Misinformation Detection in Reddit Data

Jakob Leander Müller
Philipps-Universität Marburg
muelle5t@students.uni-marburg.de

Abstract

This document summaries different analysis approaches on the FACTOID dataset (Sakketou et al., 2022) with the goal of understanding the underling structure. We will further discuss different models and their effectiveness in misinformation classification on a post level as well as proposals regarding better labeling techniques. Finally, my proposals include an improved way of labeling users as well as insights gained trough manual labeling.

1 Introduction

Fake News Detection is as difficult as it is important. Given the structure of noisy social media text and often used constructs like irony even human readers face difficulties distinguishing real news from misinformation. This report will discuss the structure and possible flaws in the FACTOID dataset, methods to train models for misinformation detection and propose improvements to the labeling process.

Current datasets and researchers classify misinformation spreaders based on the amount of misinformation links users have posted. Usually a threshold of 2 or 3 is chosen and if some user surpasses this threshold they are considered to be a misinformation spreader. To classify a post a set of known domains is used. A user would post some text including a link and if this link is considered to be directed to a domain spreading misinformation the post is classified as misinformation. This report will discuss challenges regarding the described approach and propose a different way of assigning user labels with the result that the current user labels are not considering major factors such as the number of total posts or posts including more than one link.

A strong misinformation detecting model could heavily reduce the spread of such information as well as the emergence of echo chambers. Social Media platforms such as Instagram introduced automatic warning banners related to potential misinformation about COVID-19 . These banners show up on posts related to COVID-19 and are supposed *"to keeping people safe and informed about coronavirus (COVID-19), COVID-19 vaccines and reducing the spread of false information"*[1]. FACTOID is a user centered dataset which raises the question of user labeling as well as post labeling. The report will cover statistics and models in both regards.

With the thought of extending or building another dataset we would like to understand if the used gathering approach of the FACTOID dataset yielded promising results. A split of 50:50 in real news and misinformation users was enforced during scraping using pre-defined thresholds.

2 Dataset

My work is an in depth look at the FACTOID dataset (Sakketou et al., 2022). The authors provide a dataset of around 4.000 Reddit Users and 3.4mil Posts from these Users.

If a user includes a link inside their post, that link will be used to label the post. Mediabiasfactcheck[2] is the external archive that provides a factual factor and a political bias. The post will not be assigned a label if the link is not included in the mediabiasfactcheck archive or if it has no link at all. The user will be assigned a label based on their post labels. The original method labeled a user as misinformation spreader if they had any post with a factuality of mixed, low, very low or a post with a political bias that is not between left- and right-center. The motivation behind this being that domains that publish with a strong

[1] https://help.instagram.com/2346065712360/
[2] https://mediabiasfactcheck.com/

political bias (no matter if left or right) or domains that do not provide factual news are probably spreading misinformation. This method will yield an almost perfect fifty-fifty split of real-news and misinformation spreading users. This method leaves some open questions like is there a minimal number of labeled posts a user should have or what happens if multiple links are in one post.

In the dataset section we will first go over the scraping process (Section 2.1) where the used topic groups are introduced, and some ground distributions are shown. After that we will explain what kind of additional data was used for the analysis (Section 2.2). The initial labeling process faces the issue that there are roughly 700 users with less than four labeled posts. I have worked on better ways to deal with users that have a small number of labeled posts in subsection 4.2.

2.1 Data Scraping

A focus on seven topic groups was enforced during scraping. This is implemented by handpicking subreddits that are associated with some controversial topic. Table 1 shows the topics picked by Sakketou et al. as well as the subreddits for each topic along with their complete distributions.

For each subreddit the most popular posts where collected and all participated users of these post which shared at least one post in their history containing some link where collected. Then a 50:50 split was enforced and after that the process was repeated with the posts that the extracted users have been active in.

2.2 Additional Data

In addition to the FACTOID Dataset I used the comments of posts for some analysis[3]. The additional data includes all comments that were published and have a labeled post as their direct parent. Figure 1 shows the distribution in number of comments for labeled posts. We can observe that most posts have only a small number of comments. Furthermore, comparing misinformation and real news we see that while misinformation post tend to receive more comments with over 5.000 characters this difference is so small (notice y-Axis is log scaled) that we can say misinformation and real news post are equal in regard to their mean comment length.

[3]Thank you Ezzeddine Ben hadj yahya for providing the data.

