Story Forms Detection in Text through Concept-based Co-clustering

*Sultan Alzahrani**, *Betul Ceran**, *Saud Alashri**, *Scott W. Ruston*†, *Steven R. Corman*† *and Hasan Davulcu**
**School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287-8809*
Email: {ssalzahr, betul, salashri, hdavulcu}@asu.edu
†*Hugh Downs School of Human Communication, Arizona State University, Tempe, AZ 85287-1205*
Email: {scott.ruston, steve.corman}@asu.edu

***Abstract*—A story is defined as actors taking actions that culminate in resolutions. In this paper, we extract *subject - verb - object* relationships from paragraphs and generalize them into semantic conceptual representations. Overlapping generalized concepts and relationships correspond to archetypes/targets and actions that characterize story forms. We present an analytic framework which implements co-clustering based on generalized conceptual relationships to automatically detect such story forms. Co-clustering can help in identifying similarities that exist in low-dimensional sub-spaces of sparse data such as textual paragraphs. Through co-clustering, we detect not only the clusters themselves but also their characteristic features which can be useful in describing and summarizing their contents. We perform co-clustering of stories using two different types of features: standard unigrams/bigrams and generalized concepts. We show that the residual error of factorization with concept-based features is significantly lower than the error with standard keyword-based features. Qualitative evaluations also suggest that concept-based features yield more coherent, distinctive and interesting story forms compared to those produced by using standard keyword-based features.**

***Index Terms*—Story forms, Narrative analysis, Non-negative matrix factorization, Co-clustering.**

1. Introduction

A key component of spreading an ideology is the utilization of cultural narratives tailored to specific target audiences. For example, extremists are known to adopt historically deeply rooted narratives from the cultural heritage of their target audience in order to gain their attention. Narratives are systems of stories that are linked by common archetypes, forms and themes. A story is defined as an actor(s) taking an action(s) that culminates in a resolution(s). The actors, actions and resolutions in these stories form the template of a strategic message regarding the current events which is used by extremists to justify their actions and policies, persuade their target audience and gain followers. Hence identifying and countering the story forms found in these messages is an essential part of counter violent extremism efforts. This paper presents a framework to help subject matter experts rapidly analyze large collections of extremist narratives and discover their underlying story forms. The framework streamlines the narrative analysis by providing bi-clusters of stories and their underlying characteristic features. Another contribution of this framework is the improved clustering performance obtained through the utilization of concept-based features compared to widely-used standard keyword-based features.

The recurring themes in extremist narratives can be categorized into general story forms. These story forms are characterized by archetypes and their actions. We aim to reveal information about the story forms in our dataset via clustering analysis. We observe that generalized concepts derived from extracted overlapping *subject - verb - object* relationships are better suited to be used as features in clustering since they provide information regarding the underlying semantic structure of these story forms.

Clustering is an essential step towards the analysis of data without prior labels or categories. Two critical aspects of clustering are; i) semantic quality of resulting clusters and ii) their descriptive features; i.e. clusters should be self-descriptive in order to present a meaning to the user. Conventional clustering methods provide ample ways to group data. However, they do not automatically yield descriptive features for the groups without further processing (i.e. through a classifier). Co-clustering, on the other hand, identifies both the underlying groups, out of the data, along with their characteristic features. It simultaneously clusters the rows and columns of an input matrix generating a subset of instances which exhibit high similarity across a subset of features, called bi-clusters. Since descriptions of the clusters are produced simultaneously with clustering, co-clustering presents an advantage over conventional clustering methods for our application.

The initial efforts in co-clustering text data relied on term-document matrices and lexical features, mainly n-grams. In this paper, we perform co-clustering of stories using two different types of features: standard unigrams/bigrams and generalized concepts that rely on extracted linguistic roles. We employ the model developed in Ceran et. al and Alashri et. al. [1], [2] to produce generalized concept-based representations of extremist stories. We show that the residual error of factorization with concept-based features is lower than the residual error with standard keyword-based features. Qualitative evaluations also suggest

