Using Functional Schemas to Understand Social Media Narratives

Xinru Yan **Aakanksha Naik** **Yohan Jo** **Carolyn Rosé**
Language Technologies Institute
Carnegie Mellon University
{xinruyan, anaik, yohanj, cp3a}@cs.cmu.edu

Abstract

We propose a novel take on understanding narratives in social media, focusing on learning "functional story schemas", which consist of sets of stereotypical functional structures. We develop an unsupervised pipeline to extract schemas and apply our method to Reddit posts to detect schematic structures that are characteristic of different subreddits. We validate our schemas through human interpretation and evaluate their utility via a text classification task. Our experiments show that extracted schemas capture distinctive structural patterns in different subreddits, improving classification performance of several models by 2.4% on average. We also observe that these schemas serve as lenses that reveal community norms.

1 Introduction

Narrative understanding has long been considered a central, yet challenging task in natural language understanding (Winograd, 1972). Recent advances in NLP have revived interest in this area, especially the task of story understanding (Mostafazadeh et al., 2016a). Most computational work has focused on extracting structured story representations (often called "schemas") from literary novels, folktales, movie plots or news articles (Chambers and Jurafsky, 2009; Finlayson, 2012; Chaturvedi et al., 2018). In our work, we shift the focus to understanding the structure of stories from a different data source: narratives found on social media. Table 1 provides an example story from the popular online discussion forum Reddit [1]. Prior work has studied stories of personal experiences found on social media, identifying new storytelling patterns. However, these studies have focused on how storyteller identity is conveyed (Page, 2013). In our work, we instead aim to understand novel structural patterns exhibited by such stories.

i was eating breakfast this morning while my stepfather was making his lunch to take to work. as he reached for the plastic wrap for his sandwich i subtly mentioned that he could use a reusable container. so he walked over to the container drawer and used a container realizing that it was the perfect size. i know its not much but hopefully he remembers this tomorrow when making his lunch...

Table 1: Sample personal story from Reddit

Computational work in story understanding often attempts to construct structured representations revolving around specific narrative elements. Broadly, these approaches can be divided into two classes: *event-centric* techniques (Chambers and Jurafsky, 2008) and *character-centric* techniques (Bamman, 2015). We adopt a novel take that focuses instead on extracting the "functional structure" of stories. For example, a common story can have a functional structure consisting of phases such as character introduction, conflict setup and resolution. To represent such structure, we propose the paradigm of **functional story schemas**, which consist of stereotypical sets of functional structures. A major difference between our conceptualization of functional story schemas and prior approaches is the focus on high-level narrative structure, which reduces domain-specificity in the found schemas. Studies have shown that functional narrative structures are critical in forming stories and play an important role in story understanding (Brewer and Lichtenstein, 1980, 1982).

We develop a novel unsupervised pipeline to extract functional schemas (§3), which consists of two stages: *functional structure identification* and *structure grouping for schema formation*. The first stage uses the Content word filtering and

[1] https://www.reddit.com/

Proceedings of the Second Storytelling Workshop, pages 22–33
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

Speaker preferences Model (CSM), a generative model originally applied to detect schematic progressions of speech-acts in conversations (Jo et al., 2017), while the second stage groups strongly co-occurring sets of structures using principal component analysis (PCA) (Jolliffe, 2011). To validate extracted schemas, we perform a two-phase evaluation: manual interpretation of schemas (§4.2) and automated evaluation in a downstream text classification task (§4.3).

Utilizing our pipeline to extract functional schemas from posts on three subreddits discussing environmental issues [2], namely */r/environment*, */r/ZeroWaste* and */r/Green*, we observe that our schema interpretations reflect typical posting strategies employed by users in each of these subreddits. Incorporating schema information into the feature space also boosts the performance of a variety of baseline text classification models on subreddit prediction by 2.4% on average. After validation, we use extracted schemas to gain further insight into how stories function in social media (§5). We discover that functional schemas reveal community norms, since they capture dominant and unique posting styles followed by users of each subreddit. We hope that our conceptualization of functional story schemas provides an interesting research direction for future work on story understanding, especially stories on social media.

2 Background & Related Work

2.1 Narrative Understanding

Much prior work on narrative understanding has focused on extracting structured knowledge representations ("templates" or "schemas") from narratives. These works can be divided into two major classes based on the narrative aspect they attend to: *event-centric* and *character-centric*.

