A Framework towards Computational Narrative Analysis on Blogs

Kiran Kumar Bandeli, Muhammad Nihal Hussain, and Nitin Agarwal
Collaboratorium for Social Media and Online Behavioral Studies (COSMOS)
University of Arkansas at Little Rock,
Little Rock, Arkansas, USA
{kxbandeli, mnhussain, nxagarwal}@ualr.edu

Abstract

Social media is widely used to express views and share opinions with others. With the availability of inexpensive and ubiquitous mass communication tools like social media, creating narratives, false-information and propaganda is both convenient and effective. Social media users leverage this platform to further their views by framing narratives and participating in online discourse. Almost all events, issues, crises are discussed on social media. Blogs, unlike other social media platforms, are not regulated by any authority and have no restriction on character limit, providing bloggers with not only space for richer content but also serve as a platform for agenda-setting and content framing abetting development of narratives. This innate feature of blogs makes them a valuable platform for sociologists/political scientists to gain situational awareness by tracking different opinions, political views, and narratives as they are shaped. However, the deluge of posts in blogosphere makes it impractical to manually identify narratives. In this paper, we propose a novel framework to computationally identify narratives by extracting actors/actions using NLP techniques including POS tagging, chunking, and grammar rules to identify narratives. We also employ a scoring mechanism to rank them in the order of their dominance. Later, we evaluate the efficacy of the proposed model by validating against human annotated narratives. Our framework achieved an accuracy of 66.8%. Our proposed framework can help social scientists identify narratives computationally reducing human effort reasonably. Moreover, the results from this research could also be used to build effective counter narratives to stem propaganda campaigns.

1 Introduction

Social media has been widely used to study events/campaigns by many people around the world. Almost all events online that generate discourse have narratives to have radical effects or reinforce a perspective on social

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia (eds.): Proceedings of the Text2Story'20 Workshop, Lisbon, Portugal, 14-April-2020, published at http://ceur-ws.org

media. It can force change in perceptions of consumers on social media. Additionally, weaponized narratives can have a huge impact on social media discourse such as fake-news dissemination, and misinformation. To eliminate these radical effects on social media, it is important to track these narratives so that counter measures can be taken to quickly avoid damage to society. Social scientists define that narrative texts can be represented as directed graphs (networks) of semantic triples—subject-verb-object sets" [FR04].

Blogs, with their journal/diary like design, providing ample space for bloggers to write and express their opinions on events, as they develop, around the world, are conducive to capture the chronological sequence of events and help identify narratives. As narratives are described with actors and their actions by forming semantic triplets, using the same notion by using noun phrases to detect actors and verb phrases to identify their actions from blog posts. In the proposed framework, we aim to computationally identify actors/actions with the help of NLP techniques that include - POS tagging, chunking to obtain noun/verb phrases using grammar rules that captures appropriate phrases, and finally filter the phrases/sentences based on a scoring mechanism to generate narratives that describe actors and their actions. For a given blog post, our framework can identify dominant narrative as well as less dominant ones in the decreasing order of their rank. To evaluate the efficacy of the proposed model, the narratives generated computationally are evaluated against the human annotated narratives for the representative sample. The results from the representative sample indicate our framework achieved 66.8% accuracy.

Our proposed framework reduces social scientists' effort reasonably by identifying narratives computationally. This enables the analysts to learn what resonates with the community and if those interests and views are changing with time under the influence of exogenous factors or events. The results from this study could also be used to build counter narrative measures against propaganda campaigns.

2 Literature Review

Narratives are key features of any strategic communication [CRF12]. Therefore, narratives are used for persuasion because they are culturally grounded and frame the actions of actors. There have been several studies [CRF12], [BK16], [Rie05], [Rus16], [HBAkA18] that have looked at narratives based on different themes such as looking at text, frames, and semantic triples of subject-verb-object, named-entity extraction, targeted sentiment and linguistics. Authors in [CRF12], [BK16] apply a manual approach to study narratives and analyze how specific narratives rise and diffuse. Authors from [CRF12] also talk about strategies to undermine or counter the narrative argument. However, this approach lacks scalability.

