Technische Hochschule
Brandenburg
University of
Applied Sciences
Fachbereich
Informatik und Medien

Untersuchung von Methoden zur Klassifizierung von Nachrichtenmeldungen – Transparente Modelle zur Erkennung von Fake News

Benedikt Michaelis

Masterarbeit • Studiengang Master Informatik • Fachbereich Informatik und Medien • 05.07.2021

Aufgabenstellung

Ziel der Arbeit ist es, Modelle zur Erkennung von Fake News anhand von Nachrichtenartikeln/Textdaten zu untersuchen und ob und wie diese transparent/erklärbar sein können. Frühere Arbeiten haben eine Überanpassung abhängig vom Datensatz festgestellt [1]. Diese wird ebenfalls untersucht.

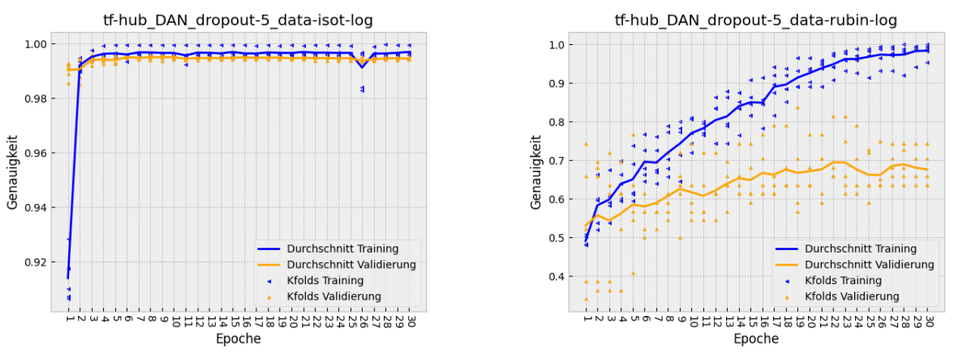
Datensätze

In dieser Arbeit werden drei Datensätze verwendet. Zwei mit einem welt- und US-politischen Bezug (ISOT [2], Rubin [3]) und einer mit Artikeln rund um den syrischen Krieg (FA-KES [4]).

Konzept

Insgesamt werden vier grundlegende Modelle untersucht. Eine logistische Regression angelehnt an Ahmed et al. [2] und drei neuronale Netze basierend auf GloVe-Einbettungen & LSTM Schichten, auf einem vortrainierten Universal Sentence Encoder und auf einem vortrainierten Neuronal Net Language Modell. Alle werden einzeln mit jedem Datensatz kreuzvalidiert und anschließend persistiert. Für einige Modelle erfolgt dafür zunächst eine Parameterauswahl. Folglich entstehen 12 persistierte Modelle (4 Modelle x 3 Datensätze).
Jedes Modell wird anschließend auf den zwei anderen unbekannten Datensätzen evaluiert. Anschließend wird versucht, die Ergebnisse einzelner Modelle post-hoc mittels LIME [5] zu erklären.

Trainingsprozess und Evaluierung auf unbekannten Daten



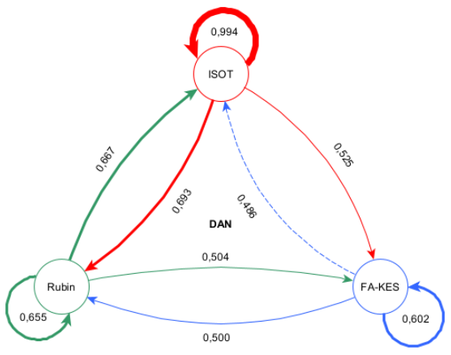


Abb. 2: Trainings- und Validierungsgenauigkeiten des Modells basierend auf dem Universal Sentence Encoder (oben links: ISOT, oben rechts: Rubin, unten links: FA-KES) und die Evaluierung auf den jeweils unbekannten Datensätzen (unten rechts)

Nach dem Training und der Persistierung aller Modelle lieferte der ISOT Datensatz konstant gute Werte im höheren 90-er Prozentbereich ähnlich wie bei Ahmed et al. [2].