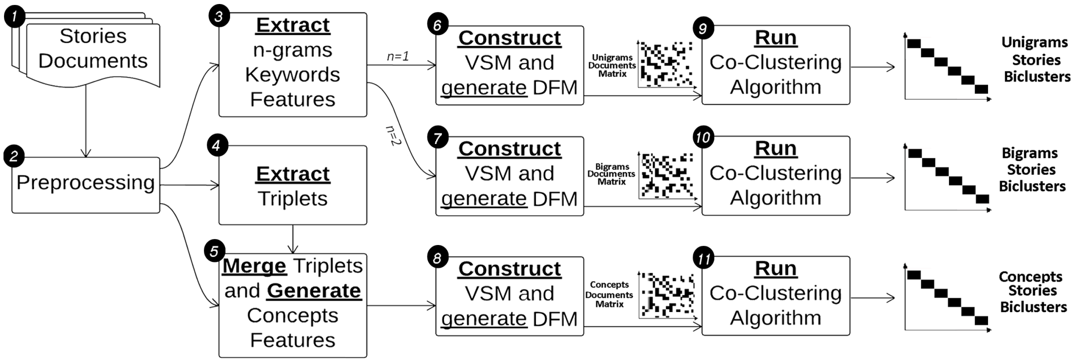

Figure 1. System Architecture

that concept-based features yield more coherent, distinctive and interesting story forms compared to those produced by utilizing standard keyword-based features.

The rest of the paper is organized as follows. § 2 presents related works. The system architecture is presented in § 3. § 4 describes the extraction of lexical and linguistic features and co-clustering algorithms. In § 5, we present experimental evaluations and results. § 6 concludes the paper.

2. Related Work

Various types of co-clustering algorithms based on matrix factorization, probabilistic and geometric models have been developed in literature [3] and they have been applied in many different domains such as text, bioinformatics, and image analysis.

One of the pioneering works in this area is Dhillon et. al. [4] which introduces spectral co-clustering of documents and their terms by leveraging the singular value decomposition of the term-document matrix. Non-negative Matrix Factorization (NMF) has also been adapted for co-clustering [5]. Different versions of NMF for co-clustering such as [6] have been developed to improve the initial model.

Kok and Domingos [7] proposed a clustering framework to produce generalized concepts and relations similar to the concept-based features utilized in this paper. Their algorithm is purely statistical and relies on second order Markov Logic. They compared their results with other algorithms on a gold-standard clustering scheme created by a human. Kang et al. [8] used tensor decomposition in order to obtain contextual synonyms, and generalized concepts/relationships; however, they do not present any formal evaluations of their results. We generate the concept-based features using the method proposed in Ceran et. al. [1] and and Alashri et. al. [2].

Jing et.al. [9] studied co-clustering text along with term and concept features, which were generated using Wikipedia. They present a higher-order co-clustering framework where they improve the performance of conventional clustering. They present an evaluation of their work on the labeled benchmark Reuters data set. The concepts used in [9] are named entities which are extracted using an external information source (Wikipedia), whereas the concept-based features that we use are in the form of generalized relational semantic triples which are produced by processing the document set itself.

3. System Architecture

The components of our framework are presented in Figure 1. Each component is briefly described below, and the technical details are presented in the following sections.

- The input document set consisting of stories. We analyze the data at the paragraph level, i.e. each document contains a single story paragraph. We apply pre-processing in order to clean and prepare the paragraphs for feature extraction. (Steps 1 and 2)
- Three different feature sets (unigrams, bigrams and generalized concepts/relations) are generated from the story paragraphs. Concepts/relations are obtained using the method proposed in [1]. Triplet generation step has been modified according to [2]. (Steps 3, 4 and 5)
- Unigrams and bigrams are ranked based on their TF-IDF values, by initially generating Vector Space Model (VSM) corresponding to each feature set and then followed by constructing the Document-Feature-Matrices (DFMs), for short, feature matrices (Steps 6 and 7).
- A binary feature matrix is also created with concept/relational features. (Step 8)
- Co-clustering algorithm is run on unigram, bigram and concept-based feature matrices to produce bi-clusters of story forms and their associated features. (Steps 9, 10 and 11)
- Quantitative and qualitative evaluations are performed on the resulting bi-clusters, § 5.

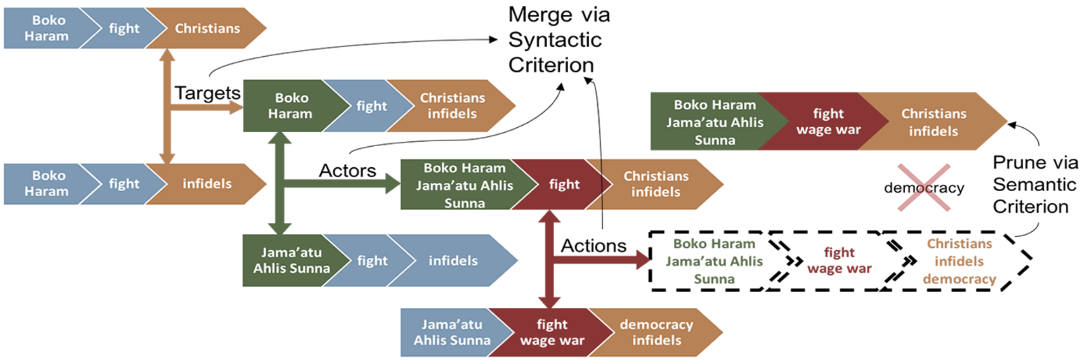

Figure 2. Generation of concepts/relations form ⟨subjects, verbs, objects⟩ triplets