Event-centric approaches primarily focus on learning "scripts", which are stereotypical sequences of events occurring in the narrative along with their participants (Schank and Abelson, 1977). While scripts were introduced in the 1970s, not much early work (with the exception of Mooney and DeJong (1985)) attempted to build models for this task due to its complexity. However, it has garnered more interest in recent years. Chambers and Jurafsky (2008) modeled scripts as narrative event chains, defined as partially ordered sets of events related to a *single* common actor, and built an evaluation called the *narrative cloze* test aimed at predicting a missing event in the script given all other events. Chambers and Jurafsky (2009) broadened the scope of event chains by defining "narrative schemas" which model all actors involved in a set of events along with their *role*. These inspired several script learning approaches (Regneri et al., 2010; Balasubramanian et al., 2013). A related line of research focused on extracting "event schemas", which store *semantic roles* for typical entities involved in an event. Several works proposed unsupervised methods for this task (Chambers and Jurafsky, 2011; Cheung et al., 2013; Chambers, 2013; Nguyen et al., 2015). Recent research identified a key problem with the narrative cloze test, namely that language modeling approaches perform well without learning about events (Pichotta and Mooney, 2014; Rudinger et al., 2015). This drove the establishment of a new task: the *story cloze* test where the goal was to select the correct ending for a story given two endings (Mostafazadeh et al., 2016a; Sharma et al., 2018). Several works showed that incorporating event sequence information provides improvement in this task (Peng et al., 2017; Chaturvedi et al., 2017b). Additionally, some work has focused on defining new script annotation schemes (Mostafazadeh et al., 2016b; Wanzare et al., 2016; Modi et al., 2016) and domain-specific script-based story understanding (Mueller, 2004; McIntyre and Lapata, 2009).

Character-centric approaches adopt the outlook that *characters* make a narrative compelling and drive the story. While no standard paradigms have been established for character representation, a common approach concentrated on learning character types or *personas* (Bamman et al., 2013, 2014). Other work proposed to model inter-character relationships (Krishnan and Eisenstein, 2015; Chaturvedi et al., 2016, 2017a). Information about character types and their relationships has been demonstrated to be useful for story understanding tasks such as identifying incorrect narratives (e.g., reordered or reversed stories) (Elsner, 2012) and detecting narrative similarity (Chaturvedi et al., 2018). Finally, an interesting line of research has focused on constructing "plot units", which are story representations consisting of affect states of characters and tensions between them. Plot units were first proposed by

[2] All data and code are available at `https://github.com/xinru1414/Reddit`

Lehnert (1981) and have recently attracted interest from the NLP community resulting in the development of computational approaches (Appling and Riedl, 2009; Goyal et al., 2010).

Our work takes a unique approach in that we propose a computational technique to learn *functional schemas* from stories. Functional schemas consist of stereotypical sets of functional structures observed in stories. The key difference between functional schemas and scripts is that scripts contain events present in the narrative, while functional schemas consist of phases in a story arc. For example, for a crime story, a script representation may contain a "murder" event, but a functional schema could represent that event as "inciting incident", based on its role in the arc. Functional structures are key to rhetorical structure theory for discourse analysis (Labov, 1996; Labov and Waletzky, 1997) and have been operationalized in discourse parsing (Li et al., 2014; Xue et al., 2015). However, not much work has explored their utility in uncovering novel narrative structures. One exception is Finlayson (2012), which learned functional structures from folktales, indicating that computational techniques could recover patterns described in Propp's theory of folktale structure (Propp, 2010). Our work differs since we aim to uncover new schemas instead of validating existing structural theories. We take this perspective because we are interested in studying stories told on social media which may not conform to existing theories of narrative structure.

2.2 Schema Induction via Topic Models

To computationally extract functional schemas, it is important to identify characteristic functional structures from stories. Topic models, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), can be used for automatic induction of such structures since they identify latent themes, which may be treated as functions, from a set of documents. However, vanilla topic models do not model transitions between themes, whereas stories tend to follow stereotypical sequences of functional structures. For example, the conflict in a story must be set up before the resolution. Hence, to account for the order of functional structures, conversation models can be employed (Ritter et al., 2010; Lee et al., 2013; Ezen-Can and Boyer, 2015; Brychcín and Král, 2017; Joty and Mohiuddin, 2018; Paul, 2012; Wallace et al., 2013; Jo et al., 2017). These models impose structure on transitions between latent themes, typically using an HMM. This uncovers latent themes that account for interactions among themselves, helping to identify dialogue acts, which these models aim to extract. A similar HMM-based framework has been used to extract story schemas from news articles (Barzilay and Lee, 2004).

