18

Our analysis identified 55 distinct clusters, with 37,425 of the clustered posts on WeChat and 1,157 of the posts on Weibo. Clusters were generally focused on broad BRI topics,

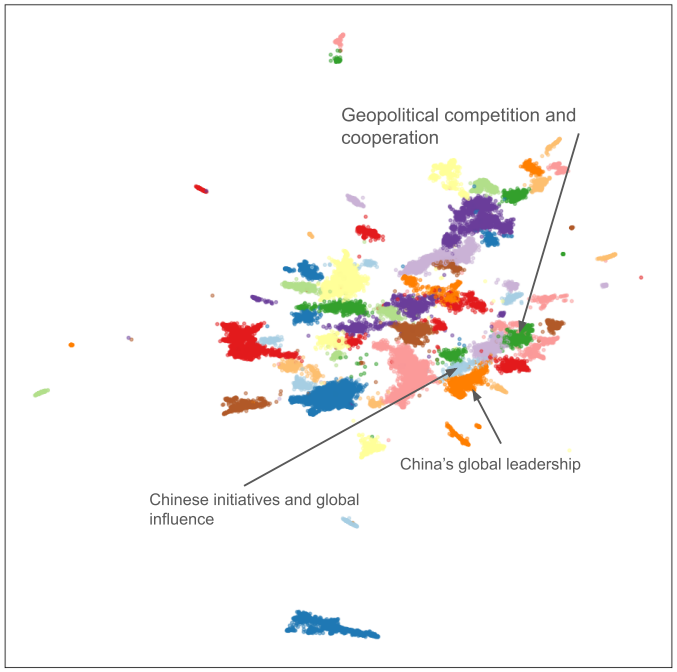

Figure 1: Clustering output visualized in two dimensions for the multi-platform dataset relating to general BRI terms.

not related to a particular real-world event. Several example general-BRI topics are the Aerial Silk Road, central Europe rail links, and BRI trade growth opportunities. Several BRI success-themed clusters were identified on the right side of the map, such as "Geopolitical competition and cooperation," "China's global leadership," and "Chinese initiatives and global influence." Texts within these clusters are all heavily biased in favor of the Chinese government and focus heavily on the success of the BRI, while using these successes to frame China as a well-positioned global leader. This is reasonable, since the Chinese government is well known to heavily censor and restrict topics on their platforms, as previously noted. This fact is reinforced by the model, due to the lack of China-critical clusters surfaced by the model. Several representative texts in these clusters include:

- TEXT: "党的理论经验 深刻认识和把握 "以史为鉴、开创未来" 的深厚意蕴… 部世界的关系。面对 "世界怎么了、应该怎么办" 的全人类发展困惑，以习近平同志为核心的党中央统筹国内国际两个大局、着眼人类发展未、提出共建 "一带一路" 倡、推动构建人类命运共同体、为建设美好世界贡献了中国智慧、中国方案。尤其是中国式现代化道路丰富了人类文明内涵，拓展了发展中国家走向"

  TRANSLATION: "The party's theoretical experience has a profound understanding and grasp of the profound implications of 'taking history as a mirror and creating the future'... the relationship between the world. Faced with the confusion of 'what's wrong with the world and what should be done about it?' in mankind's development, the Party Central Committee with Comrade Xi Jinping at its core coordinates both domestic and international situations, focuses on the future of human development, proposes the joint construction of the 'Belt and Road Initiative' to promote the construction of a community with a shared future for mankind and contributes Chinese wisdom and Chinese solutions to build a better world. In particular, the Chinese-style modernization path has enriched human civilization and expanded the direction of developing countries."

- TEXT: “policy REPORT “国际板” 呼之欲出？ 2022 进博会展望 政策展望 …” 概念得到全球认可 全球新共识 近年来，中国在国际不停的提出、呼吁、反复讲解 “人类命运共同体” 的国际发展新秩序、新模式，不仅讲，而且通过 “一带一路” 等扶持更多国家发展，参与全球的经济发展，共享人类发展的福利，受到大部分国家的欢迎，赢得了广泛的共识和人心。国内的环境更是不用说了”

  TRANSLATION: “policy REPORT ‘International Edition’ is about to come out? 2022 China International Import Expo Outlook Policy Outlook… ’The concept has gained global recognition and a new global consensus. In recent years, China has continuously proposed, called for, and repeatedly explained the new international development order and model of a ‘community with a shared future for mankind.’ In addition to talking about it, through the ‘Belt and Road Initiative’ and other initiatives, we have supported the development of more countries, participated in global economic development, and shared the benefits of human development. This has been welcomed by most countries and won widespread consensus and popular support. Needless to say, the domestic environment has [...]”

While the model did separate these topics of China-centric discussions of global cooperation and leadership, it was also able to distinguish their similarity, given that it placed these clusters near each other in the map. This is significant, as it indicates that the model is not only able to group similar posts into clusters, but also capable of grouping similar-content clusters, allowing for high-level topics to be identified. These smaller, similar clusters indicate that the model is capable of separating narratives with a high degree of granularity, rather than grouping them into a single “BRI success” cluster, for example. This is representative of the optimization over HDBSCAN parameters, ensuring that an optimal number of clusters is achieved. As we have demonstrated through detailed analysis of the texts within the clusters, and as evident by the presented labels in the Appendix, we have created a map of Chinese conversation around the BRI that focuses heavily on the Chinese government’s self-reported successes around the BRI, such as opportunities to create new trade partnerships, influence the global economy, and create green and sustainable development.

4.2 China–Pakistan Economic Corridor

While our method was capable of accurately capturing the large-scale conversations taking place, we were also interested in interrogating with greater granularity, the finer conversations that may overlap with specific real-world events. Thus, we turn to the map generated from the CPEC query shown in Figure 2; the labels are again in Table 9 (p. 23). The optimal parameters are shown in Table 1.