4. Methodology

4.1. Problem Definition

For a given a set of documents comprising stories $\{D_1, \ldots, D_N\}$ where N denotes the number of documents, we generate two sets of features: n-grams features and the generalized concepts. Our main objective is to identify which type of features yield better bi-clustering of stories into story-forms. We evaluate the quality, initially quantitatively by employing NMF residual error measure presented in § 5.3.2 and qualitatively in collaboration with a subject matter expert in § 5.4

4.2. N-gram Features

We extracted highest ranked unigrams and bigrams by utilizing term frequency - inverse document frequency (TF-IDF), a simple form of cross entropy and a popular technique used in informational retrieval tasks.

4.3. Generalized Concept-based Features

Subject - verb - object triplet extraction is the basic building block towards generalized concepts. We first process the story corpus to resolve its co-references using state-of-the-art coreference resolvers [10], [11], [12], [13]. Previously, Ceran et. al. [14] utilized ClearNLP [15] to extract triplets. However, using this triplet extractor alone resulted in poor recall. Alashri et al [2] proposed an enhanced approach that utilized additional triplet extractors: AlchemyAPI [16], Everest [17], Reverb [18] and implemented a Cartesian product of the atomized phrases in all argument positions to double the production of extracted triplets. Ceran et. al [1] utilized generalized triplets and compared their performance with keywords in a classification model to show that the triplets yield a significant 36% boost in performance for the story detection task. However, although triplets as features carried more semantic information, they were showing high sparsity during matching across the document corpus. Hence, we moved to a generalized triplet representation by suitably "merging triplets" into generalized concepts without a drift.

In [1], we utilize both syntactic and semantic corpus-based merging criteria to generalize triplets into concepts. A pair of ⟨subject⟩-⟨verbs⟩-⟨objects⟩ triplets is merged further only if (i) they share a common context among their corresponding terms (i.e. syntactic criteria) and (ii) they satisfy corpus-based support and similarity measure of "contextual synonymy" (i.e. semantic criteria) between their newly added terms and existing terms. Next, a hierarchical bottom-up merging algorithm allows information to propagate between clusters of related subjects, verbs and objects leading to a set of generalized concepts.

Figure 2 illustrates an instance of how syntactic and semantic criteria are applied on a sample set of triplets extracted from our story corpus. Initially, syntactic criterion is satisfied between the pair of triplets: ⟨Boko Haram, fight, Christians⟩ and ⟨Boko Haram, fight, infidels⟩ since they share a common (subject, verb) context (Boko Haram, fight). Hence, this pair becomes a candidate for merging if a "contextual synonymy" relationship exists between their newly added and existing terms (i.e. *Christians* and *infidels*). Contextual synonyms are not synonyms in the traditional dictionary sense, but they are phrases that may occur in similar semantic roles and associated with similar contexts. In the next step the resulting generalized concept ⟨Boko Haram, fight, {Christians, infidels}⟩ can be merged with ⟨Jama'atu Ahlis Sunna, fight, infidels⟩ due to their shared (verb, object) context: (fight, infidels) meeting the syntactic criteria, and due to the existence of contextual synonymy relationship between *Boko Haram* and *Jama'atu Ahlis Sunna*. Syntactic criterion is applied iteratively to identify candidate concepts for merging in combination with the application of semantic criterion to screen for the introduction of new topics that could cause a generalized concept to drift from it original meaning. Let us consider an additional pair of candidates for merging based on syntactic criteria: ⟨{Boko Haram, Jama'atu Ahlis Sunna} fight, {Christians, infidels}⟩ and ⟨Jama'atu Ahlis Sunna, {fight, wage war}, {infidels, democracy}⟩. First, a core component is created using only the intersections of the subject, verb, object sets of these two concepts: ⟨Jama'atu Ahlis Sunna, fight, infidels⟩. The remaining words from the two candidates can be added to the core concept only if they are among the closest contextual synonyms of at least one of the already existing members in the core item. For example, the algorithm would permit the addition of *Boko Haram*, *wage war* and *Christians* to the resulting set since the newly added terms are among the closest contextual synonyms of *Jama'atu Ahlis Sunna*, *fight* and *infidels* in their respective argument positions. However, *democracy* would be left out of the object argument position in the resulting generalized concept since it is not among the contextual synonyms of neither *infidels* nor the *Christians* according to the corpus based definition of "contextual synonymy" (i.e. semantic criteria).