Among many conversation models, we use the Content word filtering and Speaker preferences Model (CSM), which recently offered the best performance at unsupervised dialogue act identification (Jo et al., 2017). We choose this model because it has some characteristics which make it especially useful for capturing functional structures. Above all, it automatically distinguishes between topical themes and functional structures, which have different behavior. For example, a functional structure that represents asking a question would be characterized by wh-adverbs and question marks, rather than the specific content of questions. Being able to make this distinction between *topics* and *functional structures* is crucial to our task of extracting functional schemas.

3 Method

We use unsupervised algorithms to induce functional schemas from stories. More specifically our pipeline consists of the following stages:

1. **Functional Structure Identification:** We use CSM to identify typical sequences of functional structures.
2. **Story Schema Formation:** We perform PCA to form functional schemas.

3.1 Functional Structure Identification

The first step in our pipeline is to identify the typical sequences of functional structures in the corpus, which will then be clustered to form several functional schemas. Specifically, we utilize CSM to identify underlying functional structures from the corpus.

CSM is a generative model originally applied to conversation – a sequence of utterances by speakers. The model assumes that a corpus of conversations has a set of functional structures undertaken by individual sentences. Each structure is represented as a language model, i.e., a probability distribution over words. CSM can be seen as a combination of an HMM and a topic model but adopts a deliberate design choice different from other models that focus mainly on topical themes. It captures linguistic structures using multiple mechanisms. First, the model encodes that, in a conversation the content being discussed transitions more slowly than the structures that convey the content. Capturing the difference in transition paces allows the model to distinguish co-occurrence patterns of fast-changing words (functional structures) from words that occur consistently throughout (topical themes).

CSM also assumes that each utterance plays some functional role, indicated by structural elements found within it, and that the function is probabilistically conditioned on that of the preceding utterance. This captures dependencies between utterance-level functions and thus those between lower-level structural elements within sentences as well. We can see these dependencies in, for example, a tech forum, where a conversation begins with a user's utterance of "information seeking" comprising such functional structures as introduction, problem statement, and question. This utterance may be followed by another user's utterance of "providing solutions" comprising such functional structures as suggestions and references. Formally, an utterence-level function is represented as a "state", a probability distribution over functional structures.

Since each story in our task is a monologue rather than a conversation, we need to format our data in a way analogous to a conversation to apply CSM. Specifically, we treat each story as a "conversation", and each sentence in the story as an "utterance". Accordingly, each "conversation" has only one speaker. This way, we apply CSM to a corpus of stories, still benefiting from the model's ability to distinguish functional structures from topical themes and account for temporal dependencies between functional structures.

3.2 Functional Schema Formation

After determining functional structures, we identify sets of most strongly co-occurring structures to form functional story schemas. To identify co-occurring structures, we represent each story as a bag of functional structures and run PCA[3]. Each resultant principal component is treated as a schema, consisting of functional structures which have a high loading value for that component. Since principal components are orthogonal, extracted schemas will be distinct. In addition, the set of extracted schemas will be representative of most stories, because PCA retains the variance of the original data. The functional structures present in each schema (based on loading) are treated as elements of that schema.

4 Experiments

4.1 Dataset

We demonstrate the utility of our schema extraction pipeline on Reddit posts[4]. We select three active subreddits to construct our dataset, */r/environment*, */r/ZeroWaste*,and */r/Green*, which cover issues from the environmental domain. We are interested in studying how people structure their experiences and stories differently in each subreddit, though all of them discuss similar topics, as well as the extent to which our extracted functional schemas capture such subtle structural differences. We collect all posts from these subreddits since their inception until Jan 2019. Table 2 summarizes statistics for our dataset.

Subreddit	# of Posts
environment	$3,785$
ZeroWaste	$2,944$
Green	305

Table 2: Dataset Statistics

Using our schema extraction pipeline, we first extract a set of 10 functional structures using CSM[5]. Then using PCA, we derive 10 sets of co-occurring structures as our candidate functional schemas. Next, we manually inspect each set of structures and select the most salient 4 sets as our functional schemas[6]. To validate these schemas, we perform a two-fold evaluation. First, we manually interpret extracted functional structures and schemas. Second, we demonstrate the utility of our schemas by incorporating them into a downstream task: text classification.

[3] Though using PCA in the next phase removes ordering from the final schemas constructed, incorporating ordering during functional structure estimation helps in detecting more salient structures.

[4] According to the Reddit User Agreement, users grant Reddit the right to make their content available to other organizations or individuals.

[5] For CSM specific parameter settings see A.

[6] During manual inspection, we also try to ensure diversity (each set contains different structures).

