Appendix:
Design and Evaluation of Deep Learning Models for Real-Time Credibility Assessment in Twitter

Marc-André Kaufhold[1], Markus Bayer[1], Daniel Hartung[1], and Christian Reuter[1]

Technical University of Darmstadt, Science and Technology for Peace and Security (PEASEC), Darmstadt, Germany
{kaufhold,bayer,reuter}@peasec.tu-darmstadt.de
daniel.hartung@student.tu-darmstadt.de

1 Appendix

1.1 Collection and Coding of the Twitter20 Dataset

To create the dataset, about 24,000,000 tweets were gathered in two stages of the COVID-19 pandemic (February 27, 2020 to March 6, 2020 and April 14, 2020 to April 22, 2020). The set was deduplicated and all non German posts resulting in 55,600 tweets for the early stage and 75,892 tweets for the second stage were removed. Since the ratio of misleading to non-misleading content of all tweets is very unbalanced, a rough first annotation pass over a sample was performed to identify potentially misleading content. In this way, a subsample for the in-depth analysis of each tweet was created. The coding was done by researchers with expertise in political science, misinformation, and HCI. They were familiarised with plausible methods of identifying misinformation, including fact-checking pages. The linal labels were chosen by a majority decision of the three independent researchers. After the annotation process, non-misleading tweets were randomyly excluded to get a perfectly balanced set. The final set consists of 2,382 tweets in total, from which 50% are misleading informations, such as false statements or misleading satire, and 50% are labelled as non-misleading, for example, non-absolute opinions or satire that is not misleading. The Twitter20 dataset will be made available in the scope of a different paper that is currently in review.

1.2 Automatic Coding Process

As described above, the FakeNewsNet dataset only contains one headline per topic that is labeled according to the credibility. However, in order to generate a considerable amount of training data, not just one single post per topic but several posts need a assigned credibility. This leads to the major challenge of not including posts that actually expose the topic as true or false. The following steps processed on the FakeNewsNet dataset should mitigate this problem:

1. **Topic filtering**: The dataset contains articles from Politifact and GossipCop. GossipCop's data mostly relates to "gossip"; thus, it is discarded.
2. **Temporal filtering**: The first 25 tweets are determined for each topic. The intuition of this pre-filtering is that the first contributions to a topic are probably still about the topic itself and do not yet expose it as right or wrong.
3. **Query-based filtering**: Through manual analysis, we identified certain keywords that indicate inappropriate tweets (e.g., false, fake, rumor, snopes, politifact, and gossipcop). Tweets that contain one of the keywords are discarded. In order to avoid excessive filtering, the filtered tweets were analyzed. Individual topics were then defined as exceptions and the associated contributions are no longer filtered.
4. **Similarity filtering**: The selected posts should be as different as possible in order to avoid redundancies in the training data. Therefore, we use the Levenshtein distance [**?**] to find similar posts that can be discarded. Only the first 80 percent of the characters of the posts were considered. With the help of the distance metric, all posts, up to a maximum of five, are selected at random whose distance to one another is greater than 30.

Table 1. Results of the quality analysis per model on the developments of both datasets.

	MLP		RNN		BERT	
Features	MSE	Acc.	MSE	Acc.	MSE	Acc.
Base	–	–	0.1340	0.7224	0.0669	0.8675
Tweet (Tw)	0.1424	0.6215	0.1247	72.56	0.0661	0.8644
User (Us)	0.1499	0.6215	0.1231	0.7161	0.0669	0.8675
Text (Tx)	0.1393	0.6404	0.1261	0.7224	0.0664	0.8675
Timeline (Ti)	0.1422	0.6656	0.1125	**0.7729**	0.0655	0.8675
Advanced Ti (ATi)	0.1441	0.6498	0.1146	0.7382	0.0666	0.8612
Tw & Us	0.1285	**0.7066**	0.1145	0.7445	0.0660	86.12
Tw, Us, & Tx	0.1227	**0.7066**	**0.1112**	0.7476	**0.0652**	**0.8707**
Tw, Us, Tx & Ti	0.1342	0.6877	0.1133	0.7697	0.0661	0.8612
Tw, Us, Tx & ATi	**0.1244**	0.7003	0.1133	0.7445	0.0656	0.8644
Base	–	–	0.1859	0.6801	0.1424	0.7778
Tweet (Tw)	0.1959	0.6202	0.1782	0.6954	0.1427	0.7778
User (Us)	0.1916	0.6245	0.1711	0.7088	0.1406	0.7759
Text (Tx)	0.1893	0.6494	0.1747	0.6877	0.1426	0.7778
Timeline (Ti)	0.1864	0.6446	0.1719	0.6897	0.1397	0.7720
Advanced Ti (ATi)	0.1845	0.6513	0.1704	**0.7165**	0.1397	0.7778
Tw & Us	0.1888	0.6207	0.1718	0.7011	0.1409	0.7759
Tw, Us, & Tx	0.1802	0.6590	0.1710	0.7050	0.1409	**0.7797**
Tw, Us, Tx & Ti	0.1888	64.18	0.1702	0.7011	0.1404	0.7720
Tw, Us, Tx & ATi	**0.1771**	**0.6590**	**0.1688**	0.7050	**0.1393**	0.7720