Mit dem FA-KES Datensatz wurden nahezu konstant Genauigkeiten um 50 % generiert. Mit dem Rubin Datensatz liegen die Werte bei drei von vier Modellen um 65 %, ein Modell bei knapp unter 60 % (vgl. Ergebnisse des Universal Sentence Encoder Modells für alle drei Datensätze in Abbildung 1). Die guten Ergebnisse mit dem FA-KES Datensatz konnten nicht gehalten werden. Bei der Evaluierung mit den anderen beiden Datensätzen sind die Genauigkeiten jeweils deutlich eingebrochen (zwischen 30 % und 50 %) (Beispiel in Abbildung 1).

Erklärungen & Fazit.

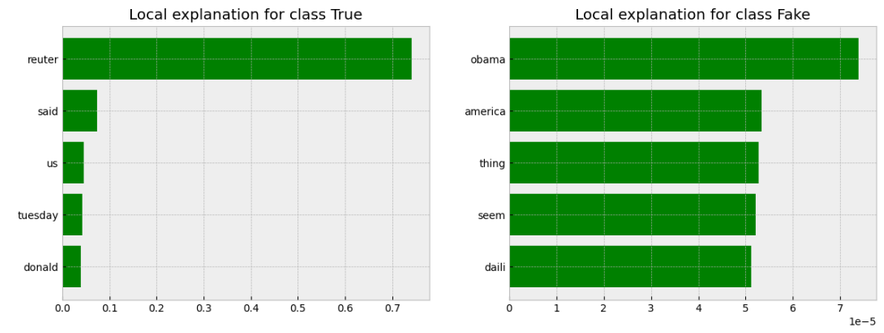


Abb. 1: Erklärungen für eine echt-negative Instanz (links) und eine falsch-positive Instanz (rechts) mit der logistischen Regression trainiert auf dem ISOT Datensatz. Die Instanzen stammen aus dem Rubin Datensatz.

Erklärungen konnten nur post-hoc lokal für die logistische Regression erzeugt werden. Die Versuche mit den anderen Modellen schlugen fehl. Die vorhandenen Erklärungen zeigten allerdings, dass das auf ISOT trainierte Modell Artikel auf Basis der Wörter „said" und „reuters" als echte Artikel bewertet und alles andere als Fake News.
Dies zeigt zum einen, dass ein besonderes Augenmerk auf die Datensätze und deren Erzeugung geworfen werden muss. Erklärungen für Modelle sind außerdem wichtig, um Fehler besser zu erkennen.

Quellen

[1] F. Torabi Asr und M. Taboada, „Big Data and quality data for fake news and misinformation detection", Big Data Soc., 2019, doi: 10.1177/2053951719843310.

[2] H. Ahmed, I. Traore, und S. Saad, „Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques", in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Okt. 2017, Bd. 10618 LNCS, S. 127–138, doi: 10.1007/978-3-319-69155-8_9.

[3] T. V. Asubiaro und V. L. Rubin, „Comparing features of fabricated and legitimate political news in digital environments (2016-2017)", Proc. Assoc.

[4] F. K. Abu Salem, R. Al Feel, S. Elbassuoni, M. Jaber, und M. Farah, „FA-KES: A Fake News Dataset around the Syrian War", Jan. 2019, doi: 10.5281/ZENODO.2607278.

[5] M. T. Ribeiro, S. Singh, und C. Guestrin, „„Why should i trust you?' Explaining the predictions of any classifier", in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, Bd. 13-17-August-2016, S. 1135–1144, doi: 10.1145/2939672.2939778.

Betreuung: Prof. Dr.-Ing Sven Buchholz, Dipl.-Inform. Ingo Boersch • Technische Hochschule Brandenburg