Searching for Structure in Unfalsifiable Claims

Peter Ebert Christensen,[1] Frederik Warburg,[2] Menglin Jia,[3] Serge Belongie[1]

[1] University of Copenhagen & Pioneer Centre for AI
[2] Technical University of Denmark
[3] Cornell University
pec@di.ku.dk, frwa@dtu.dk, mj493@cornell.edu, s.belongie@di.ku.dk

Abstract

Social media platforms give rise to an abundance of posts and comments on every topic imaginable. Many of these posts express opinions on various aspects of society, but their unfalsifiable nature makes them ill-suited to fact-checking pipelines. In this work, we aim to cluster such posts into a small set of core claims, that capture the essential claims related to a given topic. Understanding and visualizing these claims can facilitate more informed debates on social media. As a first step towards systematically identifying the underlying claims on social media, we introduce, PAPYER, a fine-grained dataset of online comments related to hygiene in public restrooms, which contains a multitude of unfalsifiable claims. We present a human-in-the-loop pipeline that uses a combination of machine and human kernels to discover the prevailing claims and show that this pipeline outperforms recent large transformer models and state-of-the-art unsupervised topic models.

Introduction

Social media platforms have changed the ways information is produced, disseminated, and consumed, creating new opportunities along with complex challenges. One of these challenges is how to grasp, use, and interpret a large corpus of text from online discussion.

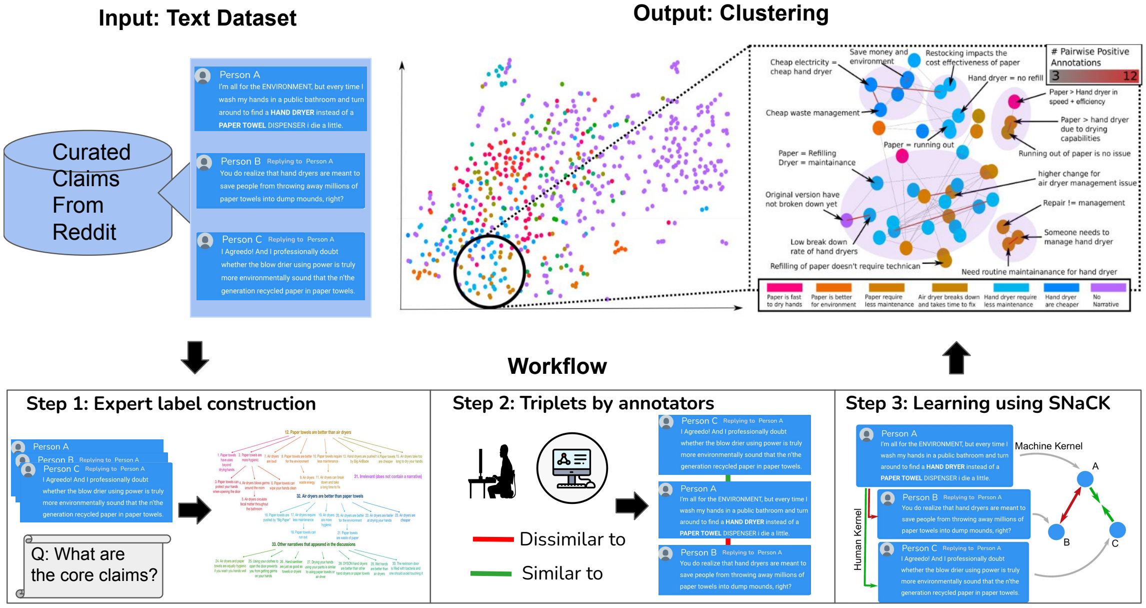
Several works (Blei, Ng, and Jordan 2003; Churchill and Singh 2021; Thompson and Mimno 2020; Moody 2016; Sia, Dalmia, and Mielke 2020) aim to distill large documents either through topic modeling or document summarisation. Our work falls into this category, however, we focus on identifying unfalsifiable claims in fine-grained topic-specific discussions. One of our long-term motivations to discover unfalsifiable claims is to complement fact-checking pipelines that currently do not have a use for unfalsifiable claims. Concretely, a statement such as "*Queen Elizabeth II was born in 1926.*" is considered falsifiable and hence check-worthy (Jaradat et al. 2018; Gencheva et al. 2017; Hassan et al. 2017), while "*The royal family is a waste of taxpayers' money.*" is an unfalsifiable claim. A common fact-checking pipeline would discard the latter type of claims being not check-worthy nor easily verifiable (Augenstein 2021). Our approach is not applied in any fact-checking pipeline, but can be seen as complementary, as annotators no longer need to consider the veracity of claims, and instead only think about underlying similarity in terms of views held by the individuals making the claims. Inspired by recent efforts for capturing human notions of similarity that remain elusive to state-of-the-art machine learning based representations (Agarwal et al. 2007; Tamuz et al. 2011; van der Maaten and Weinberger 2012), we introduce a new human-in-the-loop machine learning problem of social media claim discovery using SNaCK (Wilber et al. 2015). In doing so, we aim to discover a clusters of claims that can subsequently describe all facets of a debate. More specifically, this paper presents a crowdsourcing framework with three stages (Figure 1): (1) inferring core claims from comments (by experts), (2) triplet annotation for claim alignment (by annotators) and (3) Use the text and alignments to learn an embedding using SNaCK. We present a case study on a discussion topic related to hygiene and present PAPYER, a dataset containing claims related to the use of hand drying in public restrooms (i.e., *pap*er vs. air dr*yer*). Results show that our human-in-the-loop framework using SNaCK outperform fully automatic methods based on large Transformer models.