4.2 Manual Schema Interpretation

In order to interpret the schemas, we first need to label functional structures extracted by CSM. Labeling was performed independently by two annotators who looked at sample sentences assigned to each structure by the model and assigned structure name labels based on inspection. A consensus coding was assembled as the final interpretation after an adjudicating discussion. Table 3 gives a brief overview of the structure labels along with examples for each. We see that the detected structures indeed represent strategies commonly used by Reddit users.

Schemas can now be interpreted based on labels assigned to the structures they contain. Final schema interpretations, along with sample posts for each schema, are presented in table 4. We observe that schema 0 and schema 2 are news and fact oriented, whereas schema 1 and schema 3 include more personal experiences. Moreover, new posts can also be fit into these schemas. We assign a schema to each post P using the following formula:

$$schema(P) = \arg\max_{s \in S} \sum_{t \in s} l_t * \frac{n_t}{n_P} \qquad (1)$$

Here, S is the set of schemas, s is a schema, t is a functional structure, n_t is the number of sentences assigned t in P, n_P is the total number of sentences in P, and l_t is the absolute value of the PCA loading for t.

Figure 1 shows the proportion of posts from each subreddit assigned to each schema. We clearly see that posts from different subreddits follow different schemas. Specifically, half of the posts in subreddit */r/environment* fit into schema 0 and about $1/4$ of the posts fit into schema 2; Schema 1 dominates posts in */r/ZeroWaste*; Posts in */r/Green* occupy schemas 0, 1, 2 and 3 in decreasing numbers. This demonstrates that our extracted schemas do capture typical structures present in Reddit posts and that posts in each subreddit indeed exhibit unique structures.

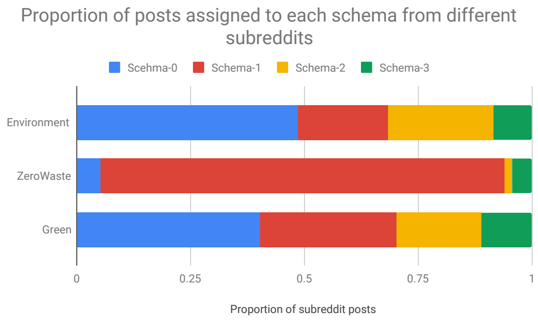

Figure 1: Proportion of schemas for each subreddit

4.3 Using Schemas for Text Classification

In addition to manual interpretation, we demonstrate the practical utility of our schema extraction pipeline by applying it in a downstream task: multi-label text classification. In our task setup, we treat each post as a document and the subreddit it belongs to as the document label. Since all subreddits in our dataset focus on environmental issues, most posts discuss similar topics, making classification using only content information hard. However, as we observed in our schema interpretation, posts from different subreddits follow different schematic structures. Hence, we hypothesize that using schema information should help on this task. As a preliminary experiment, we construct a document representation using only schema-based features. Each document is represented as a 4-dimensional vector consisting of schema scores calculated per equation (1). The performance of logistic regression (**LR**) and support vector machine (**SVM**) classifiers using these feature representations is presented in table 5. These scores demonstrate that schema information is extremely predictive for the classification task in comparison to a majority vote baseline. Encouraged by this result, we conduct further experiments in which schema information is combined with word features. We experiment with both neural and non-neural baseline models for our task. Our models and results are described below.

4.3.1 Baseline Models

We set up the following baseline models, which use only word-level information, for text classification:

- **LR:** A logistic regression classifier with two feature settings (bag-of-words or tf-idf)
- **NB:** A naive bayes classifier with two feature settings (bag-of-words or tf-idf)
- **SVM:** A support vector machine classifier with unigram bag-of-word features
- **BiLSTM:** A bi-directional LSTM with mean-pooling (Yang et al., 2016), followed by an MLP classifier
- **CNN:** A CNN with filter sizes 3,4,5 and max-pooling (Kim, 2014), followed by an MLP

Structure	Label	Examples
0	Requesting help	*any advice would be appreciated* *any ideas on how i can do this*
1	Asking for feedback & thanking	*thanks in advance for your help* *if you want to help please send me a message here*
2	Disclosing personal stories	*i teach global environmental history...* *i'm trying to learn more about being eco-friendly...*
3	Presenting news/statements	*this is called economy of scale* *solar is unreliable expensive and imported*
4	Catch-all for questions	*how happy will we be when our wells are dry* *how do we make up for these losses*
5	Presenting news/facts (numbers)	*85 of our antibiotics come from ascomycetes fungi...* *reduce the global population of bears by two thirds...*
6	Expressing personal opinions	*now i think that landfills are the devil...* *i am sure something can be done with them...*
7	Providing motivation	*we r/environment need to be a vehicle for change...* *we need to engage learn share eulogize and inform*
8	Non-English sentences	*men data siden 2005 viste veksten av disse...* *durant ces cent dernires annes...*
9	Catch-all for personal story bits	*when i asked for a carafe of water he said...* *all i wanted to do was use a cup to get some coffee...*

