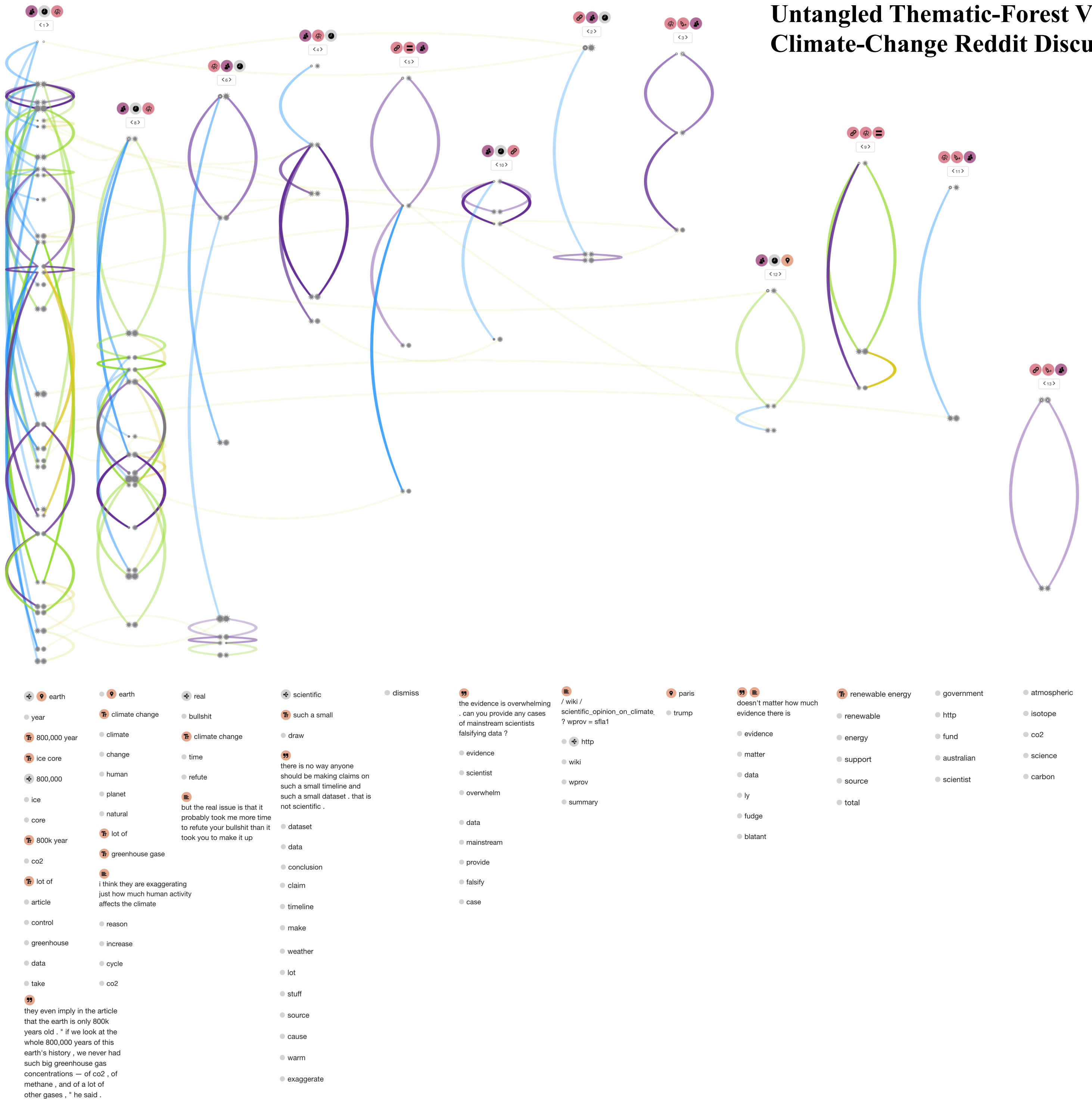
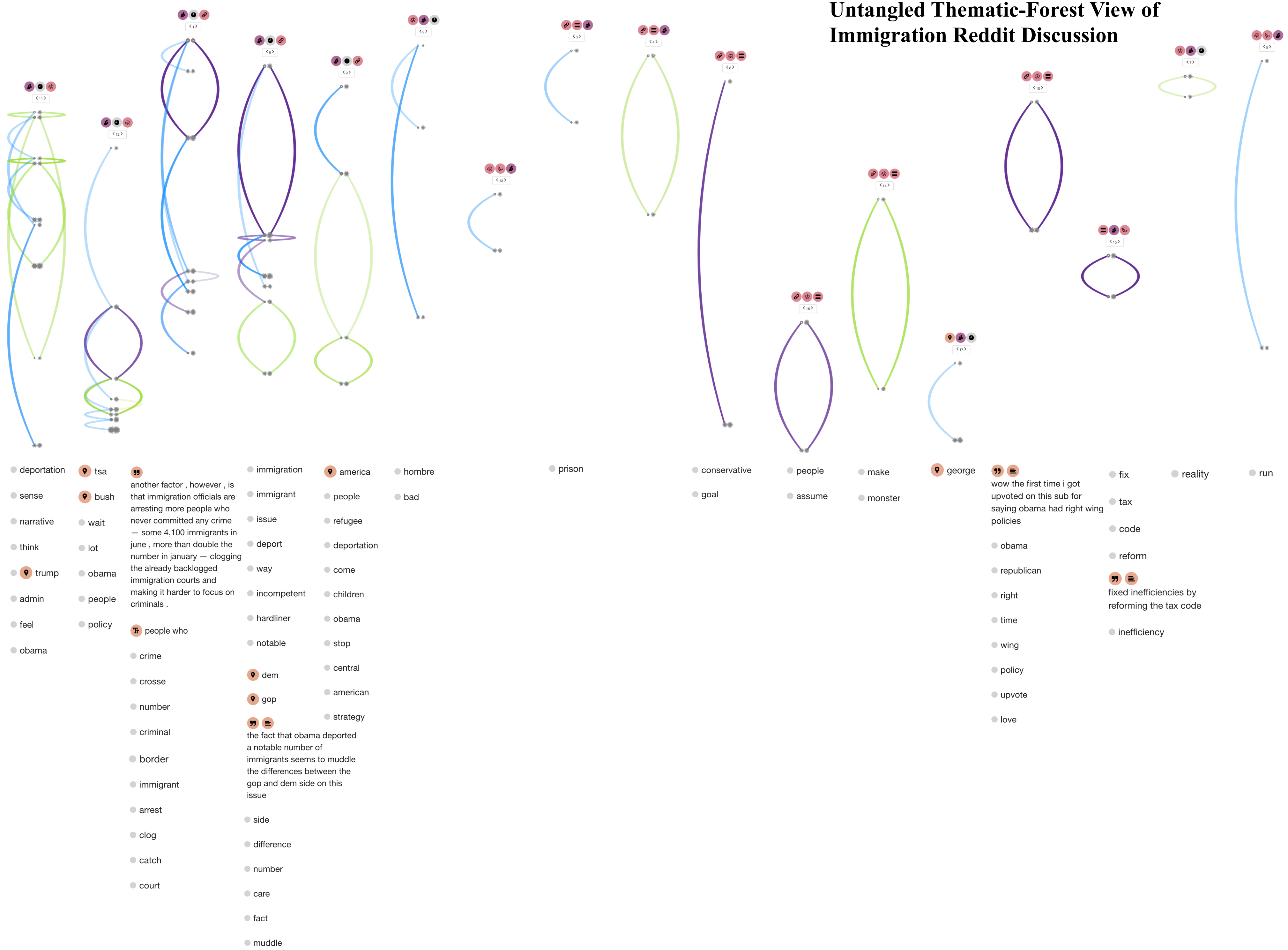
Untangled Thematic-Forest View of Climate-Change Reddit Discussion



