# **Data**

Our analysis bases on the Reddit discussion website (https://en.wikipedia.org/wiki/Reddit) and in particular on the analysis of two prominent subreddits in regards to smart home; namely, *r/smarthome* (https://www.reddit.com/r/smarthome/) and *r/homeautomation* (<https://www.reddit.com/r/smarthome/>).

The data extraction covers a period of 3 years from 2016-07-01 to 2019-06-30 (for details see Appendix 1). From the roughly 336 thousand comments and 38 thousand submissions, 5000 comments have been randomly selected from each of the subreddits (*r/smarthome* and *r/homeautomation*) for external validity and to guarantee an equal representation.

# **Data Preprocessing**

In order to add coherence to the highly intertwined conversations, which exists in online discourse (Abbasi et al., 2018), to each of the randomly selected comments the corresponding submission text and all the comments within the selected comment’s tree have been added together as one unique document, sorted by tier position and posting time (Figure 1).

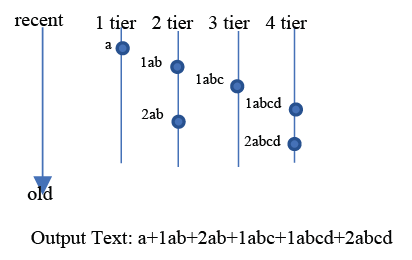


Figure 1: Document granularity

A base vocabulary was created by pre-processing documents using standard NLP tools, specifically: (1) removing comments from bots manually identified, (2) removing comments where 70% of the words were not in English, (3) running a spelling check to see if there were systematic errors, (4) dealing with stop words, lemmatization and stemming using the NLTK package, (5) coping with html formatting, internal hyphen and punctuation, (7) removing documents with less than 15 words considered too short, (8) dividing the data in 80% training and 20% testing, (9) training bigram models using SCIKIT-LEARN package, (10) filtering out words that occur less than in 5 documents, or more than in 40% of documents using GENSIM package.

# **Models**

## 1. LDA

LDA (Blei et al., 2003) is a generative model which assumes that the collection of documents (the ‘corpus’) is created based on the prior probabilities on the per-document topic distribution (alpha) and the prior probabilities on the per-topic word distribution (beta). A low alpha indicates that each document covers only few topics and a high alpha indicates that each document covers many topics. A low beta assumes that each topic consists of few words and a high beta assumes that each topic consists of many words. It is an unsupervised machine learning algorithm, since the only input provided by the researchers in the model estimation are the overall number of topics and the prior probabilities. LDA provides a powerful framework for representing and summarizing the contents of large document collections.

We run LDA topic modeling for all the combinations of numbers topics, that goes from 5 until 200 by step of 5, and the prior probabilities of alpha and beta in the values [0.01, 0.1, 1, 10]. The model with number of topics equal 20 and alpha and beta equal 1 was selected using the convergence of the following metrics (Figure 2): *coherence\_gensim\_c\_v* (Röder et al., 2015), *cao\_juan\_2009* (Cao et al., 2009), *arun\_2010* (Arun et al., 2010) and *coherence\_mimno\_2011* (Mimno et al., 2011)*.*