Subreddit	#unlabeled	#real	#fake
General political debate			
r/politics (Unbiased)	2.399.254	81.261	3.869
r/Conservative (Right)	346.042	5.165	2.784
r/JoeBiden (Left)	57.810	1.521	26
r/conservatives (Right)	24.310	526	453
r/Republican (Right)	17.797	500	256
r/democrats (Left)	11.747	338	41
r/ConservativesOnly (Right)	9.431	57	62
r/LockdownCriticalLeft (Left)	7.116	135	39
r/uspolitics (Unbiased)	6.063	356	17
r/Liberal (Left)	4.026	343	32
r/Impeach_Trump (Left)	3.122	117	4
r/RepublicanValues (Right)	1.093	18	0
r/LeftistsForMen (Left)	21	0	2
SARS-CoV-2			
r/Coronavirus (Unbiased)	92.163	2.753	54
r/NoNewNormal (Anti)	72.411	1.941	1.387
r/LockdownSkepticism (Unbiased)	62.480	1.441	275
r/CoronavirusUK (Unbiased)	8.030	167	4
r/CoronavirusUS (Unbiased)	4.028	141	4
r/COVID19 (Unbiased)	3.771	100	2
r/Masks4All (Pro)	2.140	35	9
r/NoLockdownsNoMasks (Anti)	1.887	82	61
r/EndTheLockdowns (Anti)	1.226	24	26
r/COVID19positive (Unbiased)	892	23	0
r/CoronavirusRecession (Unbiased)	756	35	0
r/CoronavirusCanada (Unbiased)	600	75	8
r/CovidVaccinated (Pro)	254	6	0
Women's and men's rights			
r/MensRights (Man)	57.654	1.636	501
r/antifeminists (Man)	1.138	44	15
r/feminisms (Woman)	654	15	3
r/MRActivism (Man)	410	29	7
r/Feminism (Woman)	238	2	1
r/Egalitarianism (General)	83	4	42
r/feminismformen (Man)	67	1	0
r/masculism (Man)	18	0	0
r/RadicalFeminism (Woman)	9	0	0
r/GenderCritical (General)	2	0	0
r/RadicalFeminismUSA (Woman)	1	0	0
Climate change			
r/climateskeptics (Questioning)	38.606	756	856
r/climatechange (Science)	7.858	622	153
r/climate (Science)	120	12	0
r/GlobalClimateChange (Science)	26	2	0
Abortion			
r/Abortiondebate (Unbiased)	7.590	84	22
r/prolife (Anti)	7.109	167	82
r/prochoice (Pro)	4.801	80	4
r/insaneprolife (Pro)	348	3	0
r/abortion (Unbiased)	60	0	0
r/AskProchoice (Pro)	18	1	0
r/ProLifeLibertarians (Anti)	1	0	0
Guns			
r/Firearms (Pro)	12.728	200	33
r/progun (Pro)	10.774	453	61
r/liberalgunowners (Pro)	9.719	336	6
r/GunsAreCool (Pro)	4.930	233	27
r/gunpolitics (Unbiased)	1.967	61	11
r/guncontrol (Anti)	1.062	206	10
r/GunResearch (Unbiased)	24	2	0
r/GunDebates (Unbiased)	1	0	0
5G			
r/5GDebate (Unbiased)	2.192	19	6
Other/Noise			
r/offmychest	1	0	0

Table 1: This table shows the names of the subreddits that belong to each topic and the corresponding number of unlabeled, real and fake news posts.

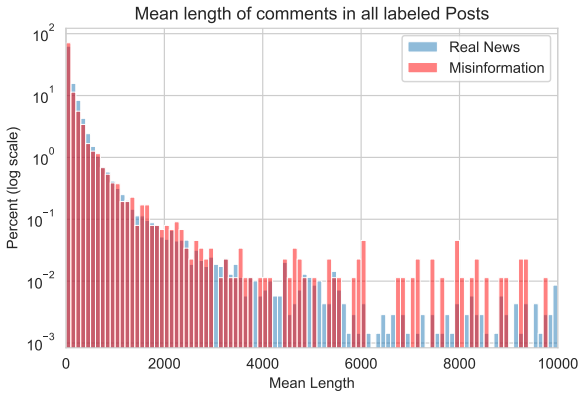

Figure 1: Distribution of number of first level comments in all labeled posts.

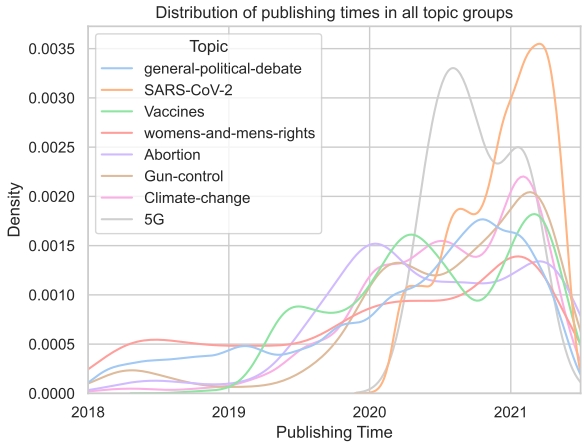

Figure 2: Distribution of post publishing times in topic groups.