Untangled Thematic-Forest View of Immigration Reddit Discussion



Model Results

Model	Precision	Recall	F-Score
Balanced Training Data (6926 Conversations)			
Decision Tree (5 features): 10-fold c.v.	0.77	0.69	**0.73**
Decision Tree (13 features): 10-fold c.v.	**0.79**	0.68	**0.73**
Random Forest (5 features): 10-fold c.v.	0.71	0.67	0.69
Random Forest (13 features): 10-fold c.v.	0.74	**0.70**	0.72
Imbalanced Real-World Data (40 Conversations)			
Decision Tree–5 (top 1 parent cand.)	0.14	0.14	0.14
Decision Tree–5 (top 10 parent cand.)	0.06	0.45	0.11
Decision Tree–13 (top 1 parent cand.)	0.16	0.16	0.16
Decision Tree–13 (top 10 parent cand.)	0.07	**0.46**	0.12
Random Forest–5 (top 1 parent cand.)	0.12	0.12	0.12
Random Forest–5 (top 10 parent cand.)	0.07	0.32	0.11
Random Forest–13 (top 1 parent cand.)	0.16	0.16	0.16
Random Forest–13 (top 10 parent cand.)	0.06	0.45	0.10
Precision Query (top 1 parent cand.)	**0.87**	0.04	0.08
Precision Query (top 10 parent cand.)	0.81	0.05	0.08
Recall Query (top 1 parent cand.)	0.27	0.28	**0.28**
Recall Query (top 10 parent cand.)	0.12	0.38	0.18
Content Query	0.36	0.29	0.32
Content Query (threads with 30 msgs)	0.56	0.48	0.51
Content Query (threads with 10 msgs)	0.70	0.66	0.68

Table 2: Summary of different model results.

