Leveraging Publicly Available Information to Analyze Information Operations

Nico Manzonelli[1], Taylor Brown[2], Antonio Avellandea-Ruiz[1], William Bagley[1], and Ian Kloo[1]

[1]United States Military Academy
Department of Systems Engineering
West Point, New York

[2]United States Military Academy
Dept. of Mathematical Sciences
West Point, New York

Corresponding author's Email: nico.manzonelli@ll.mit.edu

Author Note: We would like to thank Captain Iain Cruickshank, and Cadet Eda Hopkins, for their advice and direction in developing our methodology. Also, thank you to Lieutenant Colonel Natalie Casey for sponsoring our work at INDOPACOM.

Abstract: Traditionally, a significant part of assessing information operations (IO) relies on subject matter experts' time-intensive study of publicly available information (PAI). Now, with massive amounts PAI made available via the Internet, analysts are faced with the challenge of effectively leveraging massive quantities of PAI to draw meaningful conclusions. This paper presents an automated method for collecting and analyzing large amounts of PAI from China that could better inform assessments of IO campaigns. We implement a multi-model system that involves data acquisition via web scraping and analysis using natural language processing (NLP) techniques with a focus on topic modeling and sentiment analysis. After conducting a case study on China's current relationship with Taiwan and comparing the results to validated research by a subject matter expert, it is clear that our methodology is valuable for drawing general conclusions and pinpointing important dialogue over a massive amount of PAI.

Keywords: Information Operations, Publicly Available Information, Natural Language Processing, Web Scrapping

1. Introduction

As defined by Joint Publication 3-13, information operations (IO) are the integrated employment of information-related capabilities to influence, disrupt, corrupt, or usurp the decision-making of adversaries while protecting our own (Information Operations, 2012). Understanding IO campaigns and their effects informs high-level decision making and assists in combating adversaries' influence operations. Under the modern operating environment, assessing IO campaigns depends on manually evaluating PAI made available on the internet in the form of news reports and social media posts. This paper presents a method for aggregating and analyzing the vast amount of PAI in support of IO campaigns. First, we provide a brief background on IO, review existing techniques for analysis, and frame the capability gap that the proposed methodology aims to address. Next, we propose a methodology that leverages web scraping and natural language processing (NLP) techniques to acquire and process PAI in support of IO assessment. With the intent to support specific policy analysis and validate our methodology, we present a case study on Chinese, Taiwanese, and U.S. diplomacy and military relations. For classification and sensitivity reasons, our case study avoids directly assessing a known IO campaign in favor of addressing a broader regional focus. Finally, we conclude with a discussion on the results of the case study, the implications of the proposed methodology, and recommendations for future research in this domain.

2. Background

In 2009, U.S. Secretary of Defense James Mattis stated that capturing the perceptions of foreign audiences would replace seizing terrain in future joint operations (Rees, 2018). China's influence in Africa and Southeast Asia has exposed a new wave of IO on the global population. While U.S. IO campaigns are likely focused on trying to counteract Chinese messaging, there are limited ways to determine how friendly IOs affect adversaries' media and local populations.

DOI: https://doi.org/10.37266/ISER.2021v9i2.pp142-148

Traditionally, the Army's main tool for assessing friendly IO campaigns relies on subject matter experts' review of PAI, such as press releases, news reports, and social media (Army Publishing Directorate, 2018). The existing methods are time intensive, and the amount of information analysts can review is directly proportional to the number of personnel with the requisite expertise available. The proposed methodology intends to fill this capability gap by automating the existing process and applying cutting-edge text analytic techniques to process large quantities of data. While a close review of relevant media by a subject matter expert remains the "gold standard" for IO campaign analysis, the methodology presented in this paper improves on subject matter experts' capability to process PAI more efficiently.

2.1 Literature Review

The study team is aware of several classified efforts to conduct similar analyses of bulk data in support of US operations. While there are some approaches that are researched and disclosed for academic purposes, the majority of the work conducted across the information domain is classified (Holm, 2017). To avoid classification issues, we focused our literature review on methods in the NLP space.