Table 3: 10 functional structures extracted by CSM along with examples. These structures are more general than narrative primitives appearing in classic theoretical frameworks such as Propp's theory, but we believe that they provide a reasonable approximation.

classifier

For all models using bag-of-words or tf-idf features, we restrict the vocabulary to the most frequent $2,000$ words. All neural models use 300-dimensional GloVe embeddings (Pennington et al., 2014).

4.3.2 Schema-based Extension Models

To incorporate schema features alongside word-level features, we adopt a strategy inspired by domain adaptation techniques (Daume III, 2007; Kim et al., 2016). Daume III (2007) proposed a feature augmentation strategy for domain adaptation, which was extended to neural models by (Kim et al., 2016). It works as described: given two domains ("source" and "target"), each feature is duplicated thrice creating three versions – a general version, a source-specific version and a target-specific version. We follow the same intuition considering each schema to be a separate domain. Hence, we duplicate each feature 5 times (a general version and 4 schema-specific versions). For example, if a document contains the word "plastic", our feature space includes "general_plastic", "schema0_plastic", and so on. We experiment with several feature duplication strategies, resulting in the following settings for each model:

- **Vanilla:** Only the general domain features contain non-zero values. All schema domain features are set to zero, hence this setting contains no schema information.
- **AllSent:** Both general and schema domains contain non-zero feature values computed using sentences from the entire document. For each document, only one schema domain (i.e. assigned schema) contains non-zero values.
- **SchemaSent:** General domain feature values are computed using the entire document, while schema domain feature values are computed using only sentences which contain structures present in the assigned schema.

4.3.3 Results

To evaluate the performance of all models on our text classification task, we create a held-out test set using 10% of our data. The remaining data is divided into *train* and *dev* sets. To avoid double-dipping into the same data for both schema learning and subreddit prediction, we use *dev* set to learn schemas, and train **AllSent** and

Schema	Interpretation	Examples
0	Presenting news/facts, asking questions and providing motivation	*deforestation in the amazon can hardly be a headline for forty years running...how happy will we be when our wells are dry...right now the jaguars are on the rise and i have hope*
1	Disclosing personal problems or opinions, sharing story snippets and providing motivation	*i am not a techsavvy person...i literally know the bare minimum of how a computer works*
2	Presenting news/facts, asking questions and sharing story snippets	*the commission by environmental campaigners forecast 3 trillion euros would generate by 2050...it has yet to achieve agreement on binding targets beyond 2020...the crown report finds almost totally green energy would lead to half a million extra jobs*
3	Disclosing personal problems, presenting facts and requesting help	*i just got this job the only job i've been able to find for the last year...we work on different studies each week for the likes of bayer and monsanto...i know i should stop pestering the internet for help but you're so benevolent*

Table 4: Manual interpretation for 4 schemas extracted by PCA, along with example post sinppets. Note that the functional structures in each schema may appear in any order in the post, not necessarily the one presented here

Model	Accuracy
LR	83.64%
SVM	82.79%

Table 5: Accuracy of classifiers using only schema features for text classification. Majority vote accuracy is 53.34%

SchemaSent models on *train* data only. However for the **Vanilla** setting, we can use both *train* and *dev* sets for training since no schema information is used. Because we need a large dev set to learn good schemas, we perform a 50 : 50 split to create train and dev sets. Exact statistics are provided in table 6.

Split	# of Posts
Train	3, 166
Dev	3, 165
Test	703

Table 6: Dataset split statistics

Table 7 shows the performance of all models in different settings on the text classification task. We observe that for both neural and non-neural models, incorporating schema information helps in all cases, the only exception being **NB-BoW**.

We also notice that neural and non-neural models achieved comparable performance which is surprising. To further investigate this, we look into precision recall and F1 scores of the best model for each type respectively i.e. **NB-BoW Vanilla** and **CNN AllSent**. Our investigation shows that unlike NB-BoW, the CNN model completely ignores the minority subreddit */r/Green*, which we believe could be due to the fact that our dataset is extremely small for neural models.