4.4. Co-Clustering

Clustering is an unsupervised learning technique that tries to draw inferences from given data where data labels (i.e. classes) are concealed or unknown, as in our case). This approach is adopted to assist in benchmarking unigrams, bigrams and the generalized concepts as features when used for story forms detection in a story corpus. We aim to investigate which feature set will provide us with the highest quality bi-clustering results. Comparing different feature sets while applying co-clustering will not only allow us to determine which feature set can quantitatively perform better but also, it can prompt us about which feature set could provide more coherent, distinctive and interesting story forms as clusters.

To formalize the co-clustering problem, let's assume our story corpus contains M documents and N features provided as the matrix $A = (a_{ij})_{M \times N}$ such that a_{ij} represents the entry value of i-th story document and j-th feature. The A *feature-term-matrix* can be also written as $A = (R, C) \in \Re^{M \times N}$ where $R = \{1, 2, \dots, M\}$ denotes row indices, and $C = \{1, 2, \dots, N\}$ denotes column indices. Here, our objective is to find set of sub-matrices or bi-clusters, say $B_k(X_k, Y_k)$, such that $X = M_1, \dots, M_a \subseteq R$ and $Y = N_1, \dots, N_b \subseteq C$ as separate subsets of R and C. This task is an NP-hard problem [19], but an optimization approach with a greedy iterative search utilizing Non-negative Matrix Factorization (NMF) has been shown to produce effective results.

4.4.1. Non-negative Matrix Factorization (NMF). Co-clustering is an NP-hard problem, yet many different optimization based approximation algorithms have been developed in the literature. One of those is the non-negative matrix factorization or decomposition based method which factorizes a given matrix into multiple matrices revealing substructure patterns within the matrix. This method has been widely used in many applications such as bioinformatics, image processing, and text mining. NMF can be used to factorize our $A \in \Re^{M \times N}$ *feature-term-matrix* into a pair matrices $U \in \Re^{M \times K}$ and $V \in \Re^{K \times N}$ having non-negative elements, such that $A \simeq UV$ constructing an approximation, as shown in Equation 1. U represents the basis vectors (or factors), and V represents the coefficients on the linear combination of the factors that allows construction of the original A *feature-document-matrix*. K variable can be used as the number of clusters and it has to satisfy $K < min\{M, N\}$. The mathematical formulation of the optimization problem is written as follows:

$$\begin{aligned} \underset{U,V}{\text{minimize}} \quad & \frac{1}{2}\left\|X - UV\right\|_F^2 + \frac{1}{2}\left\|U\right\|_F^2 + \frac{\lambda}{2}\sum_{i=1}^{n}\left\|v_i\right\|_1 \\ \text{subject to} \quad & U, V \geq 0 \end{aligned}$$

The above optimization problem is a modified version of NMF proposed by [6] since the standard form of NMF has shortcomings of non-unique and scale-variance outputs. Kim et al. [6] enhanced sparseness degree of basis vectors by introducing regularizations as well as alternating negative constraints update technique based on the multiplicative update. The multiplicative update, in the standard NMF, does not necessarily yield sparse basis vectors. The sparse optimization problem can be solved using non-negative quadratic programming (NNQP). The modified NMF compares different feature sets by looking into the residual error E after factorization, where E, shown in Equation 1, is the error term after decomposing A matrix into U and V. Lower E values indicate better underlying structure detection in A. We used the software package in [20] for the implementation of the Sparse NMF.

$$A = UV + E \tag{1}$$

5. Experimental Results

5.1. Corpus

Our story corpus consists of $6,856$ paragraphs which are pulled from a database of Islamist extremist texts. Texts are collected from online sources websites, blogs and other news sources that are known to be outlets of extremist groups such as Al-Qaeda, ISIS or their followers who sympathize with their cause and methods. Extremists' texts are not entirely composed of stories. After the crawling process, subject matter experts annotated the paragraphs based on a coding system, consisting of eight mutually-exclusive categories: story, exposition, imperative, question, supplication, verse, annotation, and other. A paragraph is labeled as a story if it tells a sequence of related events, leading to a resolution or projected resolution. In our experiments, we work on the paragraphs which are coded as stories.

