Research Notebook

Weekly Entry:

Literature Review on "Using Group Membership Markers for Group Identification"

1. Title and Citation

Title: Using Group Membership Markers for Group Identification
Authors: Jean Mark Gawron, Dipak Gupta, Kellen Stephens, Ming-Hsiang Tsou, Brian Spitzberg, and Li An
Conference: Sixth International AAAI Conference on Weblogs and Social Media
Year: 2012

2. Summary of the Paper

This paper presents the automated militancy **level-ranking method applied for web documents by militant groups**, including groups belonging to the type of the white supremacy groups. Following the authors, the membership indicators for groups, the "**us-them**" terms, can also serve for the ordering and classification of documents by their militancy level. In the experiment, the authors evaluate the efficiency of the three models for the classification: one applying small manually selected vocab for membership indicators, one applying larger vocab, and one applying **Mutual Information (MI)** chosen vocab. Results confirm the best-ranking is provided by the small us-them vocab combined using the **TFIDF weighting**, beating the larger vocab and the models applying the MI vocab. To the author's surprise, the best classification models were given by the **SVM classifiers,** while for the ranking the models were not good, emphasizing the fundamental difference for text analytics classification versus ranking.

3. Relevance to Our Topic: Community Text analysis and formation

This paper is very relevant to the subject matter for the analysis by our group for text **formation by groups**, for the work discusses the influence the identity characteristics for groups has on the type of language militant groups will use. In the work, the author discusses how linguistic features, ie "**us-them" linguistic features**, can serve good predictors for membership and militancy. This is very relevant for the subject matter for the study for formation by groups and communication by groups using text. This paper also discusses the value for the text analytics features choice, something very relevant for the study for the formation by groups and communication by groups using text.

4. Key Concepts and Methods

Group Identity Markers: **"Us-them" terminology** is the domain of the study, including terms and expressions separating the membership in the in-group (us) from the membership in the out-group (them). "us-them" terminology is one of the building blocks for sublanguage and is the **identifier for the group**.

- **Examples**: Terms "our white brothers" versus "ZOG" (Zionist Occupation Government, them).

Ranking vs. classification: Ranking is distinguished from classification by the paper where documents were ranked by militancy level versus labeled militant versus non-militant. The highest ranked method applied the **small word list combined with the TFIDF weighting,** while the highest ranked classifier applied the larger word list combined with the SVM training.

Feature Selection: The study emphasizes the importance of feature engineering, showing that hand-selected features (us-them vocabulary) outperformed features selected by Mutual Information for ranking tasks.

TFIDF Weighting: Term Frequency-Inverse Document Frequency (TFIDF) weighting technique contributed significantly towards the effectiveness of the ranking algorithm, mirroring the work in sentiment analysis where the use of TFIDF over binary features proved superior.

5. Strengths and Limitations

Strengths:

Application of the membership and militancy identifiers (us-they terminology is the **innovative method** for classification by membership and militancy. This work bridges social psychology (i.e., the group identity theory by Tajfel) and text analysis by machine learning from a **unique perspective.** This methodology has **potential applications** for the study of law enforcement, counterterrorism, and the social scientific study where militant groups must be identified and studied.

Limitations:

This analysis is carried out using a comparatively **small dataset** (22 sites), the generalizability of the conclusions potentially being compromised. This analysis only dealt with **the white racist groups, perhaps not all the linguistic features** for the different forms of groups. **Human judgments** were being applied for the evaluation of the ranking system, being arbitrary and possibly not very reliable.

6. Implications for Our Project

This paper introduces some reflections for the work of the author's group for the moment concerning text formation groups:

That the **us-them terminology** is successful implies that features chosen by hand from the terminology unique to the communities can work very strongly for text analytics applications. We can perhaps make the features for the communities under study similar. To make the differentiation clearer, the need for the right measurement for our project is demonstrated by the **differentiation between classification and ranking**. If the goal is the measurement of the strength of engagement or identity, the application of ranking is perhaps the best over classification. Based on the study's conclusions about the use of **TFIDF weighing**, this method can potentially be applied for our project, especially when analyzing the term frequency and the documents' relevance for the communities. **Integrating the combination of machine learning and social psychology** here is the template for applying theoretical models in work through text analytics

7. Questions and Future Directions

- How can the "us-them" paradigm apply when studying non-militant groups? We studied militant groups, but the concept of the in-group versus the out-group terminology can also apply when studying political, culture, and internet groups.
- What other linguistic features (other than "us-them" terminology) can signal membership? For example, common references, jargon, or slang can also signal membership.
- How can the technique scale up for bigger heterogenous data? That the work is contingent upon the small corpus is the sole doubt about the conclusions drawn applying for bigger heterogenous data sets.
- What role does sentiment analysis play in identifying community identity? The paper briefly mentions parallels with sentiment analysis, which could be explored further in the context of community text formation.

8. Conclusion

This paper is a good example of the potential for text analytics to explore the analysis of community identity and processes of groups. Focus on membership indicators for groups is one concrete way by which the groups' use of language can be researched for its capacity for establishing the definition of the group and delineating the group from others. But the study’s focus on militant groups and its appeal to judgments by individuals for measurement demonstrate some pitfalls for consideration when structuring our own study. Overall, this paper has helped improve my knowledge of the potential for linguistic patterns to signal community identity and has given some inspiration for the features identified for analysis and analysis for the work being undertaken.