Our method identified fifty-two total clusters, with 4,877 clustered posts from WeChat and 105 posts from Weibo. Clusters included a “China’s Rise and US-China Competition” cluster, which projected China as an advocate for equal diplomatic relations and the US as a declining hegemony; a “China-Pakistan Friendship and Strategic Cooperation” cluster concerning China’s role in maintaining regional stability in Pakistan; and an “International Relations” cluster discussing how the United States and European countries are supposedly struggling with the Russia-Ukraine war, impacting their ability to serve as trade partners to Asian countries. Representative posts include:

- TEXT: “来源：杨风 面对俄乌战争，欧洲国家开始撑不住了。...太经济框架。即便是如此，越南仍然想要加强与中国的关系，尤其是在经济与贸易的合作。美国无法拉拢越南，更不用说其他的东南亚国家。除此之外，巴基斯坦总理夏巴兹 · 谢里夫、坦桑尼亚总统萨米娅 · 苏卢胡 · 哈桑，也都在这个星期访问中国。当然，重点是在星期五，德国总理朔尔茨访华。这将是从去年” and “点击上方蓝字 关注『国魂』...中美之间的差距将愈发明显。换句话来说，中国在过去数十年的对外政策获得了一定意义上的成功。毕竟相

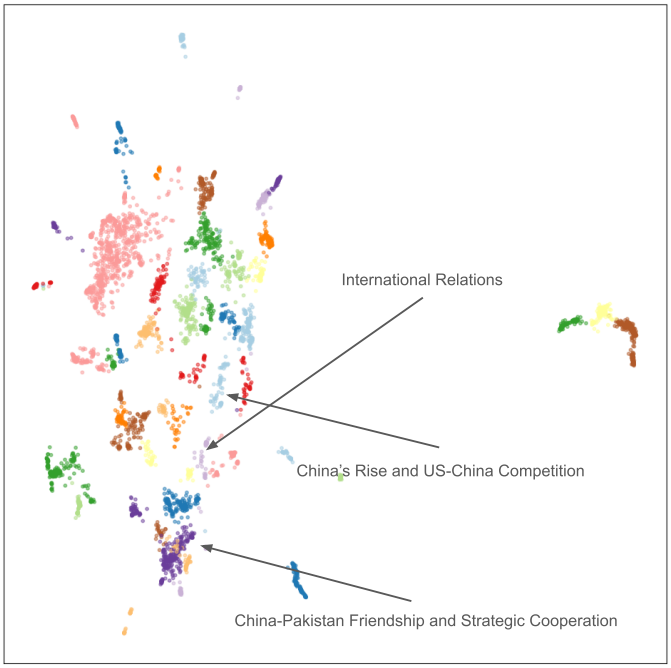

Figure 2: Clustering output visualized in two dimensions for the multi-platform dataset relating to CPEC terms.

对于美国动不动就是“十万吨”外交力量，中国“一带一路”的倡议显然要更有吸引力。中美两国到底谁会带来和平与发展，谁会带来战争和破坏，全球各国民众其实心里都非常清楚。倘若美国还继续保持着自己”

TRANSLATION: “Source: Yang Feng Facing the Russia-Ukraine war, European countries began to be unable to hold on. ...too economical framework. Even so, Vietnam still wants to strengthen relations with China, especially in economic and trade cooperation. The United States cannot win over Vietnam, let alone other Southeast Asian countries. In addition, Pakistani Prime Minister Shehbaz Sharif and Tanzanian President Samia Suluhu Hassan are also visiting China this week. Of course, the highlight is Friday, when German Chancellor Scholz visits China. This will mark from last year [...]” and “click the blue text above to follow ‘National Spirit’ ...The disparity between China and the U.S. will become even more apparent. In other words, China’s foreign policy over the past decades has achieved meaningful success. After all, compared to the United States’ frequent ‘100,000 ton’ diplomatic power, China’s ‘Belt and Road Initiative’ is more attractive. Citizens around the world know in their hearts who between China and the U.S. will bring peace and development, and who will bring war and destruction. If the U.S. continues to maintain[...]”

- TEXT: “丨来源：新华社 李克强同巴基斯坦总理夏巴兹举行会谈时强调 弘扬传统友谊 拓展全方位合作 ...策基石。巴方在涉及中方核心利益和重大关切的问题上始终坚定支持中方、衷心感谢中方为巴经济发展和灾后重建提供的宝贵帮助、愿同中方一道积极推进中巴经济走廊建设、加强农业、基础设施、清洁能源等领域合作、推动巴中关系迈上新台阶。夏巴兹强调、巴方不会容忍破坏巴中友谊的行径、将尽最大努力、采取”

  TRANSLATION: “丨 Source: Xinhua News Agency When Li Keqiang held talks with Pakistani Prime Minister Shehbaz, he emphasized that carrying forward traditional friendship and expanding all-round cooperation...policy cornerstone. Pakistan has

always firmly supported China on issues involving China's core interests and major concerns, and sincerely thanks China for its valuable help in Pakistan's economic development and post-disaster reconstruction. Pakistan is willing to work with China to actively promote the construction of the China–Pakistan Economic Corridor and strengthen agriculture, infrastructure, clean energy and other fields to promote Pakistan-China relations to a new level. Shehbaz emphasized that Pakistan will not tolerate actions that undermine the friendship between Pakistan and China and will do its best to take [...]"

Clusters in this map served as highly relevant indicators for the Chinese government's stated BRI visions, diplomatic goals, and general state interests. Several clusters suggested that conversations surrounding CPEC were frequently associated with political stability in the region. When promoting China's leadership roles, content in these clusters emphasized not only the economic benefits China can bring to certain countries and regions but also "softer" angles such as cultural and education exchange and China as a health-related public goods provider. This map also surfaced rarely-cited key terms including "six corridors and six channels serving multiple countries and ports," offering direction for future monitoring and analysis. Again, all clusters in the map were China-positive, reinforcing the China-centric bias on their country's internal social media platforms.

4.3 Southeast Asia Projects

The map for the Southeast Asia railway projects query is shown in Figure 3, and labels are shown in Table 10 (p. 24). The optimal parameters are shown in Table 1. Our method identified 87 clusters, with 2,875 clustered WeChat posts and 128 clustered Weibo posts. Most clusters were related to currently ongoing BRI projects in Southeast Asia including China–Thailand, China–Laos, and China–Indonesia railway connections. In addition, our method identified clusters containing content stating the Chinese government's ambitions for expanding infrastructure projects toward Regional Comprehensive Economic Partnership (RCEP) countries. Representative posts include:

- TEXT: "中国建材签署 阿尔及利亚埃利赞水泥粉磨站运维服务合同...。（国机工程集团公众号）中老铁路 老挝段货运站全部投入使用 11 月 1 日，老挝北部中老铁路孟赛站开办货运业务，至此中老铁路老挝段开通货运站达 7 个，标志着中老铁路货运站全部启用。中老铁路于 2021 年 12 月 3 日全线开通运营。截至 2022 年 10 月 31 日，中老铁路老挝段共运输货物 171."