Overview of the Related Work

Ref.	U/S*	Algorithm	Prec.	Rec.	F-sc.	Acc.	Characteristics
[9]	U	graph-based	-	-	0.7	-	**long messages** (avg > 60 words) **educational discussions**
[4]	U	SMSS	0.524	0.524	0.524	-	**long messages** (avg > 70 words) **manually annotated data**
[10]	U	similarity matching	-	0.8739	-	-	**reliable feature** (*quotes*) **e-mails** **short threads** (avg three e-mails)
[6]	S	Decision Tree	0.8307	0.6638	0.7379	-	**reliable feature** (only one feature: *reference to author's name*) **manually annotated data** **short threads** (4-comment threads)
[1]	**S**	**Decision Tree**	**0.939**	**0.918**	**0.928**	**-**	**reliable feature** (79.7% of the replies have a *distance* of 1) **balanced training dataset** **3-40 posts per thread**
[7]	S	Ranking SVM	-	0.9617	-	-	**reliable feature** (*quotes* as one of the main features) **e-mails** **short threads** (at least three e-mails)
[2]	S	SORTS: Ranking SVM + candidate filtering	0.5264	0.5264	0.5264	-	**long messages** (avg 63.4 words)
[3]	S	PPC + Ranking SVM	-	-	-	0.970	**e-mails** **short threads** (avg 6-12 e-mails)
[5]	S	threadCRF	-	-	-	0.635	**reliable feature** (*reference to author's name, person resolution*)
[8]	S	threadCRF	-	-	-	-	uses own set of metrics **short threads** (avg 6 messages)

Table 1: Summary of algorithms which are used to reconstruct the reply-relation structure. Listed are the **best evaluation results** of each paper, and the reasons, why these results could be achieved. The best performance for forum data is reached by [1], using Decision Tree algorithm. (* U-unsupervised, S-supervised, Prec.-precision, Rec.-recall, F-sc.-F-score, Acc.-accuracy)

References

[1] E. Aumayr and J. Chan. Reconstruction of Threaded Conversations in Online Discussion Forums. *Artificial Intelligence*, pages 26–33, 2011.

[2] A. Balali, H. Faili, M. Asadpour, and M. Dehghani. A supervised approach for reconstructing thread structure in comments on blogs and online news agencies. *Computacion y Sistemas*, 17(2):207–217, 2013.

[3] M. Dehghani, a. Shakery, M. Asadpour, and a. Koushkestani. A learning approach for email conversation thread reconstruction. *Journal of Information Science*, 39(6):846–863, 2013.

[4] C. Lin, J.-M. Yang, R. Cai, X.-J. Wang, and W. Wang. Simultaneously Modeling Semantics and Structure of Threaded Discussions: A Sparse Coding Approach and Its Applications. *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 131–138, 2009.

[5] Y. Liu, F. Chen, and Y. Chen. Learning thread reply structure on patient forums. *Proceedings of the 2013 international workshop on Data management & analytics for healthcare - DARE '13*, pages 1–4, 2013.

[6] A. Schuth, M. Marx, and M. de Rijke. Extracting the discussion structure in comments on news-articles. *Proceedings of the 9th annual ACM international workshop on Web information and data management*, pages 97–104, 2007.

[7] J. Seo, W. B. Croft, and D. a. Smith. Online community search using thread structure. *Conference on Information and Knowledge Management*, pages 1907–1910, 2009.

[8] L. Wang, M. Lui, S. N. Kim, J. Nivre, and T. Baldwin. Predicting thread discourse structure over technical web forums. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 13–25, 2011.

[9] Y.-C. Wang, M. Joshi, W. W. Cohen, and C. Rosé. Recovering Implicit Thread Structure in Newsgroup Style Conversations. *Artificial Intelligence*, pages 152–160, 2007.

[10] J.-Y. Yeh and A. Harnly. Email thread reassembly using similarity matching. *Third Conference on Email and Anti-Spam (CEAS)*, pages 64–71, 2006.