sourcing and supervised ensemble learning. The system aggregates different single-source annotators, each extracting topic labels from one part of the post (e.g., text, picture or video). We use crowdsourcing to evaluate how relevant topic labels are on a sample of Google+ posts. The crowdsourced judgments enable us to understand the varying reliability of the single-source annotators. We train an ensemble model on the data obtained from crowdsourcing process.

Evaluating on a gold standard data set, we find the ensemble model outperforms baseline method that naively combines topic labels from all annotators in classifying topic labels that are "Main or Important" topics. The ensemble model also significantly outperforms a baseline method in multiclass classification of topic labels into relevance categories.

Important user functions such as search and recommendation will benefit from better topic labels. By greatly improving the performance of how we apply topic labels to social media posts, it is our hope that users will enjoy more relevant and interesting posts.

7. ACKNOWLEDGEMENTS

We would like to thank Amazon Mechanical Turk workers for their participation in this study. We also thank Lichan Hong and Zhiyuan Cheng for thoughtful discussions and valuable feedback.

References

[1] O. Alonso, C. C. Marshall, and M. Najork. Are Some Tweets More Interesting Than Others? #HardQuestion. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval - HCIR '13*, pages 1–10, New York, New York, USA, Oct. 2013. ACM Press.

[2] H. Aradhye, G. Toderici, and J. Yagnik. Video2text: Learning to annotate video content. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pages 144–151, Washington, DC, USA, 2009. IEEE Computer Society.

[3] K. Bontcheva and D. Rout. Making sense of social media streams through semantics: a survey. *Semantic Web*.

[4] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–1002. IEEE, 2004.

[5] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in Twitter data with crowdsourcing. pages 80–88, 2010.

[6] A. Gattani, D. S. Lamba, N. Garera, M. Tiwari, X. Chai, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan. Entity extraction, linking, classification, and tagging for social media: A wikipedia-based approach. *Proc. VLDB Endow.*, 6(11):1126–1137, Aug. 2013.

[7] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126. ACM, 2003.

[8] G. Kazai, J. Kamps, and N. Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, 16(2):138–178, 2013.

[9] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[10] C. Lin and D. Weld. To Re (label), or Not To Re (label). 2014.

[11] J. Liu, R. Hu, M. Wang, Y. Wang, and E. Y. Chang. Web-scale image annotation. In *Proceedings of the 9th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, PCM '08, pages 663–674, Berlin, Heidelberg, 2008. Springer-Verlag.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. 2010.

[14] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[15] G. Shani and A. Gunawardana. Evaluating recommendation systems. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 257–297. Springer US, 2011.

[16] J. Vuurens, A. P. de Vries, and C. Eickhoff. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIRâĂŹ11)*, pages 21–26, 2011.

[17] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2764–2770. AAAI Press, 2011.

[18] S.-H. Yang, A. Kolcz, A. Schlaikjer, and P. Gupta. Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, pages 1907–1916, New York, New York, USA, Aug. 2014. ACM Press.