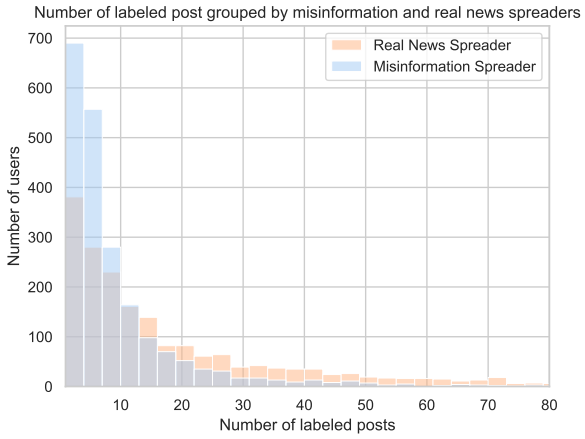

Figure 3: Distribution of annotations per user.

3 Data Analysis

To extract knowledge from the dataset different analysis methods were used. In section 3.1 we will go over statistical features present in the data set. The following subsections 3.2 - 3.5 cover different aspects of the data. Focusing on linguistic features, automatic topic detection, understanding how the hand-picked topics influence the dataset and lastly creating models using emotional features.

3.1 Data Distributions

The dataset has temporal focus on the 2020 US elections and contains most posts origination from that time. Due to the user-centric and connected post history scraping of users the total time span covered is much grater. Figure 2 shows a moving average of post publishing times using a topic group split. While the x-Axis is cut at 2018 there are posts as early as 2014 in the dataset. From these plots we can observe the time span and see the clear peak around the focus time of the US elections. Further we observe that in general all topics in the dataset have a similar distribution in time. It can be noted that *SARS-CoV-2* has its firsts posts in early 2020 which is very reasonable considering the start of the pandemic was around that time. This also implies suitable subreddits were chosen by the authors for this topic.

The number of annotated post per user varies a lot. From users with a single annotated post to users with over 1.000 annotated posts. 14.8% of all users have less than 3 annotated posts. Figure 3 also shows the high proportion of misinformation spreaders with a small number of labeled posts.

In Table 1 it is already apparent that some topic groups are heavily imbalanced. Figure 4 shows the distribution inside the topic group 'Guns'. We can see that over 90% of post originate from pro-gun communities. Further we can spot that the subreddit 'r/liberalgunowners' consists of mostly real-news where posts from 'r/progun' have a higher percentage of misinformation. Other topic groups e.g. 'Abortion' show a different distribution. Said topic is an example of a well-balanced group of subreddits.

3.2 Named Entity Recognition

As a first textual examination I performed Named Entity Recognition (NER) on all labeled posts. While we expected that there should not be major differences between real news and misinformation spreaders this process could be seen as an indicator that there is no critical fault within the annotated posts. Figure 5 displays the named entity densities

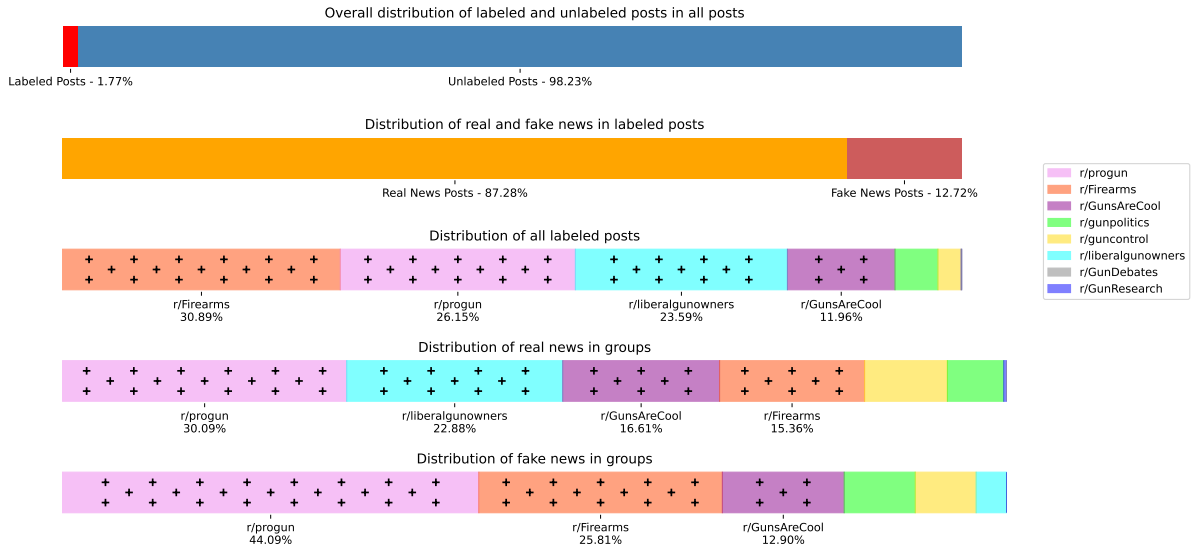

Figure 4: Distributions inside the topic group 'Guns'

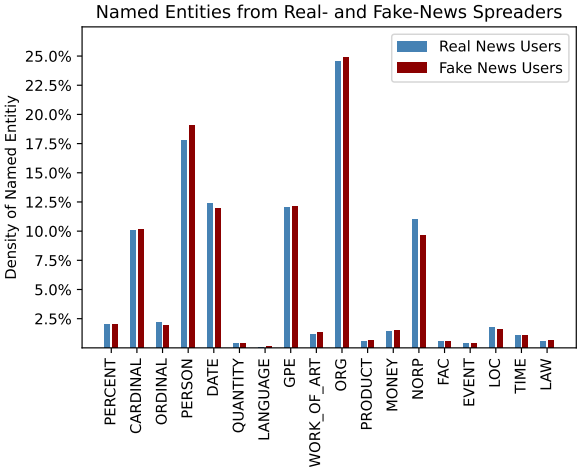

Figure 5: Named Entities in Real an Misinformation Posts

with have been extracted using spacy[4].