NLP is an area of research and development that explores how computers process natural-language. Topic modeling, a particularly useful and dynamic subfield within NLP, describes a set of algorithms used to cluster together similar ideas throughout a large set of text data. At their core, topic modeling algorithms take matrix representation of the text data as inputs and cluster text according to document-term frequency (Barde & Bainwad, 2017). Among the many topic modeling approaches, Latent Dirichlet Allocation (LDA) is popular due to its propensity for high performance across many domains. In LDA, documents are represented as random mixtures of topics, and each topic is characterized by a distribution of words (Blei, Ng, & Jordan, 2003). Understanding LDA provides us with the knowledge required to implement and interpret topic modeling in our methodology.

Effectively evaluating IO campaigns requires analyzing text in a diverse set of languages. In the absence of trained linguists, machine translation (MT) is often employed in NLP analyses on foreign languages. There are three popular forms of MT: Rule-Based, Statistical, and Neural MT (NMT). Of the three, NMT uses neural networks to create the most modern and accurate MT models (Hutchins & Somers, 1992) (Gupta, Besacier, Dymetman, & Galle, 2019). Training our own NMT models is beyond the scope of this research effort, to be beyond the scope of this research effort, we can use commercial APIs for MT to incorporate translation into our analysis.

3. Methodology

As shown in Figure 1, the first step in our proposed methodology relies on subject matter experts to identify useful sources and topics of interest. Using guidance from experts, we develop web scrapers and perform the requisite data cleaning tasks to facilitate advanced analytics. Then, with properly cleaned data, we use topic modeling and sentiment analysis to extract insights from the data. We present our results to analysts and decision makers using several advanced visualization techniques with the goal of informing the evaluation of IO campaigns. Because the problem definition steps are specific to each research effort, we will discuss the process for target identification in Section 5 as it pertains to a case study. The remainder of this section will focus on the specific techniques used to collect and analyze PAI at a large scale.

Figure 1. Method Flow Diagram

3.1 Web Scraping

Developing potential information targets requires significant domain knowledge and an accurate problem definition for the intended use case. Analyzing state-sponsored media or press releases provides a state's perception of itself and reveals surface level influence objectives. PAI in the form of posts on microblogging websites (i.e. social media sites) can be used to understand how the civilian population reacts to state-sponsored media or foreign action. In either case, capturing both sources through web scraping is important for gauging a broad perception.

Accessing information on social media from a government setting creates challenges concerning authentication of Application Programming Interfaces (APIs). Our method for scraping social media platforms operates without APIs by implementing prebuilt scrapers backed by Selenium in order to maximize information gathered and maintain information security while decreasing cost (Muthukadan, 2018). The ethics and legality of web scraping is important to consider when implementing our methodology. In the civilian sector, decisions of web-scraping-related lawsuits demonstrate that it can certainly be conducted both legally and ethically (eBay, Inc. v. Bidder's Edge, Inc., 2000). Our methods were not developed or used to collect and store any information on U.S. persons.

3.2 Data Cleaning

The raw information collected during data acquisition cannot be analyzed or interpreted without extensive cleaning. At the structural level, summarizing JSON elements and translating html responses into a flat structure is required to save data as comma-separated-value files. Additionally, we filter out text content that is not the target languages, leaving the English text as is and preparing text in alternative languages for machine translation (MT). We evaluated various methods for translation and will discuss some MT alternatives in the case study. After translating the text data, we follow standard natural language processing procedures for text cleaning which include but are not limited to normalizing to lowercase, replacing contractions, and removing symbols, punctuation, over-used terms and whitespace.