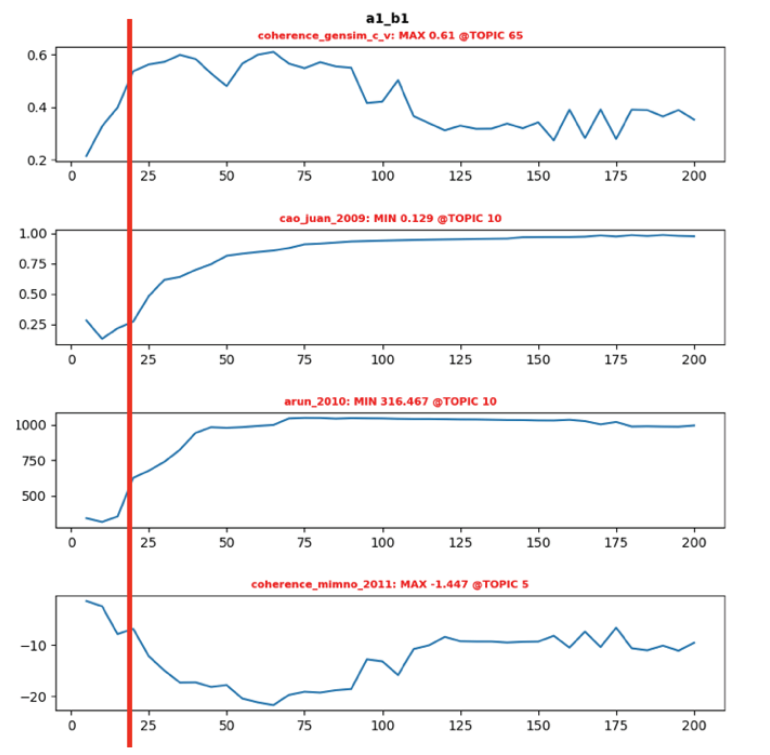


Figure 2: Metrics triangulation for alpha = 1 and beta = 1

The selected model was able to identify 9 macro topics of discussion within the two subreddits (Figure 3) (for details see Appendix 2).

Using the Jensen-Shannon distance (Endres & Schindelin, 2003), we classify the documents based on their closeness to the reference document for each of the inferred topics. The reference document represents the document with the highest percentage of the specific topic. In particular, we set a JSD <= 0.4.

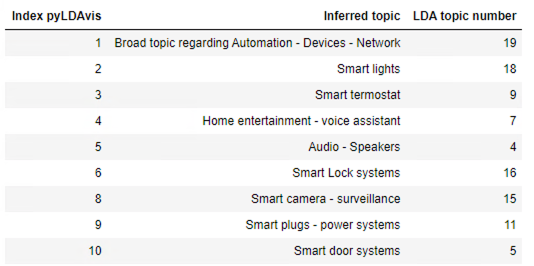


Figure 3: Inferred LDA topics

## 2. ELMo

ELMo (Peters et al., 2018) is a deep contextualized word representation that models both syntax and semantics of words taking into account the contexts where they are used.

We performed contextual vector clustering within the classified documents per topic based on the following rich query texts (Table 1). For each paper only the references part is taken out.

Table 1: Contextual vector clustering

|  |  |
| --- | --- |
| **Analyzed aspect** | **Rich query text** |
| Privacy | Kokolakis, S. (2017). Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. Computers and Security, 64, 122–134. https://doi.org/10.1016/j.cose.2015.07.002 |
| Awad, N. F., & Krishnan, M. S. (2016). The Personalization Privacy Paradox : An Empirical Evaluation Of Information Transparency And The Willingness To Be Profiled Online For Personalization. MIS Quarterly, 30(1), 13–28. https://www.jstor.org/stable/25148715 |
| Norberg, P. A., Horne, D. R., & Horne, D. A. (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. Journal of Consumer Affairs, 41(1), 100–126. https://doi.org/10.1111/j.1745-6606.2006.00070.x |
| Security | Chen, D., & Zhao, H. (2012). Data security and privacy protection issues in cloud computing. Proceedings - 2012 International Conference on Computer Science and Electronics Engineering, ICCSEE 2012, 1(973), 647–651. https://doi.org/10.1109/ICCSEE.2012.193 |

# **Results**

As shown in Figure 4 and Figure 5, the cosine similarity (Satya & Murthy, 2012) for the ELMo vectors of each document within a topic against the ELMo vectors of the privacy and security’s rich query text, seems to work as expected but there's little signal outside of the Topic 19 which includes general discussion of network setup.