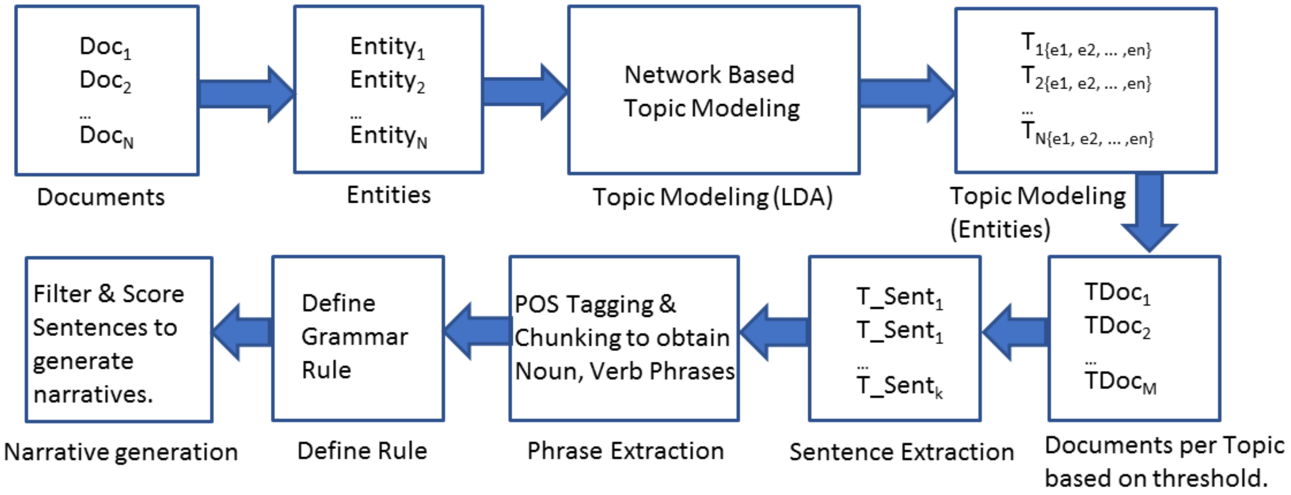
A study conducted by Alzahrani et al. [ACA$^+$16] defines story as actors taking actions that culminate in resolutions. Authors extract subject verb object relationships from paragraphs and generalize them into semantic conceptual representations. Further, the authors present an analytical framework that implements co-clustering to detect story forms. Ceran et al. [CKCD15], [CKM$^+$12] developed a novel algorithm to extract information from semantic triplets by clustering them into generalized concepts by utilizing syntactic criteria based on common contexts and semantic corpus-based based on "contextual synonyms". Furthermore, it proposes semantic features based on suitable aggregation and generalization of triplets that can be extracted using a parser. In a study by Joshua et al. [EF17], authors design a system that can automatically decide whether or not a paragraph of text contains a story. The authors say a paragraph contains a story if any portion of it expresses a significant part of a story, including the characters and events involved in major plot points. Study by Sudhahar et al. [SFC11] identifies narratives using a quantitative approach on news data. The methodology extracts semantic triplets using a parser called Minipar. Study by Hussain et al. [HBAkA18] focuses on identifying shift in narratives, rather than identifying narratives itself, using targeted sentiment analysis on blogs. Additionally, some of the studies [AB18], [BA18], [AB17], [HBN$^+$17], [MHN$^+$17] on blogs focus on several events and the role of blogs in disinformation campaign coordination, and narratives spreading during these campaigns.

Although several studies have been conducted on narratives across several domains, these are rarely computational and not scalable. Moreover, the barrage of information we see in social media and the pace at which these narratives are spread, analyzing these would require computational methods. Therefore, in this paper we introduce a preliminary framework to identify narratives computationally.

3 Research Methodology

We describe our proposed methodology as shown in Figure 1, where documents/blog posts are starting point in the framework. First, we extract named entities with the help of AlchemyAPI at the time of data collection (AlchemyAPI is deprecated now). Now, we input combination of blog posts and named entities to network topic modeling module. The output generated are topics with associated named entities. The number of topics in topic modeling module is decided based on parameters in the LDA (Latent Dirichlet Allocation) model and tuned to get optimum number of topics.Tuning refers to hyper-parameters in the LDA model which are log-likelihood and beta parameters. In LDA, both topic distributions, over documents and over words have correspondent priors, which are denoted typically with alpha and beta, and because these are the parameters of the prior distributions, are commonly referred to as hyperparameters. For our dataset, the log-likelihood is calculated for each iteration with topics as 1, 5, 10, 25, 50 etc. Ideally, the LDA model gets "better" at describing the data and increases over time. Eventually this value will level off, as subsequent iterations make negligible improvements to the model. On the other hand, low beta value places more weight on having each topic composed of only a few dominant words. Therefore, the combination of these hyperparameters helps in choosing number of topics. In our case the number of topics was 10.