3.3 Topic Modeling

We implement Latent Dirichlet Allocation (LDA) topic modeling across data gathered on news reports and social media to extract important and recurrent themes. On news reports, we found more traditional forms LDA using document-term matrices useful; however, typical LDA models do not work well on short-text collected from microblogging websites (Jones & Doane, 2019). To address LDAs short-comings for short-text documents, we implement a Gibbs Sampling Dirichlet Multinomial Model (GSDMM) designed specifically for short-text clustering (Yin & Wang, 2014). By applying different topic modeling approaches for long and short texts, we achieve better clustering on each subset.

3.4 Sentiment Analysis

Sentiment analysis is an NLP technique used to capture the overall attitude or emotion in a piece of text. We use sentiment analysis as the second piece in our analytic pipeline, applying it to the clusters identified by our LDA and GSDMM methods. First, we extract basic positive and negative sentiment for each document by implementing pre-trained sentiment models with the R package "sentimentr" (Rinker, 2019). In addition to basic positive-negative sentiment, we conduct sentiment analysis using the NRC Emotion Lexicon. The NRC Emotion Lexicon is a dictionary of English words and their related emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. (Mohammad & Turney, 2010). The annotations for NRC sentiment were completed manually via crowd sourcing and are well documented for research purposes (Mohammad & Turney, 2013). With NRC sentiment, we can better understand the emotions represented in each cluster.

By extracting the sentiment of each document in various clusters, we can make connections to the content of each document in a topic. Additionally, we can quickly sift through the data to find documents with outlying sentiment, or documents with average sentiment. Analyzing sentiment across each document in the clusters, allows us to locate specific dialogue and draw meaningful conclusions as it pertains to IO campaigns.

DOI: https://doi.org/10.37266/ISER.2021v9i2.pp142-148

4. China-Taiwan Case Study

4.1 Background

Dr. Oriana Skylar Mastro, a U.S. strategic planner with expertise on Chinese military and security policy, provided a testimony on Deterring PRC Aggression Toward Taiwan to the Congressional Review Commission presents an in-depth, human analysis of Chinese messaging and U.S. action (Mastro, 2021). Through extensive study, she answers questions like, "Do Chinese leaders believe they need to successfully land troops on Taiwan, and if so, to what extent does that deter them from initiating a military campaign?" (Mastro, 2021). Dr. Mastro concludes that despite the United States' best efforts in the region, the Chinese view themselves as a growing hegemony committed to an aggressive international policy (Mastro, 2021).

While Dr. Mastro draws important conclusions on Chinese perceptions and recommends corresponding U.S. action, the process involved in her study is extremely labor intensive. For the purposes of this case study, we apply our method to Chinese PAI to automate some of Dr. Mastro's research, using her established work to validate our findings. Specifically, we attempted to provide specific evidence of the Chinese perception that Taiwan's unification is imminent and that the Chinese have the capability to dominate the South China Sea despite U.S. intervention.

4.2 Implementation

For this case study, our sources come from Chinese news websites, and popular Chinese microblogging websites such as Twitter and Weibo. We developed target sources with direction from INDOPACOM command. Specifically, we targeted tweets sent from China and surrounding region with key terms like "China", "United States," "Taiwan", and "unification." Additionally, we targeted 10 Weibo users that primarily post about military and diplomacy. We also accessed the Chinese Ministry of Defense news website (eng.mod.gov.cn), and Global Times (globaltimes.cn) to retrieve state-sponsored media articles relevant to our study.

Using our social media scrapers we extracted 11,927 tweets and 10,009 Weibo posts dating back to November 2010. Our news web scrapers accessed 909 articles since October 2020. Data cleaning and text preprocessing followed a similar procedure for each data source. As outlined in section 3.2, we translated Chinese text to English using the Amazon Translate API (aws.amazon.com/translate).