5.2. Number of Clusters

There is no ground truth of story forms available for our story corpus therefore, we resort to additional analysis to determine the number of clusters before we present the results to subject matter experts for qualitative evaluation. Determining the right number of clusters has been a challenging problem in clustering and various techniques have been suggested in the literature to solve this problem. First, we obtain an embedding of $stories \times concepts$ feature matrix into 2-D. We use t-Distributed Stochastic Neighbor Embedding (t-SNE) [21] technique to reduce the data dimension and visualize the block diagonal sub-structures. Next, we use an external measure from literature, Calinski-Harabasz index [22], to measure the quality of a clustering across different numbers of clusters. Calinski-Harabasz index or variance ratio criterion (VRC) is proportional to the ratio of the overall between-cluster variance and the overall within-cluster variance. In this scheme, the higher corresponding VRC value, the better the clustering performance. Figure 3 (a) shows a plot of VRC values across a number of clusters ranging from 2 to 14. The rule of thumb suggested in the literature is to discern the values which cause a sharp spike in the VRC plots. In Figure 3 (b), we can see that there is a sharp spike at 6 clusters. This indicates that setting the number of clusters to 6 is a plausible choice in order to obtain a good clustering scheme. Figure 3 (b) shows the scatter plot of 6 clusters obtained using K-means after 2D embedding. The cluster centroids are also marked alongside their error ellipses representing their covariance matrices.

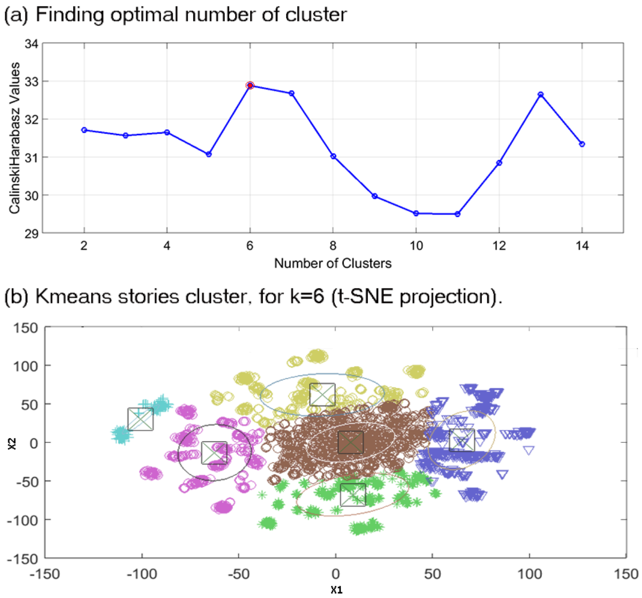

Figure 3. Analysis to determine number of clusters

5.3. Quantitative Evaluation

5.3.1. Block diagonal sub-structure. Figure 4 shows the block diagonal sparsity structures of the feature matrices after clustering, in which, the bi-clusters (unigrams vs. stories bi-clusters in (a), bigrams vs. stories in (b), concepts vs. stories in (c)) are represented along the diagonal blocks. Block diagonal plots are obtained by reordering the indices of stories and their characteristic features in each row and column for each cluster to show their groupings.

5.3.2. NMF residual error. Figure 4 (d) show the residual error of non-negative matrix factorization of the three different feature matrices across different numbers of clusters. Residual error is computed by using the formula shown in Equation 1. The error decreases as the number of clusters increases since the number of clusters also represents the dimensionality of the resulting approximation matrices. In residual error plots, it can be observed that the concept-based features consistently yield lower residual errors compared to both the unigram and bigram based features.

5.4. Qualitative Evaluation

To determine if the clusters generated by the concept-relations technique yielded a valuable analytic tool and an improvement over other clustering methods for the anticipated use case (rapidly analyzing large amounts of story text to determine themes and overarching narratives to benefit strategic communication and counter-messaging activities), a subject matter expert (SME) conducted a qualitative evaluation.

Six clusters were generated using the concept-relations technique discussed here and six other clusters were created using bi-gram co-clustering techniques- Tables 1 and 2 provide a summary for each clustering scheme based on two feature sets. The SME read the stories drawn from each of the twelve clusters (i.e. 6 for each) without cluster identification noting narratively significant features such as the protagonists and antagonists, types of actions taken, and evident and implied resolutions. Subsequently, the SME also conducted an evaluation of the feature sets of each clustering method, looking for patterns and indicators of meaning useful to a communication analyst. These efforts were synthesized to draw conclusions about the clusters.

5.4.1. Concept Clusters vs. Bigram Clusters. In general, both clustering methods produced some distinct clusters that make sense under qualitative evaluation. Notably, the dataset is dominated by stories with an overall structure described by previous analysis as a “victorious battle story”. In this story form, a protagonist (member of some extremist group) takes some form of military action killing or injuring antagonists (US forces or police) [23]. The prevalence of this basic story form within the dataset complicates identifying robustly distinct clusters in terms of narrative significance. This is because the characteristics that distinguish groups of stories tend to be the terms used for protagonists and antagonists and the settings, whereas the general meaning (successful attack by insurgent forces) remains relatively constant.