Model	Vanilla	AllSent	SchemaSent
LR-BoW	80.2%	**85.1%**	84.8%
LR-TfIdf	81.4%	80.7%	**81.7%**
NB-BoW	**86.9%**	78.0%	77.2%
NB-TfIdf	69.6%	**79.8%**	79.2%
SVM	77.8%	83.8%	**85.2%**
BiLSTM	82.4%	79.8%	**82.9%**
CNN	85.2%	**87.3%**	86.6%

Table 7: Accuracy of all models on text classification

5 Discussion

Our interpretation and experiments demonstrate that the extracted functional schemas uncover novel narrative structures employed by Reddit users. We also observe that functional schemas are differently distributed across subreddits, indicating that communities follow diverse storytelling practices, even when discussing similar topics. These subtle schema differences between narratives across subreddits can aid us in discerning how users structure stories differently when participating in different communities. In our case, extracted schemas show that users in subreddits */r/environment* and */r/Green* use more fact-oriented functions while telling stories (high abundance of stories fitting schemas 0 and 2), whereas users in subreddit */r/ZeroWaste* use more personal experience-oriented functions (high abundance of stories fitting schemas 1 and 3). We highlight this by giving prototypical example posts with assigned schema labels for each subreddit below:

> *...there is so many problems today with plastic strawsthe uk and the us use a combined total of 550 million plastic straws each day and unfortunately its safe to say that not all 550 million of these plastic items are recycled ...*
> (**/r/environment**, Schema 0)

> *...every single year plastic cards hotel key cards etc amount to 75 million pounds of pvc wasted or about 34000 tonsthe eiffel tower weighs just around 10000 tonsthis is the equivalent of burying around 3 eiffel towers a year just from used pvc cards...*
> (**/r/Green**, Schema 0)

> *...i had a few vegetables that were wilting and ready to be discarded...instead i made a soup with all of them and some broth and miso...it's good and isn't wasteful...*
> (**/r/ZeroWaste**, Schema 1)

More importantly, these narrative structures unique to each subreddit, as captured by functional schemas, can act as a lens and provide insight into community posting norms. This is analogous with previous work on computational sociolinguistics, where researchers have demonstrated that online discussion forums create community norms about language usage, and members adapt their language to conform to those norms (Nguyen et al., 2016). Especially on Reddit, language style is an essential indicator of community identity (Tran and Ostendorf, 2016; Chancellor et al., 2018). Our schemas help us make similar observations, showing that dominant user posting styles in each subreddit seem to be ones that conform to subreddit descriptions. Figure 2 presents descriptions for all subreddits which we use in our dataset. We see */r/environment* and */r/Green* specifically position themselves as platforms to discuss news and current issues, which is also recovered by our functional schemas since they contain an abundance of news and fact related functions. On the other hand, */r/ZeroWaste* positions itself as a platform for like-minded people, resulting in dominant schemas demonstrating an abundance of functional structures related to describing personal experiences. This indicates that our technique of inducing functional schemas from social media posts is useful for drawing interesting insights about how narratives align to community norms in online discussion forums.

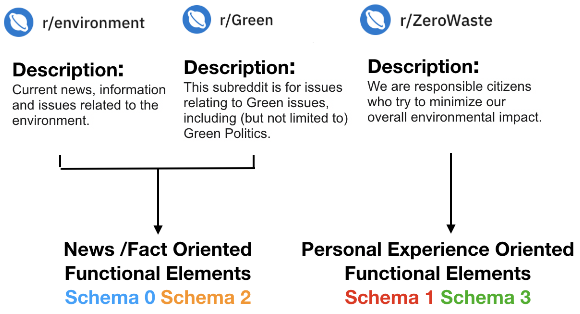
r/environment
r/Green
r/ZeroWaste
Description:
Current news, information and issues related to the environment.
Description:
This subreddit is for issues relating to Green issues, including (but not limited to) Green Politics.
Description:
We are responsible citizens who try to minimize our overall environmental impact.
News /Fact Oriented Functional Elements
Schema 0 Schema 2
Personal Experience Oriented Functional Elements
Schema 1 Schema 3

Figure 2: Subreddit description corresponding to schemas

6 Conclusion & Future Work

In this work we propose a novel computational approach to understand social media narratives. We present a unique take on story understanding, focusing on learning functional story schemas which are sets of typical functional structures. We first introduce a computational pipeline utilizing unsupervised methods such as CSM and PCA, to extract schemas and use it on social media data (posts from different communities on Reddit). We then validate learned schemas through human interpretation and a downstream text classification task. Our interpretation shows typical posting strategies used by community members and our experiments demonstrate that integrating schema information improves the performance of baseline models on subreddit prediction. Finally, we observe that functional schemas not only capture specific narrative structures existing in subreddits, but also reveal online community norms, which helps us better understand how stories function in social media.