3.3 Latent Dirichlet Allocation

The Topic groups that were picked out by (Sakketou et al., 2022) may not be the topics that the users actually talk about. To analyze which topics are most relevant inside the data set I used Latent Dirichlet allocation (LDA). The computation was done with the python package Gensim (Řehůřek and Sojka, 2010). Looking at the results when calculating nine topics we can see that the topic with number 9 is a subset of topic 7. Due to this examination eight groups will be enough to understand the major topic groups. For eight topic groups Figure 6 shows the topic locations based on the two most important PCA features as well as wordclouds based on the words that LDA found.

Since most post originated from the group 'General political debate' we can also observe that the Topics 1, 2, 4 and 8 clearly resemble topics associated with this group. Smaller topics like '5G' or 'Climate Change' do not seem to be present. Topic 3 could be mapped to 'Guns' while Topic 6 and 7 seem to be related to 'SARS-CoV-2' and Topic 5 to 'Women's and men's rights'. Even though these mappings may not be perfect, it is clear that the data set consists of the topics that the authors handpicked.

3.4 Topics in Embeddings

We would like to understand how choosing different embedding approaches influences post clustering. We suspected that sBert embedding would be very sensitive to the topic discussed in a post. This turned out to be true and therefore implies that heavily imbalanced topic groups (see Section 3.1) will lead to a model learning the discussed topics rather than the structure of misinformation. Figure 7 shows the clusters of the different embeddings. The top plot displays an user2vec approach while the bottom plot uses a non fine tuned sBert embedding. Each circle represents one cluster and the distribution of topics in that cluster. The size of a circle is proportional to the cluster size. Lastly the inner circle is filled if the cluster consists of mostly misinformation and