  TRANSLATION: "China National Building Material signed an operation and maintenance service contract for the Relizane cement grinding station in Algeria .... (China National Machinery Industry Co. Official Account) All freight stations on the Laos section of the China–Laos Railway are put into use. On November 1, the Muang Xai Station of the China–Laos Railway in northern Laos started freight operations. So far, 7 freight stations have been opened on the Laos section of the China–Laos Railway, marking the opening of all China–Laos Railway freight stations. The China–Laos Railway became fully operational on December 3, 2021. As of October 31, 2022, the Laos section of the China–Laos Railway has transported a total of 171 [...]"

- TEXT: "中新社北京 10 月 18 日电 (记者 ...新冠疫情冲击，'一带一路' 合作非但没有按下'暂停键'，反而展现出强大韧性和活力。" 他说，中欧班列跑出逆风 "加速度"，中老铁路、克罗地亚佩列沙茨大桥建成通车，雅万高铁、匈塞铁路、中泰铁路等重点项目建设稳步推进。汪文斌强调，"一带一路" 超越地缘博弈的旧思维，开创了国际合作新范式。"

> TRANSLATION: “China News Service, Beijing, October 18 (Reporter...Despite the impact of the COVID-19 pandemic, ‘Belt and Road’ cooperation has not only not pressed the ‘pause button,’ but has shown strong resilience and vitality.” He said that China–Europe freight trains have overcome headwinds “to accelerate,” the China–Laos Railway and the Pelješac Bridge in Croatia were completed and opened to traffic, and the construction of key projects such as the Jakarta–Bandung High-speed Railway, the Hungary–Serbia Railway, and the China–Thailand Railway are steadily advancing. Wang Wenbin emphasized that ‘Belt and Road’ transcends the old thinking of geopolitical games and creates a new paradigm of international cooperation.”

While we identified fewer posts about China’s Southeast Asia railway projects in comparison to general BRI discourse, this case study provided valuable information about the Chinese government’s country-specific and regional goals, including Sino-Thai relations, RCEP, and various free trade agreements. This map also provided detailed trade statistics between China and partner countries through BRI projects, figures that can be difficult to obtain in English-language contexts. Positive rhetoric about the “China Standard” appeared in multiple clusters in this map, suggesting this may be a new way to frame the quality of China’s international cooperation in an attempt to market it to potential partners.

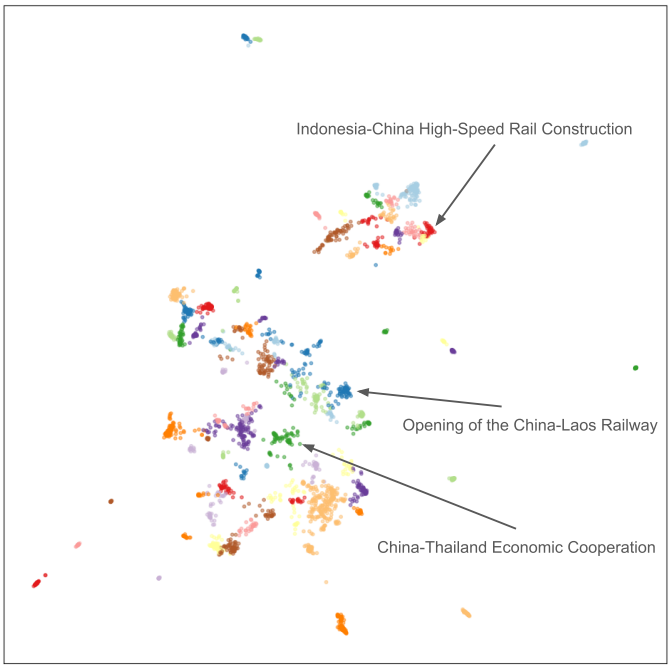

Figure 3: Clustering output visualized in two dimensions for the multi-platform dataset relating to Southeast Asia BRI terms.

4.4 Russia-Ukraine War

We also considered a use case relating to the Ukraine counteroffensive in the Russia-Ukraine war, with data in English, Russian, and Ukrainian languages. The intent behind selecting this use case is to demonstrate the ability of the method to capture multilingual content, as opposed to the strictly Chinese-language data we considered in the three BRI-related examples. The results of this use case are shown in Figure 4 (p. 13) and the labels in Table 11 (p. 25). We obtained 96 different clusters with clustered post counts

of 1,445 Facebook posts, 131 Instagram posts, 768 Reddit posts, 489 Telegram posts, 5,688 X posts, 1,089 VKontakte posts, and 730 YouTube posts. Clusters focused on Russian, Ukrainian, and Western takes on the Ukrainian counteroffensive. HDBSCAN parameters are shown in Table 1.

Due to the more balanced sizes of the platforms, we found that in the majority of cases no platform occupied more than 90% of posts for a single cluster. Conversation typically varied among platforms, with some bias depending on the platform. Qualitative review from our human analysts revealed that Instagram and X contained a roughly equal mix of Ukraine-leaning and Russia-leaning content; YouTube contained slightly more Ukraine-positive content; Facebook, Telegram, and VKontakte contained more Russia-positive content; and Reddit was mostly neutral. This is only generally the case, as at least 20% of the posts originating from any platform were included in a Russia-leaning cluster, Ukraine-leaning cluster, or non-leaning cluster.