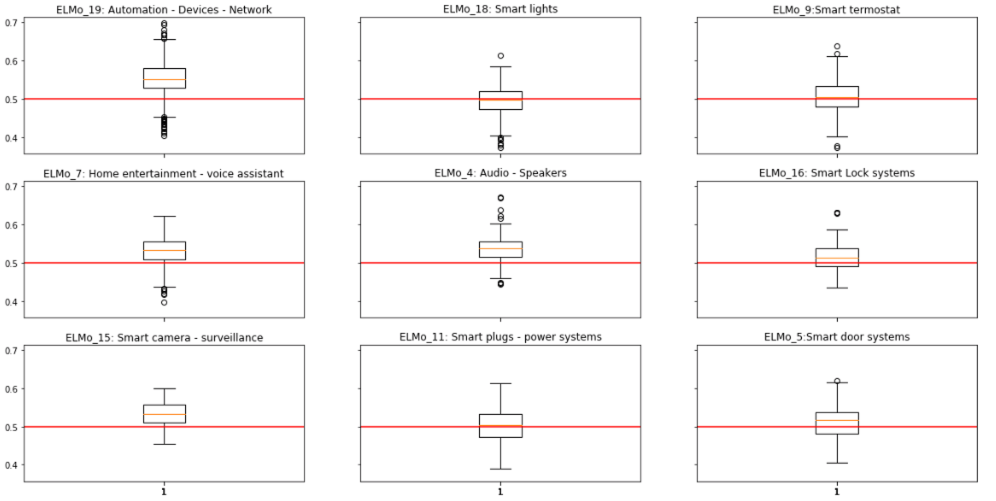


Figure 4: ELMo cosine similarity distribution of privacy within the topics

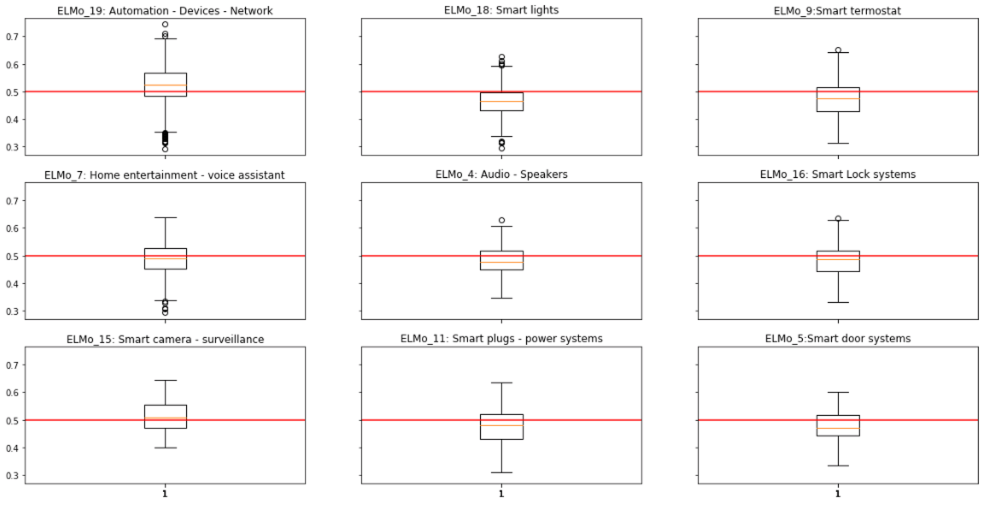


Figure 5: ELMo cosine similarity distribution of security within the topics

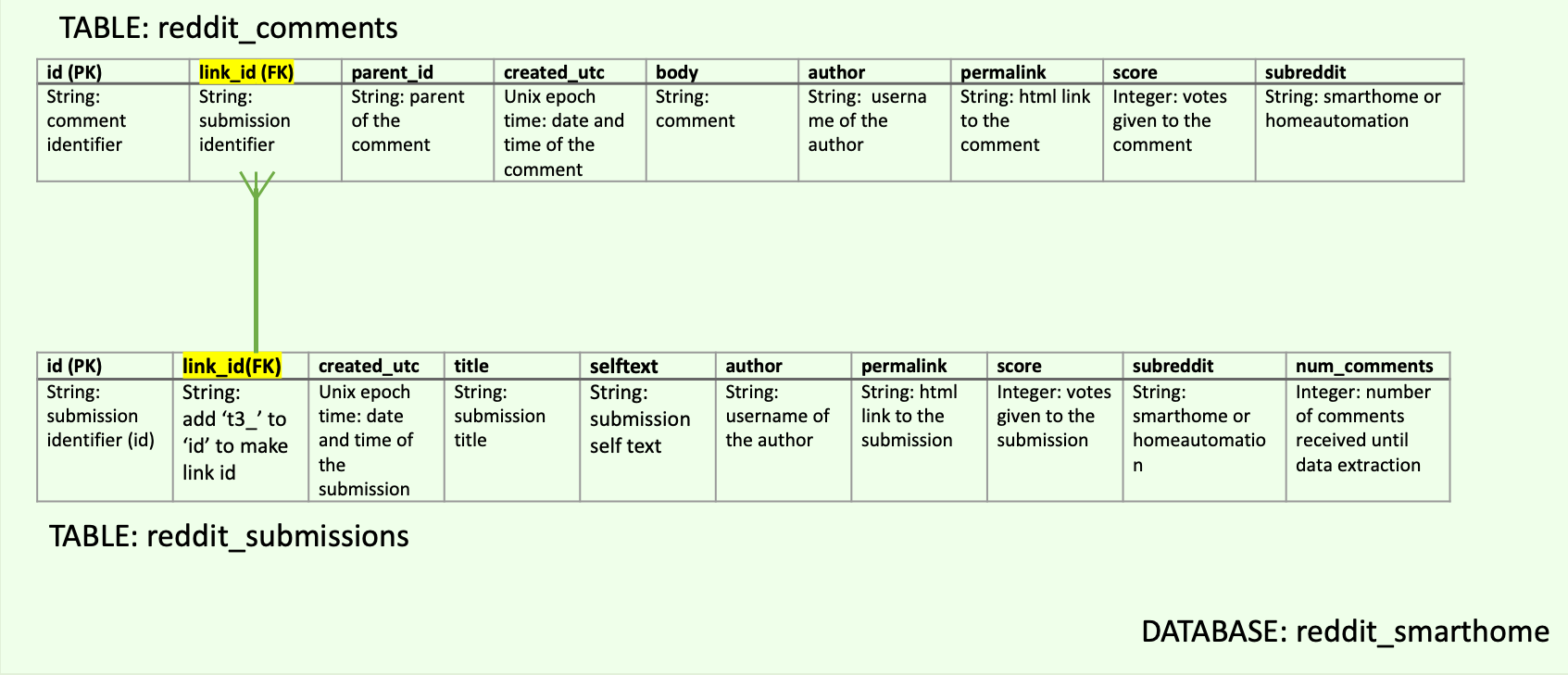
To the extent that this study is representative of the real world, an interesting observation is that the smart home community has at least a couple different subgroups. There are people whose interest in networked home tech builds on prior (possibly professional) expertise in setting up and managing networks. So they bring a lot of knowledge about network (in)security. Then there are hobbyists who are interested in the various smarthome products. Privacy and security at the network level are not a big priority with the latter group; privacy occasionally comes up in relation to voice assistants like Alexa (topics 4, 7), but - outside topic 19 - security is mostly used in sense of home security (e.g. topic 15, surveillance) (for details see Appendix 3).

For completeness, the model it's not picking up more security signal in the smart locks and doors topics, although those would also appeal to home-security types. But then again, those are pretty small topics.

# **Appendix**

## 1. Download procedure of Reddit data

1. Server repository: <https://files.pushshift.io/reddit/>
2. Enter **submissions** directory
3. Download data from 2016-07-01 to 2019-06-30
4. decompress files in csv files: SmartHome/decompress\_files/submissions/**read\_xz\_bz2\_zst.py**