A limitation of our work is that PCA-based grouping loses information about ordering of functional structures within each schema. Moving forward, we plan to tackle this to form ordered schemas. Possible applications of our work include using extracted schemas to study evolution of community norms and changes in user compliance to these norms over time.

7 Acknowledgements

This research was funded by Dow. The authors would like to thank the anonymous reviewers for their constructive feedback.

References

D Scott Appling and Mark O Riedl. 2009. Representations for learning to summarize plots. In *AAAI Spring Symposium: Intelligent Narrative Technologies II*, pages 1–4.

Niranjan Balasubramanian, Stephen Soderland, Mausam, and Oren Etzioni. 2013. Generating coherent event schemas at scale. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1721–1731, Seattle, Washington, USA. Association for Computational Linguistics.

David Bamman. 2015. *People-Centric Natural Language Processing*. Ph.D. thesis, Carnegie Mellon University.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *HLT-NAACL 2004: Main Proceedings*, pages 113–120, Boston, Massachusetts, USA. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

William F Brewer and Edward H Lichtenstein. 1980. Event schemas, story schemas, and story grammars. *Center for the Study of Reading Technical Report; no. 197.*

William F Brewer and Edward H Lichtenstein. 1982. Stories are to entertain: A structural-affect theory of stories. *Journal of pragmatics*, 6(5-6):473–486.

Tomáš Brychcín and Pavel Král. 2017. Unsupervised dialogue act induction using Gaussian mixtures. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 485–490, Valencia, Spain. Association for Computational Linguistics.

Nathanael Chambers. 2013. Event schema induction with a probabilistic entity-driven model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Seattle, Washington, USA. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 976–986, Portland, Oregon, USA. Association for Computational Linguistics.

Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms matter: contrasting social support around behavior change in online weight loss communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 666. ACM.

Snigdha Chaturvedi, Mohit Iyyer, and Hal Daume III. 2017a. Unsupervised learning of evolving relationships between literary characters. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017b. Story comprehension for predicting what happens next. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1603–1614, Copenhagen, Denmark. Association for Computational Linguistics.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daume III, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Snigdha Chaturvedi, Shashank Srivastava, and Dan Roth. 2018. Where have I heard this story before? identifying narrative similarity in movie remakes. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 673–678, New Orleans, Louisiana. Association for Computational Linguistics.

Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende. 2013. Probabilistic frame induction. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 837–846, Atlanta, Georgia. Association for Computational Linguistics.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Micha Elsner. 2012. Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644, Avignon, France. Association for Computational Linguistics.

Aysu Ezen-Can and Kristy Elizabeth Boyer. 2015. Understanding student language: An unsupervised dialogue act classification approach. *Journal of Educational Data Mining (JEDM)*, 7(1):51–78.

Mark Alan Finlayson. 2012. *Learning narrative structure from annotated folktales*. Ph.D. thesis, Massachusetts Institute of Technology.

Amit Goyal, Ellen Riloff, and Hal Daume III. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 77–86, Cambridge, MA. Association for Computational Linguistics.

Yohan Jo, Michael Yoder, Hyeju Jang, and Carolyn Rosé. 2017. Modeling dialogue acts with content word filtering and speaker preferences. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2179–2189, Copenhagen, Denmark. Association for Computational Linguistics.

Ian Jolliffe. 2011. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer.

Shafiq Joty and Tasnim Mohiuddin. 2018. Modeling speech acts in asynchronous conversations: A neural-CRF approach. *Computational Linguistics*, 44(4):859–894.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Young-Bum Kim, Karl Stratos, and Ruhi Sarikaya. 2016. Frustratingly easy neural domain adaptation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 387–396, Osaka, Japan. The COLING 2016 Organizing Committee.

Vinodh Krishnan and Jacob Eisenstein. 2015. "you're mr. lebowski, I'm the dude": Inducing address term formality in signed social networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1616–1626, Denver, Colorado. Association for Computational Linguistics.

William Labov. 1996. Some further steps in narrative analysis. *The Journal of Narrative and Life History. Special Issue: Oral Versions of Personal Experience: Three Decades of Narrative Analysis*, 7.

William Labov and Joshua Waletzky. 1997. Narrative analysis: oral versions of personal experience.

Donghyeon Lee, Minwoo Jeong, Kyungduk Kim, Seonghan Ryu, and Gary Geunbae Lee. 2013. Unsupervised Spoken Language Understanding for a Multi-Domain Dialog System. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2451–2464.