Now, from the blog posts, we extract sentences using NLP. From the sentences, we make use of NLP techniques – POS tagging, chunking to extract noun phrases and verb phrases. After this step, we define grammar rule that captures patterns of noun and verb phrases to generate narratives. Finally, we rank the narratives with dominant narrative on the top and less dominant ones in descending order. Ranking/scoring of the narratives is performed based on TF-IDF method. By doing so, we know the dominance order of the narrative from the obtained candidates. We discuss different components of NLP techniques applied in the following paragraphs.



Doc_1
Doc_2
...
Doc_N
Documents
$Entity_1$
$Entity_2$
...
$Entity_N$
Entities
Network Based Topic Modeling
Topic Modeling (LDA)
$T_{1\{e1, e2, ... ,en\}}$
$T_{2\{e1, e2, ... ,en\}}$
...
$T_{N\{e1, e2, ... ,en\}}$
Topic Modeling (Entities)
$TDoc_1$
$TDoc_2$
...
$TDoc_M$
Documents per Topic based on threshold.
T_Sent_1
T_Sent_1
...
T_Sent_k
Sentence Extraction
POS Tagging & Chunking to obtain Noun, Verb Phrases
Phrase Extraction
Define Grammar Rule
Define Rule
Filter & Score Sentences to generate narratives.
Narrative generation

Figure 1: Framework to extract narratives from blogs.

3.1 Named Entity Extraction

Named Entity Extraction is widely used in Natural Language Processing to locate and classify named entity mentions in unstructured text into a set of categories such as person, organization, location etc. In this research, we used Alchemy API to extract named entities like persons, organizations, locations, etc. For instance, if there is a sentence – 'Adam lives in USA and goes to work at the University of Arkansas'. The API tags 'Adam' as person, 'USA' as location, and 'The University of Arkansas' as organization.

3.2 Network Topic Modeling

Topic modeling is the process of identifying topics in a corpus with a set of documents. This can be useful for search engines, customer service automation, and any other instance where knowing the topics of documents is important. There are multiple ways in which we can achieve this, we use Latent Dirichlet Allocation (LDA) in our framework. In LDA, each document may be viewed as a mixture of topics and each topic is a mixture of words. In our framework, we use network-based topic models which internally uses LDA and generate a network with topics and named entities associated with each topic. The documents in each topic are further filtered based on topic percentage threshold so that we capture documents belonging to the topic.

3.3 Sentence Extraction

From NLTK (Natural Language Toolkit), we used the method - sent_tokenize to extract sentences.

3.4 POS Tagging

POS tagging refers to parts of speech and identifies how a word is used in a sentence.

3.5 Chunking

Chunking is a process of extracting phrases from unstructured text. Instead of just simple tokens which may not represent the actual meaning of the text, it's desirable to use phrases such as "United States" as a single word instead of 'United' and 'States' separate words. Like POS tags, there are a standard set of Chunk tags like Noun Phrase (NP), Verb Phrase (VP), etc. Chunking is important to extract information from text such as Locations, Person Names etc. The same process in NLP referred to as Named Entity Extraction. We will consider Noun Phrase Chunking and searching for chunks relating to an individual noun phrase. In order to create NP chunk, we define the chunk grammar using POS tags. We will define this using a single regular expression rule. The rule states that whenever the chunk finds an optional determiner (DT) followed by any number of adjectives (JJ) and then a noun (NN) then the Noun Phrase (NP) chunk should be formed. The following example in Figure 2 explains tree structure of a sentence with NP phrases. Similarly, we can extract Verb Phrase (VB) by defining chunk to find verb form (VBD) followed by any preposition (IN).