Labels for several of the pro-Ukraine clusters include "Ukrainian counteroffensive gains momentum," "Russia's last reserves deployed in Ukraine," and "Ukraine's Breakthrough Near Zaporizhzhia." Texts within these clusters typically focus on framing the Ukrainian counteroffensive as a slow but steady success, discussing things like Russian troop losses and incoming military aid from Western countries. Pro-Russia cluster labels include "Biden Admin Failing in Ukraine," "Cost to Taxpayers of US Support to Ukraine," and "American Taxpayers Deserve Honesty About Tax Dollars Spent," which are all either Russian-language texts, or English-language texts that can be attributed to known pro-Kremlin actors, e.g., previously identified covert actors. Texts in these clusters tend to be highly inflammatory, and may have been intended to incite emotional reactions from readers, as evidenced by discussions of Ukraine aid costs to US taxpayers. The distinctions between these pro-Russia and pro-Ukraine topics is significant, as it shows that the model is able to accurately break a conversation space down into more granular topics, helping analysts understand the divisions and nuances of a highly polarized conversation. Combined with qualitative assessments of the returned clusters by human analysts, the model can serve as a powerful tool for surfacing potentially violative content produced and propagated by state and non-state actors seeking to influence online conversations.

In addition, the model identified one Russian-language cluster mocking the success of the Ukrainian counteroffensive, as well as another Ukrainian-language cluster stating that Russian claims of a counteroffensive failure actually indicate the operation's success, given the Russian government has previously spread false or misleading information. Respective posts include:

- "THIS! ; I'm having a hard time reporting on Ukraine, because I don't know how much harder we can beat this dead horse. Ukraine is in shambles. The holy counter-offensive failed catastrophically, and everything the West claimed about the war turned out to be a lie. Hundreds of thousands [...]"
- "Most likely alot of truth to this ; The more the Kremlin & Western vatniks say 'Ukraine's counteroffensive is stalling' and '#Ukraine should negotiate', the more you know the counteroffensive is making gains." [1]

This is particularly interesting, since both clusters could easily be considered as part of the same topic, yet they are treated as distinct clusters because they are differently-leaning. Similar results were noted in the BRI-related examples, and this is again confirmation that the optimization of HDBSCAN parameters is sufficient to produce clusters with fine points of interest, as opposed to large high-level clusters containing very broad

1. vatnik: Russian government-propaganda follower

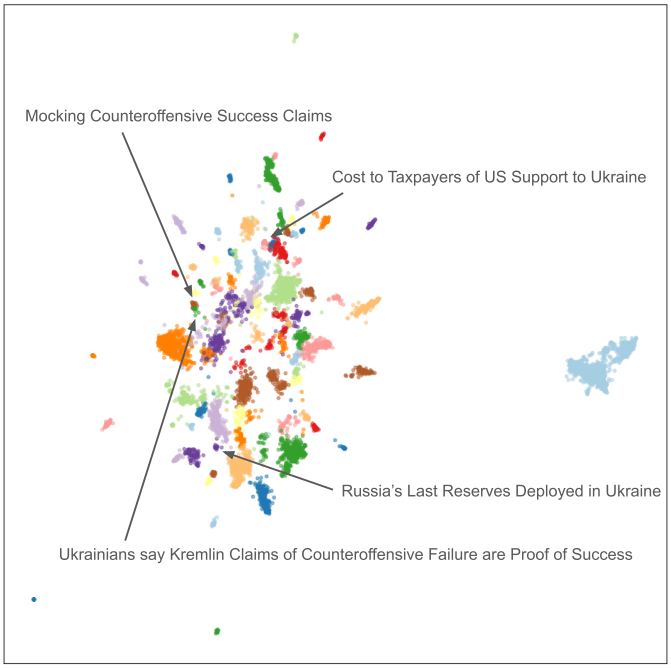

Figure 4: Clustering output visualized in two dimensions for the multi-platform dataset relating to the Ukrainian counteroffensive.

conversation topics.

Over the four separate use cases, we observed that the model was not only able to identify topics of social media discourse, but also capable of surfacing topics that could be considered misleading or harmful, such as the inability of Western governments to serve as successful Asian trade partners due to the Ukraine war, inaccuracies in claims of US military aid to Ukraine, and China's role as an economic partner for other Asian countries due to expanding Chinese infrastructure projects. This is arguably the most powerful capability of the method, as it serves in conjunction with human analysis to not only monitor and detect emergent online conversation, but also divide them into fine-grained narratives based on their viewpoints, allowing for an understanding of how small-scale topics are being discussed in the larger conversational picture across multiple platforms. A key aspect of our method is the fact that the input data is simply a list of users' posts and not dependent on their various follower or friend relationships, which allows narratives to form between groups of users that may not typically interact, e.g., on different platforms.

5 Conclusion

Starting from a list of posts across various social media platforms, we have demonstrated how this data can be processed and clustered to produce sets of coherent narratives across platforms that can be analyzed by human experts to glean valuable insights. The key aspect of our approach is that we look for similarities in posts embedded via a text embedding model, rather than look for posts shared by users with common connections. This allows our approach to be agnostic of the underlying user connection network, providing clusters with more complete narratives focused around differing viewpoints. This network independence also allows us to apply our method across datasets and

languages from multiple social media platforms, providing rich narratives with multi-platform viewpoints and allowing one to examine how content across different platforms can spread or vary.

We have explicitly applied our method to three Chinese-language Chinese-infrastructure-related use cases and one Russia-Ukraine war-related use case in English, Russian, and Ukrainian languages. We have demonstrated that the model is capable of surfacing fine-grained narratives in all cases, which are platform- and language-agnostic. This includes narratives that may merit additional review for signs of coordination or potential harms, such as incitement to violence. Analysis of these narratives can serve to provide insights into the types of conversations that take place in a social media conversation space, as well as how these conversations are responded to by platform users.

References

Acheampong, Francisca A., Henry Nunoo-Mensah, and Wenyu Chen. 2021. "Transformer models for text-based emotion detection: a review of BERT-based approaches." *Artificial Intelligence Review* 54 (February 8, 2021). https://doi.org/10.1007/s10462-021-09958-2.

Benigni, Matthew, Kenneth Joseph, and Kathleen Carley. 2017. "Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter." *PLoS ONE* (December 1, 2017). https://doi.org/10.1371/journal.pone.0181405.

Guo, Lei, and Yiyan Zhang. 2020. "Information Flow Within and Across Online Media Platforms: An Agenda-setting Analysis of Rumor Diffusion on News Websites, Weibo, and WeChat in China." *Journalism Studies* 21 (15): 2176–95. https://doi.org/10.1080/1461670X.2020.1827012.