With the data cleaned we applied the GDSMM model to the Twitter and Weibo data with the model parameters alpha = .01, beta = .15. The model parameters correspond to how the GDSMM model learns from the text data. Alpha indicates the probability of a document falling into an empty cluster, and beta indicates the prior probability that a document will seek an exact topic-match (Yin & Wang, 2014). Because GDSMM is designed for short-text clustering, we used traditional LDA to perform topic modeling on the news articles. Our LDA model clustered on 5 topics from the document-term-matrix. After creating each model, we applied sentiment analysis across clusters for each document. Implementing our method for the Taiwan-Chinese case study resulted in 3 separate models and their corresponding document sentiments.

4.3 Results and Discussion

Interpreting topic models requires acknowledgement that every document is a mixture of topics, and each topic is a mixture of words (Blei, Ng, & Jordan, 2003). This implies that words in various topics may belong to different documents, and document classification is based on the probability of that document appearing in a topic. Visualization tools allow us to gain a better understanding of our topics. Static visualizations, like word clouds, assist in interpreting LDA. Figure 2 displays the word clouds for the first 4 clusters in the news data. In addition to static visualizations, we use interactive visualizations backed by LDAvis to explore each cluster in detail (Sievert & Shirley, 2014).As seen in Figure 2, a distinct military-related cluster is prevalent throughout the news articles. We can dig into the sentiment of the articles that have a high probability of occurring within the military topic to gain a better understanding of their contents. Within the topic, we find many articles related to our case study like, "Taiwan's Display of new missile 'wrongly boosts courage of secessionists,'" "Taiwan islands intensive military exercises a political show to cover its weakness: analysts," and "PLA expels trespassing US Warships from Xisha Islands" (Xuanzun, 2021; Shumei & Lin, 2021; Xuanzun, 2020; Global Times, 2021).

Even by simply examining these articles' titles we can support three assertions on Chinese messaging: 1) China portrays confidence in their ability to exercise imperialist military action, 2) China considers Taiwan to be weak 3) China views US intervention as futile due to their military or diplomatic response. These assertions align with Dr. Mastro's argument that China is expanding its military capabilities and leaning toward Taiwanese reunification. While we found the messaging in Chinese state-sponsored media to be very direct, results from clustering social media are less clear.

DOI: https://doi.org/10.37266/ISER.2021v9i2.pp142-148

The microblogging data, clustered using GDSMM, indicates a wider variance in public opinion. Unlike state-sponsored media, where China can present a unified front on their stance, social media reveals the differing opinions on diplomacy between China, US, and Taiwan. The sentiment across microblogging political clusters is mixed with high emotions of trust and anticipation. Figure 3 includes a graph of the NRC emotion sentiment through the 3-largest clusters of twitter data. We discovered high levels of fear throughout political topics, but found that fear becomes more prevalent in military related topics. We assume that although China blocks most social media use, some citizens or visitors in the region can publish their uncensored thoughts onto social media. Allowing for oppressed ethnic groups to voice their opinion through social media, as evident by the military topic's fear proportion in Figure 3.

Overall, Dr. Mastro's findings validate our results. Chinese messaging is mostly aimed at creating confidence among their populous. China is committed to outwardly expressing their means to dominate the region through state-sponsored media. This case study was not intended to assess a specific IO campaign. Instead, we demonstrated that this methodology is capable of accurately detecting the same Chinese messaging that Dr. Mastro highlighted in her work. However, if we were analyzing a counter IO aimed at dispelling China's exuberant confidence and promoting resistance to aggressive government action, some evidence of this Io campaign shows through in the social media analysis.