[4]spaCy - Industrial-Strength Natural Language Processing https://spacy.io/

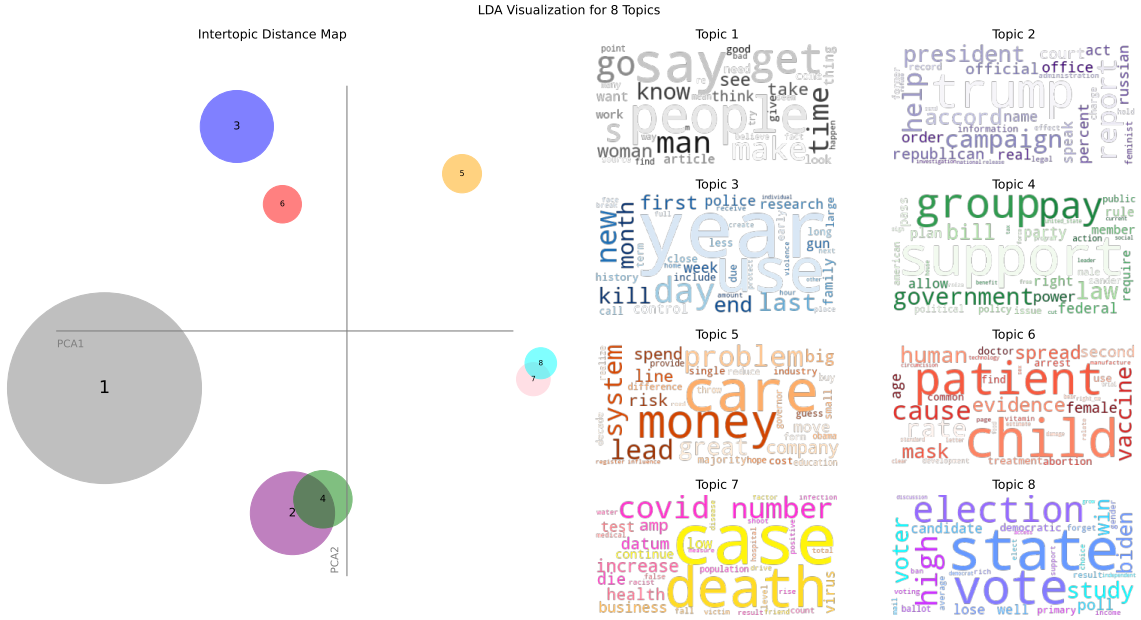

Figure 6: Topics found by LDA

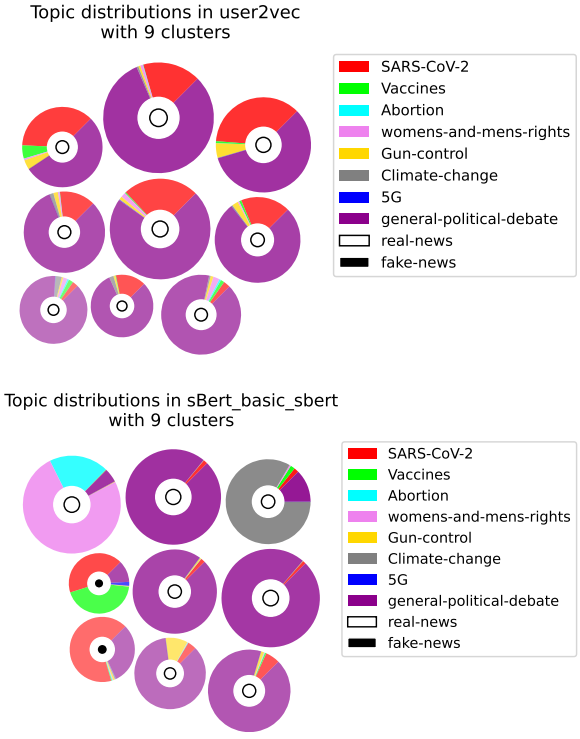

Figure 7: Topic clusters in different post embeddings. User2Vec (Top) and sBert fine-tuned to the FACTOID Dataset (bottom), both set to 9 clusters.

not filled else. The shade is proportional to the inter cluster similarity. Meaning the alpha value is equal to one if all posts are either misinformation or real news and 0.5 (minimum value set by me) if the cluster has a 50:50 split of real news and misinformation.

In the user2vec plot we can observe that all clusters carry posts from all topic groups. Further the proportion of topics is roughly the same in each cluster and equal to the overall proportion in the dataset. Contrary to that the sBert plot shows clusters with very different distributions. For example, we can clearly identify an Abortion/wonams-and-mens-rights cluster in the top left corner. Intuitively it makes sense that these two clusters probably have overlaps in discussed topics. We find a Climat-Change cluster in the top right. The left middle and bottom clusters are both mostly SARS-CoV-2 and Vaccines clusters. Considering the creation date of the dataset it again makes sense that these topics share a cluster. This observation highlights the importance for balanced topics if an sBert approach should be considered in the future. Similar results were also observed when using fine-tuned sBert modles to different datasets.

The cluster composition in terms of topic distribution is not the only interesting observation. We are interested in Misinformation Detection therefore we would like to know if the cluster have a good quality in that regard. Using data that was again provided by Ezzaddine I created visualizations for different embeddings. The Heatmaps are structured with the different embeddings on the y-

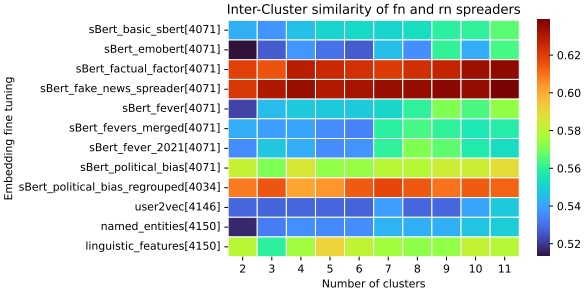

Figure 8: Heatmap of cluster qualities.

Axis and the number of clusters on the x-Axis. The number of clusters ranges from two to eleven. The quality of the clusters was judged based on there homogeneity regarding misinformation. The exact formula is

$$\frac{2 \cdot \max\{\#RN, \#MISINF\}}{\#cluster} - 1.$$

$\#RN$ is the number of real news users and $\#MISINF$ is the number of misinformation users. This ensures that every cluster is assigned a score between zero and one. Where zero would mean that the cluster carries the same number of misinformation and real-news spreaders. One means that the cluster is either consisting completely of misinformation or real news spreaders.

The clustering was performed using k-means. In general the sBert based clusters outperform other methods. Further we observe that embeddings on a user level increases when the embedding were fine-tuned on the post level. When fine-tuned on other dataset that tackle misinformation like the fever datasets (Thorne et al., 2018) we can also so an improvement when compared to a basic sBert model. Figure 8 displays the described plot.

3.5 Dual Emotion Models

When using sBert embeddings on a post level to classify rather a post is misinformation or real-news another student evaluated a F_1 score of around 50%. Putting this score into perspective is complicated. In subsection 4.1 we will take a look at some problems found while manually labeling posts of the data set. There I found that it is often a difficult, complex and time-consuming task to give a single label. If someone with more resources at hand would make a study of human performance in fake-news detection I would highly suspect that the participants will not majorly outperform a computer model. The task of experimenting with more complicated models still remains. (Zhang et al., 2021) proposed a model based on textual context as well as emotional features. Especially the relation between the emotions of the author and the users that commented. I used a slightly modified version of their model which can be seen in Figure 9. The extract of the textual features I used a small pre-trained sBert model (Bhargava et al., 2021), (Turc et al., 2019). The emotional features were extracted using NVIDIAs sentiment-discovery model [5]. I did not make any changes to the procedural of (Zhang et al., 2021) but provided code and documentation to apply it to custom data sets in an easier way.