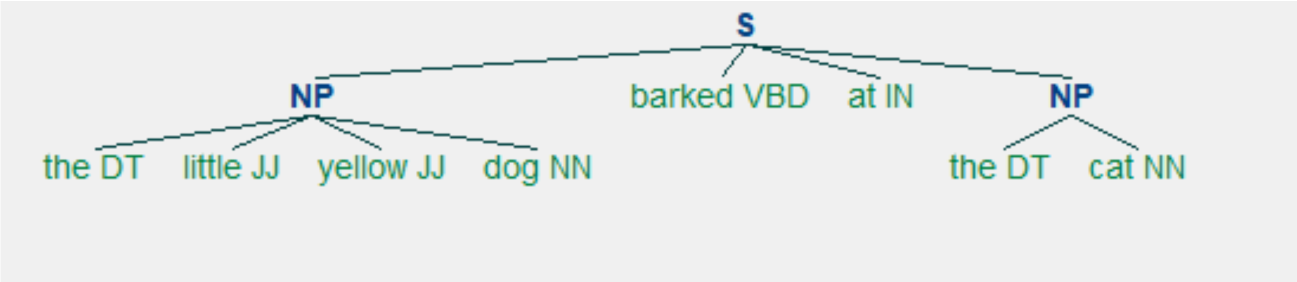
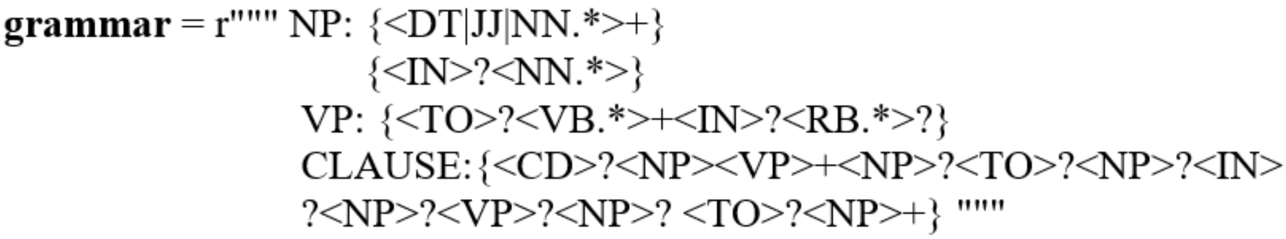


Figure 2: Chunking example in NLP

3.6 Grammar Rules

Grammar rules are the regular expressions written to capture actors and their actions. Based on POS tagging, we extract noun phrases and verb phrases. By defining grammar rules, we extract chunks of sentences to identify candidates of narratives for a given text. For our dataset, we followed below grammar rule in Figure 3.



```
grammar = r""" NP: {<DT|JJ|NN.*>+}
                   {<IN>?<NN.*>}
               VP: {<TO>?<VB.*>+<IN>?<RB.*>?}
               CLAUSE:{<CD>?<NP><VP>+<NP>?<TO>?<NP>?<IN>
               ?<NP>?<VP>?<NP>? <TO>?<NP>+} """
```

Figure 3: Grammar rule extracting narratives from blogpost

3.7 Scoring Sentences

While sentences can be scored using various algorithms, in this study, we use TF-IDF score of words in a sentence to assign a weight to the narratives. Later, when the narratives are generated, this scoring method is used to rank dominant to less dominant narratives. To briefly explain TF-IDF stands for Term Frequency-Inverse Document Frequency. TF-IDF weight is commonly used in information retrieval and text mining fields.

Table 1: Narratives from the sample blog posts.

Breitbart london was the first news site in the english speaking world to report on the horrific.
German police have admitted to losing several urban areas to migrant gangs.
The cologne city council to submit a travel warning for the carnival time.
Breitbart london is still waiting on comment from cologne police.
The deteriorating state of control the local government has over the city.
An opposition council member has today sounded the alarm bell.
A priest described munich on new years eve as a warzone.
Local police to hold a press conference on monday afternoon.
A political scandal is developing in germany as ordinary citizens wake up to the scale.
Cologne to mass sex crime on new years.

3.8 Narrative Generation

By following all the above steps in our framework from Fig. 1, we generate narratives computationally for any given blogpost. Depending on the grammar rules defined, there might be some noise in the sentences. To overcome this, we can include more rules and filter the narratives obtained. In this paper, our goal is to identify narratives computationally. First, we use the EU migrant crisis dataset collected using methodology described in [HOB+17], [KBWA19] and extend our previous work [HBAkA18]. We selected 30 blog posts that were published in January 2016 and related to the events around the New Year's Eve incident in Cologne. From these 30 blog posts, we extracted more than 3,000 sentences and filter down to capture semantic triplets having actors and their actions. Using NLP techniques, we obtain candidate narratives. Below Table 1 shows a list of narratives identified computationally using our methodology.