5. Create a table in the database with the newly downloaded files (**MySQL/CreateTable.py**)
6. Insert data into the table (MySQL/**MySQLinsert.py**)
7. Add submission.link\_id = ‘t3\_’ + submission.id
8. Delete all orphan comments (comments without a submission)
9. Add foreign key to comment.link\_id



## 2. Model Inspection

NLP\_inspection.html

## 3. ELMo visualization

ELMo\_privacy\_viz.html

ELMo\_security\_viz.html

# **Reference**

Abbasi, A., Zhou, Y., Deng, S., & Zhang, P. (2018). Text analytics to support sense-making in social media: A language-action perspective. *MIS Quarterly: Management Information Systems*, *42*(2), 427–464. https://doi.org/10.25300/MISQ/2018/13239

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. *In: Zaki M.J., Yu J.X., Ravindran B., Pudi V. (Eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2010. Lecture Notes in Computer Science*, *vol 6118*. https://doi.org/https://doi-org.ez.statsbiblioteket.dk:12048/10.1007/978-3-642-13657-3\_43

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Endres, D. M., & Schindelin, J. E. (2003). A New Metric for Probability Distributions. *IEEE TRANSACTIONS ON INFORMATION THEORY*, *49*(7), 1858–1860. https://doi.org/doi:10.1109/TIT.2003.813506

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, *2*, 262–272.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 2227–2237. https://doi.org/10.18653/v1/n18-1202

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Satya, K. P. N. V, & Murthy, J. V. R. (2012). Clustering based on cosine similarity measure. *International Journal of Engineering Science & Advanced Technology*, *3*, 508–512.