However, despite that limitation, the concept clustering method produced meaningful clusters with notable distinctions and with useful implications for communication analysis. The bigram cluster method produced clusters with less distinctiveness and significance in terms of overall meaning. For example, bigram clusters 2 and 3 are nearly impossible to distinguish, involving similar stories, nearly identical actions, and having a wide range of protagonists and antagonists. In the set of concept clusters, clusters 1, 2, 5 and 6 were the most distinctive clusters, especially 5 and 6. The stories in cluster 5 are very similar: mujahidin in

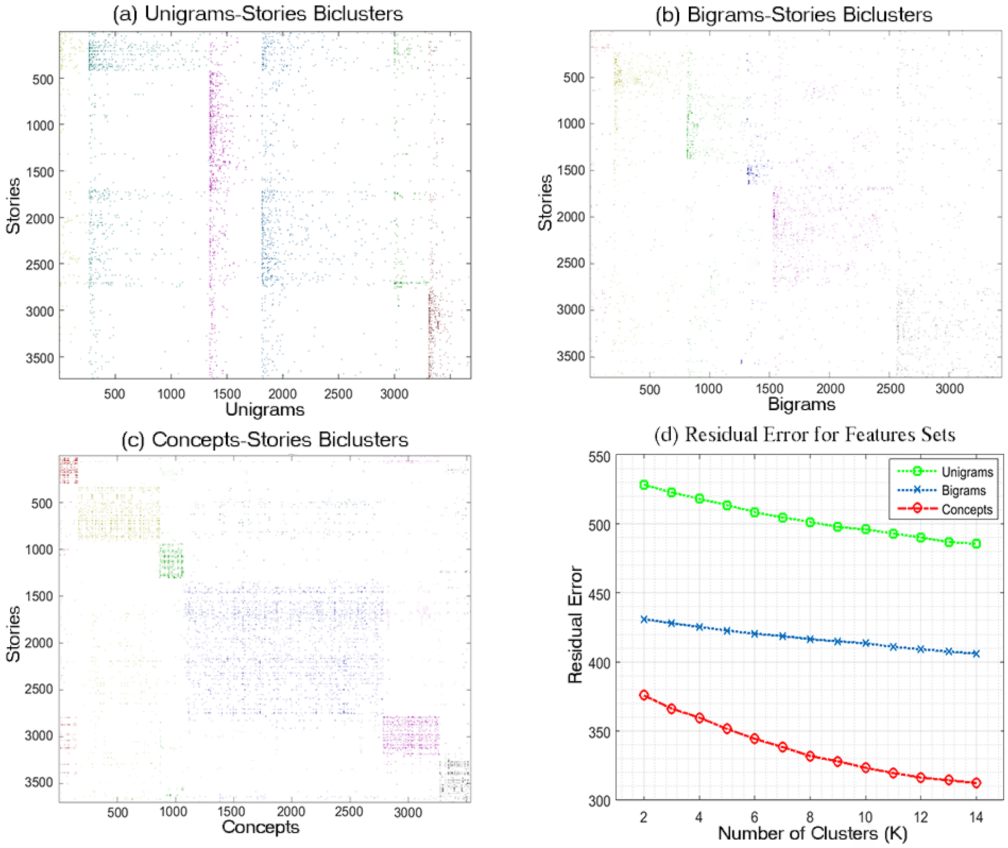

Figure 4. Feature matrices after co-clustering for all feature sets - block diagonal substructure is more coherent for the concept stories biclusters as in (c)

Afghanistan attack US and Afghan forces with improvised explosives. The stories consistently refer to the Afghan forces as “puppets”.

5.4.2. Notable clusters. While the variations across the dataset are subtle (as noted above), the concept clustering method did usefully identify some meaningful clusters. Concept Cluster 2, for example, contains stories with a particular subject-verb construction as this example illustrates: “Another martyr operation was carried out by Mujahid Abdul Wali, who carried out the attacks on military bases of puppet Afghan soldiers that were still in the same district with the first martyr operations.” This construction contributes to a semi-objective news-narrative, belying the propagandistic content. This format contributes to the positioning of the mujahidin as champions of Islam defending the *Ummah*- the Islamic Community, and also conveys that they are winning the war [24].

Concept Cluster 1 exhibits another very consistent formulation that contributes to strategic communication goals. Stories in this cluster are dominated by the verb construction of attack-result in which the stories describe an attack and include the results such as the example: “At 11:45 [08:45 GMT] on 5 August, one of our combat groups detonated a guided explosive device against a patrol belonging to the Crusader occupation forces on Kirkuk-Al-Riyad Road in western Kirkuk. The explosion resulted in destroying a specialized vehicle and killing or wounding all those on board. Praise be to God, the Lord of all creation.” Like Concept Cluster 2’s dispassionate, news-style reporting of operations, this cluster’s emphasis on the successful results of the attacks convey the meaning that the insurgents are a strong force, a strong champion and are winning the conflict. Importantly, this meaning is contained in the semantic combination of *attack* and *result*, but these words are often separated by dependent clauses or in completely separate sentences. This association of *attack* and *result* would not be detected and clustered by bigram clustering techniques, giving concept-based features an edge as it is capable of extending the features into a higher semantic level.