From the narratives identified, we can notice that the blog posts discuss about migrants and reactions by the local police in Cologne. Further, narratives show Munich being described as a warzone. Finally, discuss how political scandal is developing in Germany due to attacks of scale. Overall, we see narratives strongly talking against migrants and the consequences of allowing migrants in Europe.

To validate our approach, we asked 3 human analystss to list at least 5 narratives in the descending order of their dominance (i.e., the dominant narrative will be listed in the top and then the less-dominant one will be listed later) from the 30 posts. Later, we provided each of the 3 analystss with narrative identified from our framework to compare against human identified narratives. This effort resulted in 70% match with analysts 1, 64% match with analysts 2, and 66.6% match with analysts 3. To compute the inter-annotator agreement, we used krippendorff-alpha and obtained krippendorff-alpha of 0.82 indicating almost perfect reliability. **Note:** Krippendorff's alpha is a reliability coefficient to measure the agreement among analystss when coding a set of units of analysis in terms of the values of a variable. In our research, three analystss classifying narratives achieved $0.74 < \text{alpha} < 0.86$ with average alpha = 0.82

Table 2: Average Precision, Recall and F measure of identified narratives.

Top K Narrative	Avg. Precision	Avg. Recall	F Measure
1	0.70	0.14	0.23
2	0.69	0.28	0.39
3	0.70	0.42	0.53
4	0.69	0.55	0.61
5	0.67	0.67	0.67

Lastly, we compute precision, recall and F-measure (accuracy) to validate our methodology and select optimal Top K number of narratives. Here, K refers to number of narratives. When say K =1, it means we consider only top 1 narrative from the blog posts. Similarly, K = 5 refers to top 5 narratives from the blog posts. Overall, our framework generated more than five narratives per post. From the analystss' side, we had two analystss who gave exactly 5 narratives per post and third analysts gave more than 5 narratives per post. For calculations, we considered top 5 narratives from analystss as well as our framework. Now, based on this information, we calculated precision, recall and F measure between our framework output and each one of the analystss. Here, precision is the fraction of retrieved narratives that are relevant narrative from the analysts. Recall is the fraction of the relevant narratives that are successfully retrieved. F measure is the harmonic mean of precision and recall.

Overall, our framework achieved a maximum accuracy of 67% as shown in Table 2 at K = 5.

4 Challenges & Limitations

Although our framework computationally identifies narratives, it has several challenges and limitations. The grammar rule used is generically written to cover most of the phrase patterns of nouns and verbs. If the sentence is complex and there are different noun and verb phrase patterns, the grammar rule needs to be updated. For this reason, chunking becomes the crucial part in our framework, and we need to trial and error with grammar rules to obtain meaningful candidates of narratives. For validation, we relied on three analystss, more analystss should be included. Also, there is always subjective bias in this type of research which needs to be eliminated. For this reason, the proposed framework can be tested against several accepted datasets in the research. The candidate narratives generated may not be fully accurate but can assist analyst to understand. There will always be some noise generated by the system.

5 Conclusion and Future Work

With social media affecting almost every aspect of our life, it is important, now more than ever, to study social behavior and online campaigns. There is a critical need for models, tools and frameworks that can assist conduct these studies computationally. In this paper, we provide a framework to identify narratives on blogs. And using the EU migrant crisis dataset, we demonstrated its efficacy and robustness in identifying narratives. The framework achieved an accuracy of 66.8%. Moreover, the scoring feature embedded within the framework also helps in identifying dominant and non-dominant narratives. Although the framework's accuracy is not 100%, it provides the analyst with reasonable candidates to help gain insights into what resonates in the blogosphere.

Furthermore, the framework's accuracy can be increased using several improvements including the addition of better rules for chunking. Another future direction could be to explore abstractive summarization to improve the overall framework. The framework can also be applied to other social media platforms or even news sources. Another future research direction could be to study evolution of different narratives, i.e., does it become fringe or master narrative.

Acknowledgements

This research is funded in part by the U.S. National Science Foundation (OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2605, N00014-17-1-2675, N00014-19-1-2336), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, and the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

[AB17] Nitin Agarwal and Kiran Kumar Bandeli. Blogs, Fake News, and Information Activities. In *Digital Hydra: Security Implications of False Information Online*, pages 31–45. NATO Strategic Communications Center of Excellence (StratCom COE), 2017.