Howard, Philip N., Bence Kollanyi, Samantha Bradshaw, and Lisa-Maria Neudert. 2017. "Social media, news and political information during the US election: Was polarizing content concentrated in swing states?" *Computational Propaganda Project,* https://ora.ox.ac.uk/objects/uuid:ae140d10-edc6-4b96-98b5-ab3347470882.

Jachim, Peter, Filipo Sharevski, and Emma Pieroni. 2020. "TrollHunter2020: Real-Time Detection of Trolling Narratives on Twitter During the 2020 US Elections." *CoRR* (December 4, 2020). arXiv: 2012.02606.

Kalyan, Katikapalli Subramanyam, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. "AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing." *CoRR* (August 12, 2021). arXiv: 2108.05542.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107 (2): 326–43. https://doi.org/10.1017/S0003055413000014.

Kumar Kaliyar, Rohit, Anurag Goswami, and Pratik Narang. 2021. "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach." *Multimedia Tools and Applications* (January 7, 2021). https://doi.org/10.1007/s11042-020-10183-2.

Lin, Tianyang, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. "A survey of transformers." *AI Open* 3:111–32. ISSN: 2666-6510. https://doi.org/10.1016/j.aiopen.2022.10.001.

McInnes, Leland, John Healy, and Steve Astels. 2017. "hdbscan: Hierarchical density based clustering." *Journal of Open Source Software* 2, no. 11 (February 26, 2017): 205. https://doi.org/10.21105/joss.00205.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," February 9, 2018. arXiv: 1802.03426 [stat.ML].

Meltwater. n.d. "Meltwater." Accessed December 8, 2022. https://www.meltwater.com/.

Mozafari, Marzieh, Reza Farahbakhsh, and Noël Crespi. 2019. "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media." *SCI* (February 26, 2019). https://doi.org/10.1007/978-3-030-36687-2_77.

OpenAI. 2022. "text-embedding-ada-002," December 15, 2022. https://openai.com/blog/new-and-improved-embedding-model.

OpenAI. n.d. “OpenAI.” Accessed June 1, 2023. https://openai.com/product.

Reimers, Nils, and Iryna Gurevych. 2019. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, November. https://doi.org/10.18653/v1/D19-1410.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2019. “Calls to Action on Social Media: Detection, Social Impact, and Censorship Potential.” In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda,* edited by Anna Feldman, Giovanni Da San Martino, Alberto Barrón-Cedeño, Chris Brew, Chris Leberknight, and Preslav Nakov, 36–44. Hong Kong, China: Association for Computational Linguistics, November. https://doi.org/10.18653/v1/D19-5005.

Sawhney, Ramit, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. “A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 7685–97. Online: Association for Computational Linguistics, November. https://doi.org/10.18653/v1/2020.emnlp-main.619.

Shahsavari, Shadi, Pavan Holur, Tianyi Wang, Tangherlini Timothy, and Roychowdhury Vwani. 2020. “Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news.” *Journal of Computational Social Science* (October 28, 2020). https://doi.org/10.1007/s42001-020-00086-5.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. “Fake News Detection on Social Media: A Data Mining Perspective.” *CoRR* (August 7, 2017). arXiv: 1708.01967.

Vayansky, Ike, and Sathish A.P. Kumar. 2020. “A review of topic modeling methods.” *Information Systems* 94:101582. ISSN: 0306-4379. https://doi.org/10.1016/j.is.2020.101582.

Xu, Qing, Ziyi Shen, Neal Shah, Raphael Cuomo, Mingxiang Cai, Matthew Brown, Jiawei Li, and Tim Mackey. 2020. “Characterizing Weibo Social Media Posts From Wuhan, China During the Early Stages of the COVID-19 Pandemic: Qualitative Content Analysis.” *JMIR Public Health Surveill.* (December 7, 2020). https://doi.org/10.2196/24125.

Zhang, Han, and Jennifer Pan. 2019. “CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media.” *Sociological Methodology* 49, no. 1 (July 19, 2019): 1–57. https://doi.org/10.1177/0081175019860244.

Authors

Arman Setser (arman.setser@graphika.com) is a machine learning researcher at Graphika.

Libby Lange (libby.lange@graphika.com) is an analyst at Graphika.

Kyle Weiss (kyle.weiss@graphika.com) is a senior analyst at Graphika.

Vladimir Barash (vlad.barash@graphika.com) is the chief scientist at Graphika.

Funding statement

This work was supported through a research grant from the US government.

Keywords

Content modeling; natural language processing; sentence embedding; social media data

Appendices

Appendix A: Data Queries

Here we present the traditional and simplified Chinese-language queries used to pull data from WeChat and Weibo with Meltwater. Although the English translations are also provided, these are for information purposes only. Only the Chinese-language queries were used to gather data. The queries for the general BRI projects, China–Pakistan economic corridor-related projects, and Southeast Asia projects are shown, respectively, in Tables 2, 3, and 4.

Table 2: Chinese-language queries and returned post counts for the general BRI projects dataset. English translations were not used for gathering data.

Query	Translation	Returned posts
一带一路	one belt one road	325,510
一帶一路	one belt one road	1,132
一带一路项目	one belt one road project	147
一帶一路項目	one belt one road project	0
一带一路 AND 绿色	one belt one road AND green	16,520
一帶一路 AND 綠色	one belt one road AND green	10
数字丝绸之路	digital silk road	780
數字絲綢之路	digital silk road	0
海上丝绸之路	maritime silk road	13,647
丝绸之路经济带	maritime silk road	11,164
海上絲綢之路	silk road economic belt	70
絲綢之路經濟帶	silk road economic belt	4
一带一路國際合作高峰论坛	one belt one road international cooperation summit forum	0
一带一路国际合作高峰论坛	one belt one road international cooperation summit forum	47

Table 3: Chinese-language queries and returned post counts for the China–Pakistan economic corridor-related dataset. English translations were not used for gathering data.