Crowdsourcing Workflow for Claim Discovery

Previous research have shown the success of reframing the complex task of human taste and intent into a similarity problem using triplets (Jia et al. 2021; Wilber et al. 2015). Following this spirit, we decompose this task into three interconnected steps as described below:

Stage 1: Inferring Core Claims from Comments The experts constructs a set of labels, structured as a tree, after inspecting the dataset. The tree highlights the granularity of claims, and every excerpt is assigned one of these labels.

Stage 2: Triplet Annotation for Claim Alignment We frame our problem as gathering members for a debate team and ask the annotators "who would be on the same side of a debate on this topic?" to uncover a latent claim space. The annotators are provided a grid UI (Wilber, Kwak, and Belongie 2014) containing 6 text snippets, 1 prompt and 5 options, and asked to select 2 out of 5 which would be on the same side of a debate on this topic. We provide a simple way to suggest text snippets: Embed the text excerpts using a pre


Input: Text Dataset
Curated Claims From Reddit
Output: Clustering
Cheap electricity = cheap hand dryer
Save money and environment
Restocking impacts the cost effectiveness of paper
# Pairwise Positive Annotations
3
12
Hand dryer = no refill
Paper > Hand dryer in speed + efficiency
Cheap waste management
Paper > hand dryer due to drying capabilities
Paper = running out
Running out of paper is no issue
Paper = Refilling
Dryer = maintainance
higher change for air dryer management issue
Original version have not broken down yet
Repair != management
Low break down rate of hand dryers
Someone needs to manage hand dryer
Refilling of paper doesn't require technican
Need routine maintainanance for hand dryer
Paper is fast to dry hands
Paper is better for environment
Paper require less maintenance
Air dryer breaks down and takes time to fix
Hand dryer require less maintenance
Hand dryer are cheaper
No Narrative
Workflow
Step 1: Expert label construction
Q: What are the core claims?
12. Paper towels are better than air dryers
31. Irrelevant (does not contain a narrative)
32. Air dryers are better than paper towels
33. Other narratives that appeared in the discussions
Step 2: Triplets by annotators
Dissimilar to
Similar to
Step 3: Learning using SNaCK
Machine Kernel
Human Kernel
A
B
C

Person A
I'm all for the ENVIRONMENT, but every time I wash my hands in a public bathroom and turn around to find a **HAND DRYER** instead of a **PAPER TOWEL** DISPENSER i die a little.

Person B Replying to Person A
You do realize that hand dryers are meant to save people from throwing away millions of paper towels into dump mounds, right?

Person C Replying to Person A
I Agreedo! And I professionally doubt whether the blow drier using power is truly more environmentally sound that the n'the generation recycled paper in paper towels.

Figure 1: A crowdsourcing workflow to discover and cluster unfalsifiable claims presented in text datasets.

trained transformer and sample the nearest 4 neighbours and 1 random. The text prompt is also drawn at random.

Stage 3: Learning from Text and Triplets After annotation, the suggestions are converted from a grid to triplets and we can learn to project the text excerpts to a 2D plane using SNaCK. We utilise a pretrained transformer to first encode the text after which SNaCK can be applied on this encoding and the triplets by jointly projecting the transformer embedding to 2D using *t*-SNE and uses the triplet constraints to group the similar text excerpts using the *t*-STE.

Case Study: PAPYER

Study Design As a proof of concept, we construct a new dataset that focuses on the topic of hand drying in public restrooms. As discussions on this topic largely center on the *pap*er vs. air dr*yer* debate, we name the dataset PAPYER.

We scrape Reddit for such comments, manually filter them and split them into short text excerpts (1-2 sentences). Based on the excerpts, we manually define 33 core claims across 4 supercategories: 15 pro-paper towel, 9 pro-air dryer, 8 other (related to hand drying), and 1 for irrelevant (not related to hand drying) and assign each except to one core claim. The core claims are illustrated as a tree structure in Figure 1. Additionally for this stage we made a pre-test consisting of 5 questions where annotators had to selecting similar claims from the same topic. We then hired 50 crowd workers from Amazon Mechanical Turk as annotators. Each worker annotated 12 HITS images and was paid approximately $12 for one hour of work. We benchmark this workflow against 7 baselines, consisting of LDA (Blei, Ng, and Jordan 2003) topic models using pre-trained embeddings networks (Sia, Dalmia, and Mielke 2020; Thompson and Mimno 2020) such as BERT and T5, and a mixture of these inspired from (Steve Shao 2022), all of which are projected using either *t*-SNE (Van Der Maaten 2014) or UMAP projections (McInnes, Healy, and Melville 2018). The expert labels are used to calculate the Triplet generalisation ratio (TGR) and KNN generalisation ratio (KNNGR) (Wilber, Kwak, and Belongie 2014).