Wendy G Lehnert. 1981. Plot units and narrative summarization. *Cognitive science*, 5(4):293–331.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 2061–2069, Doha, Qatar. Association for Computational Linguistics.

Neil McIntyre and Mirella Lapata. 2009. Learning to tell tales: A data-driven approach to story generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 217–225, Suntec, Singapore. Association for Computational Linguistics.

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2016. InScript: Narrative texts annotated with script information. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3485–3493, Portorož, Slovenia. European Language Resources Association (ELRA).

Raymond J Mooney and Gerald DeJong. 1985. Learning schemata for natural language processing. In *IJCAI*, pages 681–687.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016a. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016b. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics.

Erik T Mueller. 2004. Understanding script-based stories using commonsense reasoning. *Cognitive Systems Research*, 5(4):307–340.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Survey: Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, and Romaric Besançon. 2015. Generative event schema induction with entity disambiguation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 188–197, Beijing, China. Association for Computational Linguistics.

Ruth Page. 2013. *Stories and social media: Identities and interaction*. Routledge.

Michael J. Paul. 2012. Mixed membership Markov models for unsupervised conversation modeling. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 94–104, Jeju Island, Korea. Association for Computational Linguistics.

Haoruo Peng, Snigdha Chaturvedi, and Dan Roth. 2017. A joint model for semantic sequences: Frames, entities, sentiments. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 173–183, Vancouver, Canada. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Karl Pichotta and Raymond Mooney. 2014. Statistical script learning with multi-argument events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 220–229, Gothenburg, Sweden. Association for Computational Linguistics.

Vladimir Propp. 2010. *Morphology of the Folktale*, volume 9. University of Texas Press.

Michaela Regneri, Alexander Koller, and Manfred Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 979–988, Uppsala, Sweden. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

Rachel Rudinger, Pushpendre Rastogi, Francis Ferraro, and Benjamin Van Durme. 2015. Script induction as language modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1681–1686, Lisbon, Portugal. Association for Computational Linguistics.

Roger C Schank and Robert P Abelson. 1977. Scripts. *Plans, Goals and Understanding*.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the story ending biases in the story cloze test. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 752–757, Melbourne, Australia. Association for Computational Linguistics.

Trang Tran and Mari Ostendorf. 2016. Characterizing the language of online communities and its relation to community reception. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1035, Austin, Texas. Association for Computational Linguistics.

Byron C. Wallace, Thomas A. Trikalinos, M. Barton Laws, Ira B. Wilson, and Eugene Charniak. 2013. A generative joint, additive, sequential model of topics and speech acts in patient-doctor communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1765–1775, Seattle, Washington, USA. Association for Computational Linguistics.

Lilian DA Wanzare, Alessandra Zarcone, Stefan Thater, and Manfred Pinkal. 2016. Descript: A crowdsourced corpus for the acquisition of high-quality script knowledge. In *Proceedings of the*

Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 3494–3501.

Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

A Appendices

A.1 CSM Parameter Values

Various parameter values were tested and the final parameter setting was chosen based on model performance and the parameter setting suggested in the original paper (Jo et al., 2017).

We found the optimal number of functional structures to be 10. Higher numbers tend to capture too content-specific structures, and lower numbers too general structures. The optimal number of content topics is 5, which indicates that the corpus is focused on environmental related issues and the content is relatively common across the corpus. The number of states reflects different patterns of structure composition within a post, and 5 states were found to be optimal. More states tend to capture too post-specific structures, and less states cannot account for the diversity of structures.

Parameter $\nu \in [0, 1]$ is the weight on state transition probabilities (as opposed to speaker preferences) for determining an utterance's state. 1 means only state transition probabilities are considered, and 0 means only speaker preferences are considered. In our study, we treat each post as a "conversation" that has only one speaker. Therefore, a low weight would identify functional structures that distinguish between posts rather than between sentences. We find a high weight ($\nu = 0.9$) drives the model to identify sentence structures well that also account for some consistency within each post. Parameter $\eta \in [0, 1]$ is the weight on structure language models (as opposed to content topics) for generating words. 1 means that all words are generated from structure language models, and 0 means only from content topics. Our setting ($\eta = 0.8$) filters out 20% of words as content. This is quite a large proportion compared to the original paper, meaning that the corpus has a relatively large proportion of words that constitute functional structures.

Other hyperparameters for the model were set as per the original paper: $\alpha^F = \gamma^A = 0.1, \alpha^B = \gamma^S = 1, \beta = 0.001$.