Query	Translation	Returned posts
中巴战略合作	China-Pakistan strategic cooperation	2
中巴战略伙伴	China-Pakistan strategic partnership	1
中巴全天候战略	China-Pakistan all-weather strategy	574
中巴经济走廊	China–Pakistan Economic Corridor	2,691
中国 AND 巴基斯坦 AND (基础设施投资 OR 基础建设 OR 基建)	China AND Pakistan AND (infrastructure investment OR infrastructure OR infrastructure)	133
一带一路 AND (巴基斯坦 OR 中巴 OR 巴铁)	One Belt, One Road AND (Pakistan OR China-Pakistan OR Pakistan Railway)	3,918

Table 4: Chinese-language queries and returned post counts for the Southeast Asia-related projects dataset. English translations were not used for gathering data.

Query	Translation	Returned posts
中老铁路	China–Laos railway	8,396
中老鐵路	China–Laos railway	3
磨万铁路	Mowan Railway	9
磨萬鐵路	Mowan Railway	0
雅万高铁	Jakarta–Bandung high speed railway	2,954
雅萬高鐵	Jakarta–Bandung high speed railway	1
中國-中南半島經濟走廊	China–Indochina economic corridor	0
中国-中南半岛经济走廊	China–Indochina economic corridor	5

Here we present the English, Russian, and Ukrainian queries for the dataset relating to the Russia-Ukraine war, respectively, in Tables 5, 6, and 7. The Russian and Ukrainian queries are in the same order as the English queries, for comparative purposes.

Table 5: English queries and returned post counts for the Russia-Ukraine war dataset.

Query	Returned posts
(counteroffensive OR counter-offensive) AND failure	14,303
(counteroffensive OR counter-offensive) AND failed	15,078
(counteroffensive OR counter-offensive) AND falling apart	444
(counteroffensive OR counter-offensive) AND success	15,823
(counteroffensive OR counter-offensive) AND victory	9,301
(counteroffensive OR counter-offensive) AND defeat	5,375
(counteroffensive OR counter-offensive) AND winning	3,497
(counteroffensive OR counter-offensive) AND losing	7,105
(counteroffensive OR counter-offensive) AND liberated	4,616
(counteroffensive OR counter-offensive) AND catastrophe	474
(counteroffensive OR counter-offensive) AND result	8,190
(counteroffensive OR counter-offensive) AND slaughtered	396
(counteroffensive OR counter-offensive) AND destroyed	9,483
(counteroffensive OR counter-offensive) AND strong	5,989
(counteroffensive OR counter-offensive) AND weak	5,564
(counteroffensive OR counter-offensive) AND endgame	647
(counteroffensive OR counter-offensive) AND loss	13,997
(counteroffensive OR counter-offensive) AND long war	970
(counteroffensive OR counter-offensive) AND prolonged	268
(counteroffensive OR counter-offensive) AND progress	26,942
(counteroffensive OR counter-offensive) AND patience	3,196
(counteroffensive OR counter-offensive) AND slog	257
(counteroffensive OR counter-offensive) AND cost	6,514
(counteroffensive OR counter-offensive) AND expense	595
(counteroffensive OR counter-offensive) AND imminent	511
(counteroffensive OR counter-offensive) AND gains	25,464
(counteroffensive OR counter-offensive) AND funding	1,841

Table 6: Russian queries and returned post counts for the Russia-Ukraine war dataset.

Query	Returned posts
(контрнаступление OR контр-наступление) AND провал	1,341
(контрнаступление OR контр-наступление) AND провалившийся	0
(контрнаступление OR контр-наступление) AND распад	59
(контрнаступление OR контр-наступление) AND успех	994
(контрнаступление OR контр-наступление) AND победа	331
(контрнаступление OR контр-наступление) AND поражение	454
(контрнаступление OR контр-наступление) AND побеждать	3
(контрнаступление OR контр-наступление) AND проиграть	5
(контрнаступление OR контр-наступление) AND освобожденный	0
(контрнаступление OR контр-наступление) AND успех	19
(контрнаступление OR контр-наступление) AND катастрофа	1,229
(контрнаступление OR контр-наступление) AND результат	13
(контрнаступление OR контр-наступление) AND убитый	3
(контрнаступление OR контр-наступление) AND уничтоженный	86
(контрнаступление OR контр-наступление) AND сильный	5
(контрнаступление OR контр-наступление) AND слабый	0
(контрнаступление OR контр-наступление) AND финальная стадия	726
(контрнаступление OR контр-наступление) AND потеря	0
(контрнаступление OR контр-наступление) AND долгая война	5
(контрнаступление OR контр-наступление) AND продолжительный	151
(контрнаступление OR контр-наступление) AND прогресс	57
(контрнаступление OR контр-наступление) AND терпение	250
(контрнаступление OR контр-наступление) AND марш	75
(контрнаступление OR контр-наступление) AND стоимость	230
(контрнаступление OR контр-наступление) AND расход	0
(контрнаступление OR контр-наступление) AND неминуемый	14
(контрнаступление OR контр-наступление) AND прирост	61
(контрнаступление OR контр-наступление) AND финансирование	10

Table 7: Ukrainian queries and returned post counts for the Russia-Ukraine war dataset.

Query	Returned posts
(контрнаступ OR контр-наступ) AND провал	2,453
(контрнаступ OR контр-наступ) AND провалився	233
(контрнаступ OR контр-наступ) AND розпад	2
(контрнаступ OR контр-наступ) AND успіх	178
(контрнаступ OR контр-наступ) AND перемога	201
(контрнаступ OR контр-наступ) AND поразка	6
(контрнаступ OR контр-наступ) AND перемагати	8
(контрнаступ OR контр-наступ) AND програти	1
(контрнаступ OR контр-наступ) AND визволений	0
(контрнаступ OR контр-наступ) AND успіх	61
(контрнаступ OR контр-наступ) AND катастрофа	2,236
(контрнаступ OR контр-наступ) AND результат	0
(контрнаступ OR контр-наступ) AND вбитий	2
(контрнаступ OR контр-наступ) AND знищений	8
(контрнаступ OR контр-наступ) AND сильний	12
(контрнаступ OR контр-наступ) AND слабкий	0
(контрнаступ OR контр-наступ) AND кінцева гра	2
(контрнаступ OR контр-наступ) AND втрата	1
(контрнаступ OR контр-наступ) AND довга війна	0
(контрнаступ OR контр-наступ) AND продовжений	465
(контрнаступ OR контр-наступ) AND прогрес	12
(контрнаступ OR контр-наступ) AND терпіння	421
(контрнаступ OR контр-наступ) AND марш	1
(контрнаступ OR контр-наступ) AND вартість	2
(контрнаступ OR контр-наступ) AND витрати	1
(контрнаступ OR контр-наступ) AND неминучий	0
(контрнаступ OR контр-наступ) AND приріст	16
(контрнаступ OR контр-наступ) AND фінансування	5

Appendix B: Cluster Labels

Here we present the complete set of cluster labels for all clusters in the four datasets we analyzed. Each of the following tables corresponds to one dataset.