Study Results Table 1 shows that the SNaCK and UMAP-T5 achieve the highest triplet generalization and k-NN ratio compared to the other baselines. From the KNNGR we find that the LDA baseline is unable to cluster the core claims, but in contrast, Figure 1 shows that incorporating human annotated triplets into the representation highlights interesting clusters that obey the core claims.

Method	TGR(↑)	KNNGR (↑)
t-SNE-BERT	55.33 ± 1.55	14.30 ± 2.63
t-SNE-T5	58.93 ± 2.28	31.05 ± 3.24
UMAP-BERT	54.39 ± 1.32	15.91 ± 2.71
UMAP-T5	61.44 ± 2.61	33.44 ± 4.25
t-SNE-LDA	53.34 ± 0.51	7.31 ± 1.42
t-SNE-BERT-LDA	54.01 ± 2.47	8.17 ± 3.01
t-SNE-T5-LDA	52.56 ± 1.14	9.56 ± 3.54
SNaCK-T5	67.61 ± 1.13	33.11 ± 3.07

Table 1: Discovery of prevailing claims. All models are evaluated 10 times using 1 trained model (ratios × 100).

Acknowledgements We thank Sagie Benaim, Stella Frank, and Vésteinn Snæbjarnarson for their comments on an earlier draft. This work was supported by the Pioneer Centre for AI, DNRF grant number P1.

References

Agarwal, S.; Wills, J.; Cayton, L.; Lanckriet, G.; Kriegman, D.; and Belongie, S. 2007. Generalized Non-metric Multidimensional Scaling. In Meila, M.; and Shen, X., eds., *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, 11–18. San Juan, Puerto Rico: PMLR.

Augenstein, I. 2021. Towards Explainable Fact Checking. arXiv:2108.10274.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3(null): 993–1022.

Churchill, R.; and Singh, L. 2021. The Evolution of Topic Modeling. *ACM Comput. Surv.* Just Accepted.

Gencheva, P.; Nakov, P.; Màrquez, L.; Barrón-Cedeño, A.; and Koychev, I. 2017. A Context-Aware Approach for Detecting Worth-Checking Claims in Political Debates. In Mitkov, R.; and Angelova, G., eds., *RANLP*, 267–276. INCOMA Ltd. ISBN 978-954-452-049-6.

Hassan, N.; Arslan, F.; Li, C.; and Tremayne, M. 2017. Toward Automated Fact-Checking: Detecting Check-Worthy Factual Claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, 1803–1812. New York, NY, USA: Association for Computing Machinery. ISBN 9781450348874.

Jaradat, I.; Gencheva, P.; Barrón-Cedeño, A.; Màrquez, L.; and Nakov, P. 2018. ClaimRank: Detecting Check-Worthy Claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 26–30. New Orleans, Louisiana: Association for Computational Linguistics.

Jia, M.; Wu, Z.; Reiter, A.; Cardie, C.; Belongie, S. J.; and Lim, S.-N. 2021. Intentonomy: a Dataset and Study towards Human Intent Understanding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12981–12991.

McInnes, L.; Healy, J.; and Melville, J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*.

Moody, C. E. 2016. Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. *ArXiv*, abs/1605.02019.

Sia, S.; Dalmia, A.; and Mielke, S. J. 2020. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1728–1736. Online: Association for Computational Linguistics.

Steve Shao. 2022. Contextual Topic Identification: Identifying meaningful topics for sparse Steam reviews. Medium. https://blog.insightdatascience.com/contextual-topic-identification-4291d256a032. Accessed: 2022-06-10.

Tamuz, O.; Liu, C.; Belongie, S.; Shamir, O.; and Kalai, A. T. 2011. Adaptively Learning the Crowd Kernel. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, 673–680. Madison, WI, USA: Omnipress. ISBN 9781450306195.

Thompson, L.; and Mimno, D. 2020. Topic Modeling with Contextualized Word Representation Clusters.

Van Der Maaten, L. 2014. Accelerating T-SNE Using Tree-Based Algorithms. *J. Mach. Learn. Res.*, 15(1): 3221–3245.

van der Maaten, L.; and Weinberger, K. 2012. Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, 1–6. ISBN 978-1-4673-1024-6.

Wilber, M.; Kwak, I.; and Belongie, S. 2014. Cost-Effective HITs for Relative Similarity Comparisons. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2: 227–233.

Wilber, M. J.; Kwak, I. S.; Kriegman, D. J.; and Belongie, S. J. 2015. Learning Concept Embeddings with Combined Human-Machine Expertise. *2015 IEEE International Conference on Computer Vision (ICCV)*, 981–989.