Concept Cluster 1 has an additional significant feature: the antagonists are almost exclusively referred to by derogatory epithets (“apostates”, “pagan army”, “safavids”, “Crusaders”). This rhetorical technique dehumanizes these groups and assigns unsavory and immoral characteristics to them, thereby emphasizing the threat to the Ummah by their very existent. Violence by the community’s champion is therefore justified against these groups that threaten the community. Identifying the rhetorical techniques is the first step to defusing their inflammatory and radicalizing power, and thus a technique that can distill these constructions from a body of text data is valuable.

TABLE 1. BIGRAMS-BASED CLUSTERS

Cluster	Key narrative features	Notes and significance
1	Protagonists: often unspecified, mujahidin; Actions: attacking with emphasis on bombs and vehicles, Note: name of Taliban spokesman frequent; emphasis is Afghanistan;	Protagonists and antagonists are not consistent; stories contain repeated phrases; mention of spokesman name a distinguishing feature
2	Protagonists: lions of ansar; Antagonists: inconsistent names, locations action: wide range, with most frequent being detonation destroying vehicles, with praise and gratitude to God. Emphasis on date	No significant difference in action between cluster 1, 2 and 3
3	Actors: security detachment (presumed subject/attacker) Action: emphasis on attack/result, detonation, and killing result; strong emphasis on date	Only difference between cluster 2 and cluster 3 is frequent actor "security detachment"
4	Action: attacks in Afghanistan; against Actors: US, NATO, invaders and puppets, vehicle/military base	Wide variety of actions within general category of 'attack'; clear focus on Afghanistan;
5	Actors: Shield of,Islam Brigades, mujahidin, AQIM; Iraqi National Guard, Mahdi army, and police Action: detonate explosive	Similar to clusters 1-3, but with greater emphasis on claims of attacks, potentially indicating purpose/intent of story
6	Actors: US, ISIS, God, Bin Laden, Shabaab, mujahidin, messenger; Frequent invocations of god, god's grace and praise; action: frequent construction 'carried out' operation	No consistency to the locations or actors in this cluster; variety points of view (POV)

TABLE 2. CONCEPT-BASED CLUSTERS

Cluster	Key narrative features	Notes and significance
1	Protagonists either unspecified or "Lions of Islam"; frequent construction of attack-result; antagonists always labeled with epithets (apostates, pagans, safavids, crusaders)	Function: justify the threat to Muslims by 'others' who are not to be respected
2	Protagonists: Lions or mujahidin; actions: attacks carried out; news format	Function: legitimize the insurgent/extremist actions by formatting in a news report format; convey the extremists are winning the war and are champions of Islam
3	Protagonists: mujahidin, Shabaab, Lions; Antagonists: US forces, apostates, Federal Police; Action: detonation of IEDs, car bombs, landmines and other explosives; settings: Afghanistan, Iraq, Somalia	Similar to cluster 5, but with much more variation; Highlights the vulnerability of adversary forces and highlights effectiveness across Muslim regions
4	Significant variation in protagonists, antagonists, settings and actions; Minor emphasis on the killing of women and children (by US/allies)	Very loose cluster with no discernable patterns
5	Protagonist: mujahidin; antagonists: US and puppets; action: bomb blast, detonation	Very tight cluster of stories of IED attacks against US and Afghan forces (puppets) set primarily in Afghanistan
6	Protagonists: Lions of Ansar Islam; actions: plant or detonate bombs; antagonists: US, apostates, crusaders	Another very tight cluster, analogous to Cluster 5 but set in Iraq; illustrates geographically specific epithet

6. Conclusion

Narratives are systems of stories that construct meaning [25]. That meaning is constructed in part by the patterns of relationships created by the actors and actions that make up the constituent stories [26]. In order to analyze the narratives circulating within a discursive environment, the ability to distill stories from a larger corpus of information, and then cluster those stories into meaningful groups of story forms is necessary. The clustering method must account for patterns of relationships of actors and actions. In this paper, we show that the concept-based co-clustering method described here, with its attention to subjects and objects (actors) and verbs (actions) makes a step towards a robust method that meets this strategic communication analysis goal. Concept-based features yield a better clustering scheme compared to bigram features both quantitatively and qualitatively. The lower residual error in NMF achieved by concept-based features shows that concept-based features can capture more information than high level noisy outcomes compared to the other feature sets. Additionally, the content of the clusters produced by concept-based features presents a better semantic pattern in terms of strategic communication as opposed to the repetitive and scattered content of the bigram clusters as outlined by the SME.