Figure 10 shows the distribution of different emotions in the categories' publisher (the emotion within the labeled post), mean_com (the mean emotion in all comments), max_com (the maximum emotion with all comments), mean_gap (the mean absolute difference between publisher and comment) and max_gap (the maximum absolute difference in a given emotion between the post and any comment). The bars are further grouped into correctly and falsely classified posts. The plot was generated to analyse if there are significant differences in the emotionality of the user that posts something and the users that comment on posts. Further, including the spilt between correctly and incorrectly classified post by the trained model we would like to see if there is a bias within the model.

Most importantly we do not find a strong bias regarding emotions in any of the four categories of posts. We can also observe that there is no significant difference between publishers in any of the emotion categories. It appears that comments on real-news posts are more emotional than comments directed at misinformation posts. These differences are small, and an interpretation would require a deeper manual look into the assigned labels.

The model was trained in different settings to conduct an ablation analysis. The first setting uses the complete flow described in Figure 9. For the second run all textual features encoded by the sBert

[5] https://github.com/NVIDIA/sentiment-discovery

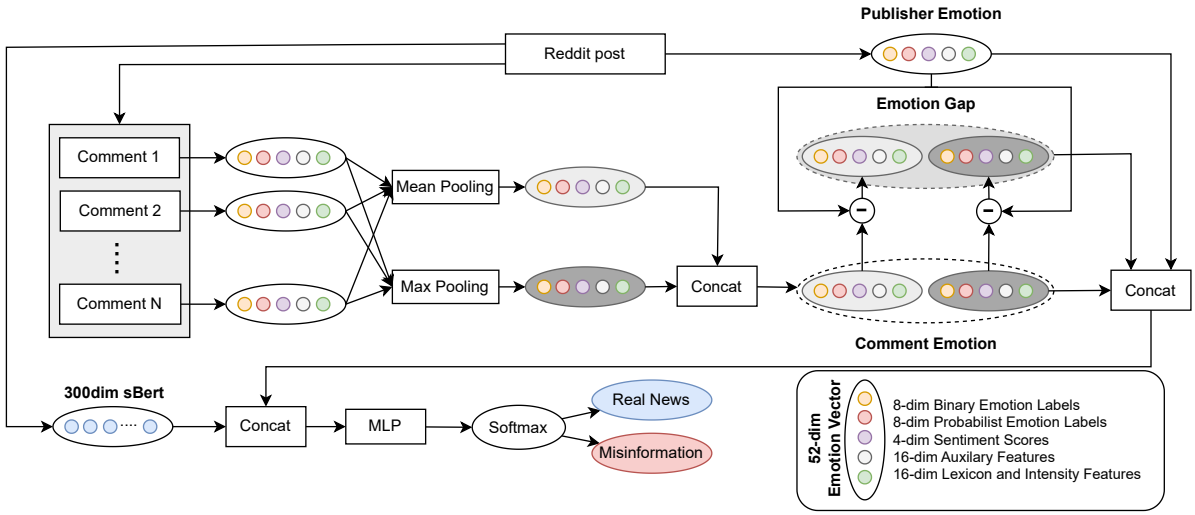

Figure 9: Structure of the Dual-Emotion Model used to classify misinformation posts.

Model	F1	Accuracy	Recall	Precision
Large sBert	53.05	**89.10**	**74.20**	53.30
Dual-Emotion	**63.41**	72.18	66.72	**57.88**
Random sBert	45.47	56.47	54.41	51.84
Random emotion	53.95	64.96	67.27	57.30

Table 2: Results of the ablation analysis.

model are replaced by random numbers. In the third run the dual-emotion features are replaced. We find that the model performs best when all features are real and second best when only using the textual features. Since we already discussed that there are no major differences in emotions this was expected. Table 2 show the results regarding F1, accuracy, precision and recall of the different runs. The training set was sub sampled to enforce a balanced training set which significantly increased the test results. Testing was always performed on the imbalanced dataset.

4 Proposals

4.1 Manual Labeling

During my work on the FACTOID data set I labeled a total of 500 post manually. Half of the posts were correctly classified by the best performing model and the other half were not. During this process we came to the conclusion that the automatic annotations are connected with a couple of issues. First, some posts do not contain enough context to understand what the user wants to express with the shared link. We decided to interpret this behavior as 'The user agrees with the contents of the shared link'. Secondly some users explicitly state that they do not agree with a link. While this raises questions on how to treat such behavior, in the end said user still shared the misinformation and raised attention for it instead of sharing real news which invalidates the misinformation. The most important observation though lies within the labeling of posts that contain more than one link. The original method considered any user that shared at least one misinformation link as misinformation spreader. This is especially problematic as there are users with 300+ real news and 1 misinformation link. And that one link could just be from a mixed source as far as labeling is executed. The gathered knowledge about these structures and connected problems caused the search for a better user labeling technique. In subsection 4.2 I will go over the changes made and the impacts they had.