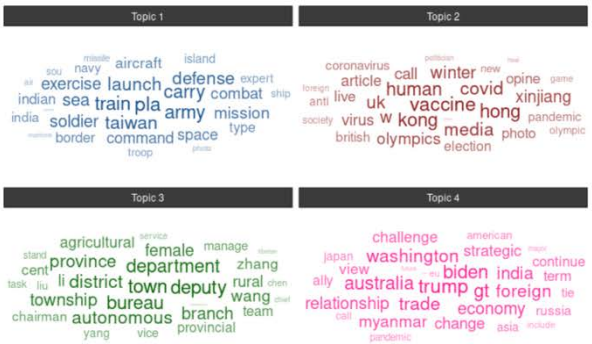

Figure 3. News Topics Word Clouds

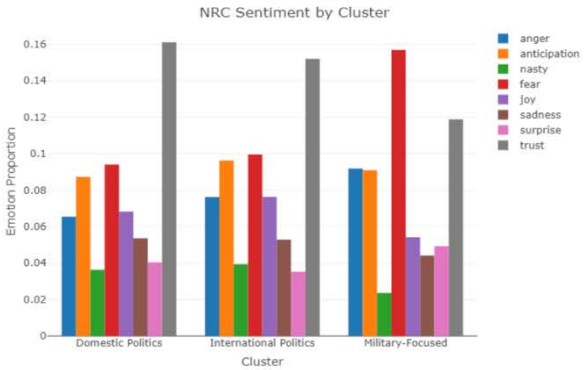

Figure 4. Twitter NRC Sentiment by Cluster

5. Limitations and Future Work

While the results from our case study on Taiwan are promising, some analysts may have difficulty applying our methodology to assess IO campaigns. Tweaking model parameters and data inconstancies create variance that makes recreating our work an intricate process that requires significant expertise in coding and applying NLP methods. It would be useful to wrap our web scrapers and analytic tools into a single web-application for ease of implementation. Even with a well-developed web application, there are many nuanced portions of data munging between sources that require careful consideration and are not suited to a one-size-fits-all approach. Additionally, in application, this method requires a subject matter expert's interpretation of the results using classified information to draw conclusions on IO campaigns. The methodology presented in this paper should be further developed and iterated upon before it is codified in an application (which creates obstacles for rapid analytic development). Therefore, our methodology is currently limited by general applicability and ease of access.

In addition to applicability, our method calls for implementing machine translation (MT) prior to analysis. Translating text from the source language to English drastically simplifies the data and, important information can be lost during MT. While we demonstrated that MT presents a viable method for analysis in the absence of linguists or Chinese-specific NLP tools, it is all but certain that applying analytics to the untranslated text would yield superior results. To address this limitation, significant research should be devoted to developing and operationalizing existing Chinese specific NLP tools. At the very least, training and implementing IO specific Chinese to English MT models would allow the practicing organization to achieve better translations without paying for access to a commercial translation API.

There are many NLP techniques beyond topic modeling and sentiment analysis that could be used to analyze information from adversaries that should also be explored. For example, researchers at MIT Lincoln Laboratory recently developed new ways to measure influence within a network (Smith, et al., 2021). Recognizing that there is a significant

DOI: https://doi.org/10.37266/ISER.2021v9i2.pp142-148

research effort behind graph theory and representing social media networks computationally, our methodology could benefit from implementing new methods of graph analytics to supplement our NLP findings and target future web scraping efforts.

Finally, any methodology that leverages PAI comes with policy and legal implications that should be considered by the implementing organization. While our methodology avoids accessing authenticated APIs (which is commonly problematic for government agencies), it would be worth considering some sort of managed attribution strategy to mask the interests and protect the tradecraft of government analysts.

6. Conclusion

The ability to accurate assess IO campaigns is a key aspect crafting successful IO campaigns and combating adversaries' influence operations. Traditional methods of evaluating IO campaigns require time-intensive study by subject matter experts. However, with massive amounts of PAI available on the internet, this task is becoming increasingly difficult. Our methodology aims to automate as much of the process as possible by using a multi-model system to analyze news articles and social media to assist in evaluating IO campaigns. Collecting PAI through web scraping proved to be an effective method to collect large quantities of data. Additionally, the implementation of NLP techniques, such as topic modeling and sentiment analysis, allowed us to cluster similar documents and make conclusions on the nature of their content. In operation, an analyst and IO subject matter expert should cooperate to apply our method to inform assessments of specific IO campaigns. By narrowing the focus and providing key insights on specific dialogue in PAI, this method has the potential to save time and effort while dramatically increasing analytic capability when evaluating IO campaigns.