Table 8: Cluster labels for the general BRI-related projects data.

International Anti-Corruption Cooperation	International cooperation in virtual reality
China's Communist Party and International Relations	习近平推动高质量发展的一带一路倡议
China's Global Influence	China's Global Connectivity
Emerging Cross-border E-commerce Opportunities	China-ASEAN Satellite Remote Sensing Cooperation
Opportunities in International E-commerce Market	Green and Low-carbon Development
Development of Intelligent Tax System	Insurance company growth and recruitment.
China's Global Development	China–Indonesia High-Speed Rail
Desert solar power project	Development of Chinese Pig Industry
Investment recommendations	Environmental Protection and Green Development
Knowledge Intellectual Property Cooperation	China's Steel Industry Expansion
Legal Development in China	Chinese Initiatives and Global Influence
Healthcare development and international collaboration	China's Green Development Initiatives
Student's Transformation through Skills Competition	Green and Low-carbon Development
International Education Institutions	Green development
Green and sustainable development	China's Belt and Road Initiative
Accelerating Chinese Export Growth	China's Global Leadership
Green and Low-carbon Development	China's Belt and Road Initiative
Green and Sustainable Development	Cross-border e-commerce
Chinese energy company with international presence and recognition.	Enhancing Maritime Trade Efficiency
Green and sustainable economic development.	Economic development and international cooperation
International Trade Events	China–Laos Railway: A Symbol of Belt and Road Initiative
China–Pakistan Economic Corridor	Significance of High-Speed Rail in National Development
China-Vietnam Tea Diplomacy	China's Openness and Economic Development
Hong Kong's Development and Integration	Regional Development
Chinese Airlines and the Belt and Road Initiative	New Era Achievements
Geopolitical competition and cooperation	Maritime Silk Road Cultural Heritage
China's Belt and Road Initiative	Artistic achievements and awards.
Building the Belt and Road Initiative	

Table 9: Cluster labels for the China–Pakistan economic corridor dataset.

Promoting Belt and Road Initiative	International Trade Exhibitions
Success of the Belt and Road Initiative	China–Pakistan Economic Corridor
China's Achievements and Contributions	Sino-Pakistani educational exchange
China's Global Influence	Cultural Exchange
China's Development and Cooperation	International expansion
China's Achievements in Infrastructure and Economic Development	China's regional diplomacy and challenges
China's Support for Pakistan	China-Brazil Relations
Greenhouse gas verification and sustainability services.	China–Pakistan Economic Corridor
China–Pakistan Economic Corridor Forum	China-Pakistan Strategic Partnership
China exports railway technology to Pakistan	China-Pakistan Friendship and Strategic Cooperation
China's Global Influence	China's Economic Support for Pakistan
China's Global Influence	中巴经济走廊合作
China's Economic Rise	China's Rise and US-China Competition
Port Investment in China	Orange Line Metro Project
RMB Settlement in Belt and Road Countries	China's Belt and Road Initiative
Macroeconomic trends in China	China–Pakistan Economic Corridor
North Star: Global Impact	China–Pakistan Economic Corridor
Climate Change Awareness and Action	China's Belt and Road Initiative
China's role in Afghanistan's economic reconstruction	Development of Southern Xinjiang Transportation Network
China-Nepal-India Geopolitical Relations	China's Belt and Road Initiative
International Relations	China's Belt and Road Initiative
China-Pakistan Flood Relief Cooperation	Global cooperation on the Belt and Road Initiative
China's Strategic Partnerships	China's Belt and Road Initiative
China-Pakistan Friendship and Cooperation	一带一路茶文化交流
China's Global Cooperation	China's 20th National Congress
Promoting the Belt and Road Initiative	China's Global Influence

Table 10: Cluster labels for the Southeast Asia projects dataset.

Chinese People's Daily news stories	China's Communist Party 20th National Congress
Chinese high-speed rail workers' dedication	Development of rural industries and infrastructure in China
Resilience	China-Latin America Relations
Language learning advertisement	China's Modern Logistics Achievements
China's Green Transformation and Global Low-carbon Support	China's Contribution to Belt and Road Initiative
China's Environmental Responsibility	China–Laos Railway
Fuzhou's Economic Development and Connectivity	China–Laos Railway
Railway schedule adjustment and improved transportation services	Chinese railway workers on the Kunming-Vientiane Railway.
王宁调研西双版纳州发展	Sino-Lao Cooperation
Recruitment for International Business Expansion	BIM Applications in Construction Projects
Chinese railway technology company	China-Latin America Relations
Carbon neutrality and green development	China-Latin America Relations
New Silk Road Economic Development	Importation of Laotian Beer to Chongqing
China's Global Leadership	Development of China–Laos Railway and its Impact on Trade and Connectivity
Investment and Cooperation in Southeast Asia	Development of China–Laos Railway
Construction Projects and Job Recruitment	Challenges and Achievements of Laos-China Railway
Current News and Developments	Development of Modern Logistics Infrastructure in Sichuan Province
Stock market updates	习近平外交思想的魅力和实践
China's Automotive and Infrastructure Development	China-Latin America Relations
Digital Innovation in Cultural and Tourism Services	China-Latin America Relations
China-South Asia Expo and Kunming Import and Export Fair	China–Laos Railway
China's Efforts for Development	Resilience of One Belt, One Road Initiative
Financial and Infrastructure Developments	Promoting Openness and International Cooperation
Current News and Events	Railway Freight Growth
Indonesia–China High-Speed Rail	Red River Delegates at Party Congress
Regional Economic Integration and Connectivity	Development and integration along the China–Laos Railway
China–Laos Railway	Launching the China–Laos Railway
Current Events	Development of China's foreign trade and transportation networks.
China-Indonesia High-Speed Rail Electrification Milestone	Dedication to the Development of the China–Laos Railway
Technological advancements and global developments.	Celebrating Chinese Railway Workers
China's High-Speed Rail Expansion	Opening of the China–Laos Railway
青岛资本市场迎来高铁出口历史性突破	Transformation through infrastructure development
Opening of the China–Laos Railway	Backpacking in Laos
China's First Exported High-Speed Train	Transportation development in Yunnan province
Inspection of Indonesian High-Speed Rail Project	Development of China–Europe Railway Connection in Guizhou
China's Future Development Direction	Transformation of Yunnan's Logistics Infrastructure
Completion of Indonesia's Yawan High-Speed Rail	China's Economic Achievements
Rapid Development of Chinese Railways	China's Belt and Road Initiative
China's Contribution to Global Development	China's Belt and Road Initiative
China-Thailand Economic Cooperation	Development of the China–Laos Railway
China–Indonesia High-Speed Rail Project	Development of the China–Laos Railway
Indonesia–China High-Speed Rail Construction	Connecting Lives: Chinese–Lao Railway
China's International Connectivity and Economic Growth	Friendship and Cooperation in Building the China–Laos Railway
China–Laos Railway	