4.2 Relabeling Users

To label a user based on their posts we will take into account all the links present in their posts. Using external work the domains of links can be labeled as fake- or real-news. Comparable approaches have used a strict cutoff of at least three fake-news-links to be considered a fake-news-spreader and a user can not have a single fake-news-link to be classified as a real-news-spreader. In the process of manual labeling we found that these thresholds are especially problematic for posts with more than one link (multi-link posts). We have seen examples of users referring to fake-news and using credible sources to invalidate the counterpart. This led to the decision that it is not sensible to strictly label a user as a fake-news spreader if they have posted a

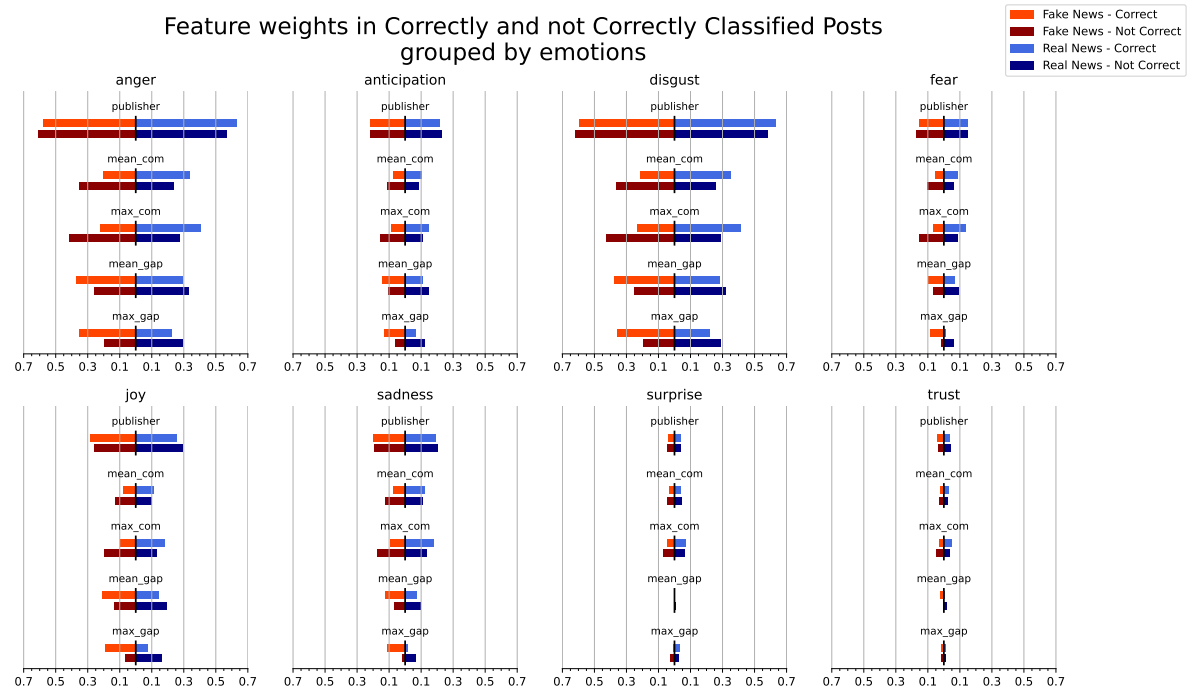

Figure 10: Distributions of emotional features

set number of links from uncredible domains. We now want to propose our thresholds which take the total number of posts into account.

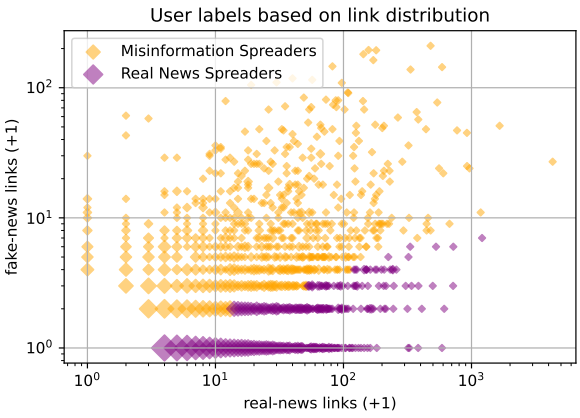

Figure 11: Label assignment when using the proposed user labeling. Diamond size scaled to the number of users at the given position.

For the relabeling task we decided to classify a user based on the link-label distribution rather than the post-label distribution. Before calculating the new labels I started by disregarding all users with less than three links. Users with too few posts do not provide significant context and therefore we see them as some kind of unwanted noise. Applying the filter drops 666 users from the dataset. The decision to use links directly removes the problem of multi-link posts.