6. Acknowledgments

This paper was previously published and presented in the Donald R. Keith Memorial Capstone Conference at USMA in May of 2021. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the United States Military Academy, United States Army, or United States Department of Defense.

7. References

Army Publishing Directorate. (2018, October 4). *The Conduct of Information Operations*. Retrieved from armypubs.army.mil/epubs/DR_pubs/DR_a/pdf/web/ARN13138_ATP%203-13x1%20FINAL%20Web%201.pdf

Barde, B., & Bainwad, A. (2017). An Overview of Topic Modeling and Tools. 2017 *International Conference on Intelligent Computing and Control Systems (ICICCS)*, (pp. 745-750). Madurai, India.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res*, 993–1022.

eBay, Inc. v. Bidder's Edge, Inc., C-99-21200RMW (US District Court for the Northern District of California May 24, 2000).

Global Times. (2021, February 5). *PLA expels trespassing US warship from Xisha Islands*. Retrieved from https://www.globaltimes.cn/page/202102/1215073.shtml

Gupta, R., Besacier, L., Dymetman, M., & Galle, M. (2019). Charecter-based NMT with Transformer. arXiv: 1911.04997.

Holm, R. R. (2017, March). *Natural Language Processing of Online Propoganda as a Means of Passivley Monitoring an Adversarial Ideology.* Retrieved from [Master's thesis, Naval Postgraduate School]: https://apps.dtic.mil/sti/pdfs/AD1045878.pdf

Hutchins, J. W., & Somers, H. L. (1992). *An Introduction to Machine Translation*. London: Academic Press.

Information Operations. (2012). In *Joint Publication* 3-13 (p. 87). Washington D.C.

Jones, T., & Doane, W. (2019). *textmineR*. Retrieved from https://www.rtextminer.com/

Mastro, O. S. (2021). The Precarious State of Cross-Strait Deterrence. *Statement before the U.S. China Economic and Security Review Commission on "Deterring PRC Aggression Toward Taiwan*."

Mohammad, S. M., & Turney, P. D. (2010). *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon*. Los Angeles: Association for Computational Linguistics.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a Word-Emotion Association Lexicon. 1308.6297.

Muthukadan, B. (2018). *Selenium with Python*. Retrieved from https://selenium-python.readthedocs.io/

Rees, B. (2018). Dismantling Contemporary Military Thinking and Reconstructing Patterns of Information: Thinking Deeper About Future War and Warfighting. *Small Wars Journal*, smallwarsjournal.com/jrnl/art/dismantling-contemporary-military-thinking-and-reconstructing-patterns-information.

Richardson, L. (2020). *Beautiful Soup Documentation*. Retrieved from https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Rinker, T. (2019). s*entimentr*. Retrieved from https://github.com/trinker/sentimentr

Shumei, L., & Lin, W. (2021, February 18). *Taiwan island's intensive military exercises a political show to cover its weakness: analysts.* Retrieved from Global Times: https://www.globaltimes.cn/page/202102/1215898.shtml

Sievert, C., & Shirley, K. (2014). *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces.* (pp. 63-70). Baltimore: Association for Computational Linguistics.

Smith, S. T., Kao, E. K., Mackin, E. D., Shah, D. C., Simek, O., & Rubin, D. B. (2021). Automatic detection of influential actors in disinformation networks. *National Academy of Sciences* (pp. 118-122). DOI: 10.1073/pnas.2011216118.

Xuanzun, L. (2020, December 22). *PLA expels US warship trespassing South China Sea*. Retrieved from Global Times: https://www.globaltimes.cn/page/202012/1210657.shtml

Xuanzun, L. (2021, January 27). *Taiwan's display of new missile 'wrongly boosts courage of secessionists'*. Retrieved from Global Times: https://www.globaltimes.cn/page/202101/1214177.shtml

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.*