[AB18] Nitin Agarwal and Kiran Kumar Bandeli. Examining strategic integration of social media platforms in disinformation campaign coordination. *Stratcom Centre of Excellence (CoE), Riga, Latvia*, 2018.

[ACA^{+}16] Sultan Alzahrani, Betul Ceran, Saud Alashri, Scott W. Ruston, Steven R. Corman, and Hasan Davulcu. Story forms detection in text through concept-based co-clustering. In *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom)*, pages 258–265. IEEE, 2016.

[BA18] Kiran Kumar Bandeli and Nitin Agarwal. Analyzing the role of media orchestration in conducting disinformation campaigns on blogs. *Computational and Mathematical Organization Theory*, pages 1–27, 2018.

[BK16] Jan Babinský and Václav Krejčíř. Representation of Ukrainian Crisis in Czech Media: Explicit and Implicit Bias in the News Coverage of the Ukranian-Russian Conflict. *Mediterranean Journal of Social Sciences*, 7(4):435, 2016.

[CKCD15] Betul Ceran, Nitesh Kedia, Steven R. Corman, and Hasan Davulcu. Story detection using generalized concepts and relations. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 942–949. IEEE, 2015.

[CKM$^+$12] Betul Ceran, Ravi Karad, Ajay Mandvekar, Steven R. Corman, and Hasan Davulcu. A semantic triplet based story classifier. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 573–580. IEEE Computer Society, 2012.

[CRF12] S. Corman, Scott W. Ruston, and Megan Fisk. A pragmatic framework for studying extremists' use of cultural narrative. In *2nd International Conference on Cross-Cultural Decision Making: Focus*, volume 2012, pages 21–25, 2012.

[EF17] Joshua Eisenberg and Mark Finlayson. A simpler and more generalizable story detector using verb and character features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2708–2715, 2017.

[FR04] Roberto Franzosi and Franzosi Roberto. *From words to numbers: Narrative, data, and social science*, volume 22. Cambridge University Press, 2004.

[HBAkA18] Muhammad Nihal Hussain, Kiran Kumar Bandeli, Samer Al-khateeb, and Nitin Agarwal. Analyzing Shift in Narratives Regarding Migrants in Europe via Blogosphere. In *Text2Story@ ECIR*, pages 33–40, 2018.

[HBN$^+$17] M Hussain, KK Bandeli, Mohammad Nooman, Samer Al-khateeb, and Nitin Agarwal. Analyzing the voices during european migrant crisis in blogosphere. In *The 2nd International Workshop on Event Analytics using Social Media Data*, 2017.

[HOB$^+$17] Muhammad Nihal Hussain, Adewale Obadimu, Kiran Kumar Bandeli, Mohammad Nooman, Samer Al-khateeb, and Nitin Agarwal. A framework for blog data collection: challenges and opportunities. In *The IARIA international symposium on designing, validating, and using datasets (DATASETS 2017)*, 2017.

[KBWA19] Tuja Khaund, Kiran Kumar Bandeli, Oluwaseun Walter, and Nitin Agarwal. A Novel Methodology to Identify and Collect Data from Relevant Blogs Leveraging Multiple Social Media Platforms and Cyber Forensics. *ALLDATA 2019*, page 49, 2019.

[MHN$^+$17] Esther Ledelle Mead, Muhammad Nihal Hussain, Mohammad Nooman, Samer Al-khateeb, and Nitin Agarwal. Assessing situation awareness through blogosphere: a case study on venezuelan socio-political crisis and the migrant influx. In *The Seventh International Conference on Social Media Technologies, Communication, and Informatics (SOTICS 2017)*, pages 22–29, 2017.

[Rie05] C. K. Riessman. *Narrative analysis: University of Huddersfield.* 2005.

[Rus16] Scott W. Ruston. More than Just a Story: Narrative Insights into Comprehension, Ideology, and Decision Making. In *Modeling Sociocultural Influences on Decision Making*, pages 57–72. CRC Press, 2016.

[SFC11] Saatviga Sudhahar, Roberto Franzosi, and Nello Cristianini. Automating quantitative narrative analysis of news data. In *Proceedings of the Second Workshop on Applications of Pattern Analysis*, pages 63–71, 2011.