Table 11: Cluster labels for the Russia-Ukraine war dataset.

Russian Discussion of Ukraine War	DOD Assessment of Ukraine Porogress and Ingenuity
Bias Against Elon	Mark Milley Assessment
US getting it's 'Money's Worth' in Ukraine While Ukranians Die	Mixed Assessments of Kherson Counteroffensive Success
Unnoticed Progress in Bakhmut	NYT and WSJ Articles on Counteroffensive Failure
Ukraine is Running Out of Time	Blaming Kiev for Failed Counteroffensive
US Sec Defense Good Meeting with Ukraine's New Defense Minister	White House Proclaims Counteroffensive Success
Criticism of Elon Musk's Success and Influence	Counteroffensive Debate
Ukraine's Strategic Successes	Failed Western Intervention in a One-sided Conflict
Anti-Kremlin Mockery	Russia's Last Reserves Deployed in Ukraine
Ukrainian President's Grim Outlook	Ukraine is Doomed
Ukraine Failing because Foreign Fighters are Abandoning	Counteroffensive Stalemate
Multicultural Business Collaboration	Preghozin Actions and Intentions
Ukrainian Diplomats Telling Critics to Shut Up	Is the Counteroffensive making progress?
Equipment losses in Zaporizhzhia counter-offensive	Shoigu Claims on Ukraine Failures
Claims of Stalled Counteroffensive are Pro-Russia Propaganda	Russian President Vladimir Putin's Economic and Political Agenda
Failed Wars and Lessons Unlearned	Putin Says Counteroffensive Failure
Ukranian Progress Miniscule on a Map	Ukraine's Long-Term Battle for Victory
Failure to Change Map	NATO's Ineffective Training of Ukrainian Soldiers
Debate over Ukraine Successes	Ukraine's Slow Counteroffensive Progress
Ukraine Showing Progress Under Extreme Conditions	Ukrainian Counteroffensive Gains Momentum
MSM Lies About Ukraine Exposed	How Ukraine Can Win a Long War
Glenn Greenwald Claiming West is Admitting Defeat	Possible Turning Point for Ukraine
Conspiracist BioeClandestine Claims Ukraine in Shambles	Zelensky's Failed Counteroffensive and Cabinet Shake-Up
Mocking Counteroffensive Success Claims	Biden Admin Failing in Ukraine
Ukrainian Forces Breaking Russian Defense	Cost to Taxpayers of US Support to Ukraine
Need to Find Even More Weapons for Ukraine	Challenges for Zelensky and Counteroffensive
Evaluating Ukraine's Counteroffensive	US Military Aid to Ukraine Updates
Kremlin Claims on Counteroffensive Failure are Proof of Success	American Taxpayers Deserve Honesty About Tax Dollars Spent
Jack Posobiec Claims Zelensky Washington Trip Huge Failure	Zelensky Facing Republican Dissent
Russian Running Disinformation Campaigns	Progress in Bakhmut Counteroffensive
Italian Assessments and Claims of Ukraine Corruption	Ukraine's Breakthrough Near Zaporizhzhia
Conflicting Claims of Zaporizhzhya Nuclear Power Plant Attack from Both Sides	West Believes Ukraine has Given Up Counteroffensive
Comparing Ukrainian Counteroffensive to Allied Invasion	West Must Prepare for Humiliation
Misc	Ukraine Proxy War: Disastrous Defeat for US/NATO
Kharkiv Counteroffensive Anniversary	MSM Propaganda has Misled the West
The Kharkiv Counteroffensive: Ukraine's Triumph	NATO Says Prepare for a Long War
Ukraine Success Necessary for European Security	Failure of War is Proof of NATO and Ukraine Corruption
Ukraine Counteroffensive Gains in Robotyne	Ukraine Gaining Ground in Counteroffensive
Secret Meetings to Reset Counteroffensive Strategy	Live Ukraine Battlefield Updates and Analysis
Analysis of Blinken Ukraine Visit	Destroyed Foreign Vehicles in Ukraine War
Criticism of Blinken Ukraine Support	Russian Destruction of British Challenger 2 Tank
Italy PM Claims NATO Failed in Ukraine	Ukraine's Leopard Tank Losses Exaggerated
Responses to Economist Article on Counteroffensive Challenges	Lack of Counteroffensive Media Shows Russia Winning
West Underestimated Russia	US Professor Claims Counteroffensive Failed
Criticism of McConnell's Ukraine Support	Counteroffensive Status is Failed
Limitations of F-16 Support for Ukrainian Counteroffensive	Putin Claims 71k Ukraine Soldiers Lost
US delivers Abrams tanks to Ukraine	Ukraine's Devastating Losses in Counteroffensive
US Cluster Munitions Supplied to Ukraine	US considering ATACMS missile aid to Ukraine