Acknowledgment

This research was partially supported by DoD Minerva Research Initiative Grant N00014-15-1-2821.

References

[1] B. Ceran, N. Kedia, S. R. Corman, and H. Davulcu, "Story detection using generalized concepts and relations," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015, pp. 942–949.

[2] S. Alashri, S. Alzahrani, J.-Y. Tsai, S. R. Corman, and H. Davulcu, "Climate change frames detection and categorization based on generalized concepts," *International Journal of Semantic Computing*, vol. 10, no. 02, pp. 1–20, 2016.

[3] H. Zhao, A. Wee-Chung Liew, D. Z Wang, and H. Yan, "Biclustering analysis for pattern discovery: current techniques, comparative studies and applications," *Current Bioinformatics*, vol. 7, no. 1, pp. 43–55, 2012.

[4] I. S. Dhillon, “Co-clustering documents and words using bipartite spectral graph partitioning,” in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2001, pp. 269–274.

[5] C. Ding, T. Li, W. Peng, and H. Park, “Orthogonal nonnegative matrix t-factorizations for clustering,” in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2006, pp. 126–135.

[6] H. Kim and H. Park, “Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis,” *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.

[7] S. Kok and P. Domingos, “Extracting semantic networks from text via relational clustering,” in *Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I*, 2008, pp. 624–639.

[8] U. Kang, E. Papalexakis, A. Harpale, and C. Faloutsos, “Gigatensor: Scaling tensor analysis up by 100 times - algorithms and discoveries,” in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2012, pp. 316–324.

[9] L. Jing, J. Yun, J. Yu, and J. Huang, “High-order co-clustering text data on semantics-based representation model,” in *Advances in Knowledge Discovery and Data Mining*. Springer, 2011, pp. 171–182.

[10] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning, “A multi-pass sieve for coreference resolution,” in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 492–501.

[11] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky, “Stanford’s multi-pass sieve coreference resolution system at the conll-2011 shared task,” in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, 2011, pp. 28–34.

[12] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky, “Deterministic coreference resolution based on entity-centric, precision-ranked rules,” *Computational Linguistics*, vol. 39, no. 4, pp. 885–916, 2013.

[13] M. Recasens, M.-C. de Marneffe, and C. Potts, “The life and death of discourse entities: Identifying singleton mentions.” in *HLT-NAACL*, 2013, pp. 627–633.

[14] B. Ceran, R. Karad, A. Mandvekar, S. R. Corman, and H. Davulcu, “A semantic triplet based story classifier,” in *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2012, pp. 573–580.

[15] J. D. Choi, “Optimization of natural language processing components for robustness and scalability,” Ph.D. dissertation, University of Colorado Boulder, 2014.

[16] (2015) Alchemy api language features. AlchemyAPI, Inc. [Online]. Available: http://www.alchemyapi.com/products/alchemylanguage

[17] (2013) Everest triplet extraction. Next Century Corporation. [Online]. Available: https://github.com/NextCenturyCorporation/ EVEREST-TripletExtraction

[18] A. Fader, S. Soderland, and O. Etzioni, “Identifying relations for open information extraction,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1535–1545.

[19] S. Busygin, O. Prokopyev, and P. M. Pardalos, “Biclustering in data mining,” *Computers & Operations Research*, vol. 35, no. 9, pp. 2964–2987, 2008.

[20] Y. Li and A. Ngom, “The non-negative matrix factorization toolbox for biological data mining,” *Source code for biology and medicine*, vol. 8, no. 1, p. 1, 2013.

[21] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[22] T. Caliński and J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[23] C. Lundry, S. R. Corman, R. B. Furlow, and K. W. Errickson, “Cooking the books: Strategic inflation of casualty reports by extremists in the afghanistan conflict,” *Studies in Conflict & Terrorism*, vol. 35, no. 5, pp. 369–381, 2012.

[24] S. Corman, S. Ruston, and M. Fisk, “A pragmatic framework for studying extremists’ use of cultural narrative,” in *2nd International Conference on Cross-Cultural Decision Making: Focus 2012*, 2012, pp. 21–25.

[25] J. R. Halverson, S. R. Corman, and H. Goodall Jr, *Master narratives of Islamist extremism*. Palgrave Macmillan, 2011.

[26] S. W. Ruston, “More than just a story: Narrative insights into comprehension, ideology and decision-making,” in *Modeling sociocultural influences on decision making: Understanding conflict, enabling stability*, J. V. Cohn, S. Schatz, H. Freeman, and D. J. Y. Combs, Eds. CRC Press.