Let us look at the link distributions on a user level. In Figure 11 each diamond represents one or more users. The position is based on the ratio between real-news and fake-news links. The size of the diamond is square root correlated with the amount of users that have the given distribution. Notice that both axes are log-scaled and to show values of zero all circles are shifted by (+1,+1). We can clearly see that most users have only a few posts and typically only real-news links. If we would cut off all users with at least one fake-news link that would yield a ratio of 48.68% real-news-spreaders and 51.32% fake-news-spreaders. This is a ratio close to the balanced version and something we want to see but it is not fair or justifiable. A User who posted 140 real-news links and 1 fake-news link will now be considered a fake-news spreader. This is not what we would intuitively expect when reading the numbers. I propose the multi level thresholds seen in Figure 12. When applying those thresholds we achieve a ratio of 59.13% real-news-spreaders and 40.87% fake-news-spreaders. See Figure 11 for the recoloured plot based on the new thresholds.

Conclusion

In conclusion we have observed some major biases contained in the current version of the FACTOID dataset. These biases range from uneven topic group sizes to biased topics. If a future version of this dataset is constructed there should not only be a consideration of balanced user labels but also balanced topics.

Through manual labeling and visualizations of automatic labels we have learned the importance of posts containing multiple links. Further we learned that there is large group of users with a small number of (labeled) post. These users have an impact on the dataset as a whole which we cannot argue to be sensible due to the lack of data. Again, in a future iteration a user should have at least some, for example 3 as used in similar datasets build on twitter data, labeled posts.

Through the experiments conducted using a dual-emotion model, we have learned that we are able to get similar results to (Zhang et al., 2021). We can see a clear improvement when comparing the model to a fine tuned sBert model. This result could be further improved by using more recent sentiment analysis models like NVIDIAS Megatron-LM [6].

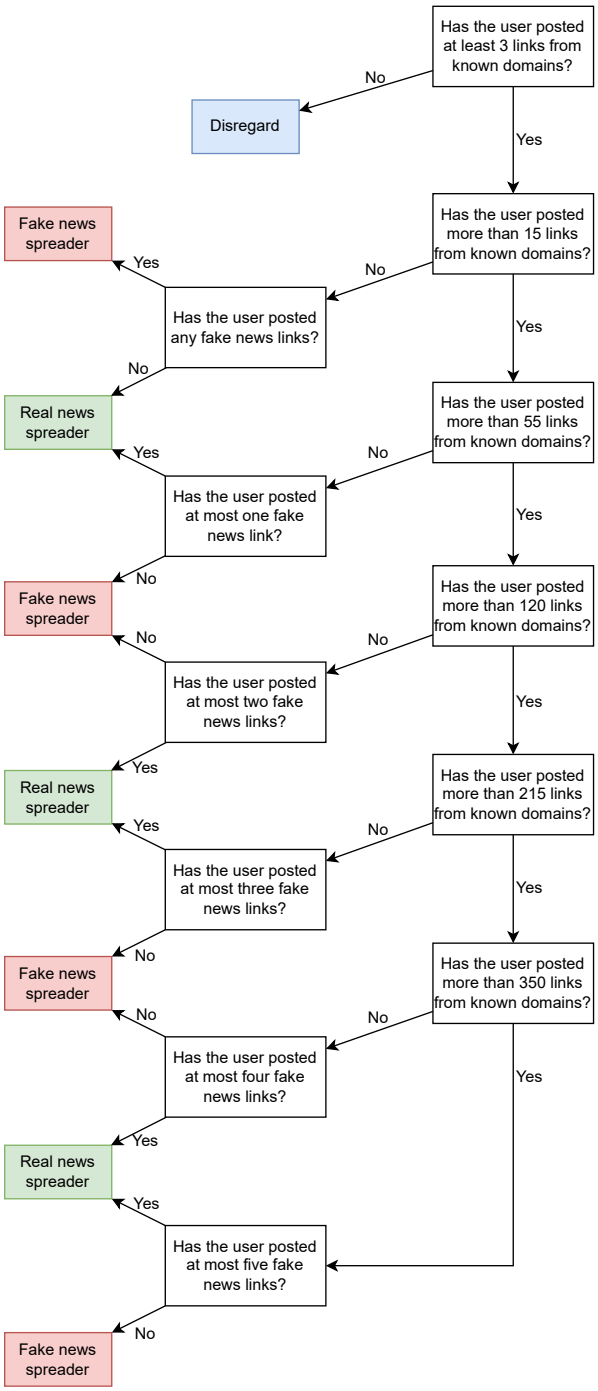

Figure 12: Proposed user labeling thresholds.

Acknowledgements

Many thanks to Flora Sakketou and Joan Plepi for the continuous support during this project and thank you to all the CAISA-Lab members that provided feedback and inspirations during the milestone presentations.

References

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in nli: Ways (not) to go beyond simple heuristics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Flora Sakketou, Joan Plepi, Riccardo Cervero, Henri-Jacques Geiss, Paolo Rosso, and Lucie Flek. 2022. Factoid: A new dataset for identifying misinformation spreaders and political bias.

[6] https://github.com/NVIDIA/Megatron-LM

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*. ACM.