Nutzung der Dependenzsyntax zur Erkennung von Hassrede

Ivo de Souza Bueno Júnior

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Centrum für Informations- und Sprachverabeitung (CIS), LMU München

24. Januar, 2024

Gliederung

Motivation

- 41% der amerikanischen Internetnutzer haben schonmal online Belästigung erlebt (Vogels, 2021).
- Soziale Medien erlauben es zwar, den Usern Hassreden zu melden, aber die Anzahl von Meldungen ist zu groß, um jede einzelne Meldung manuell zu überprüfen.
- Hassrede zu erkennen ist nicht einfach:
  - Ziel von beleidigenden Wörtern;
  - Wer der Sprecher ist;
  - Impliziter Hass.

hate speech

noun [U]

public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation (= the fact of being gay, etc.):

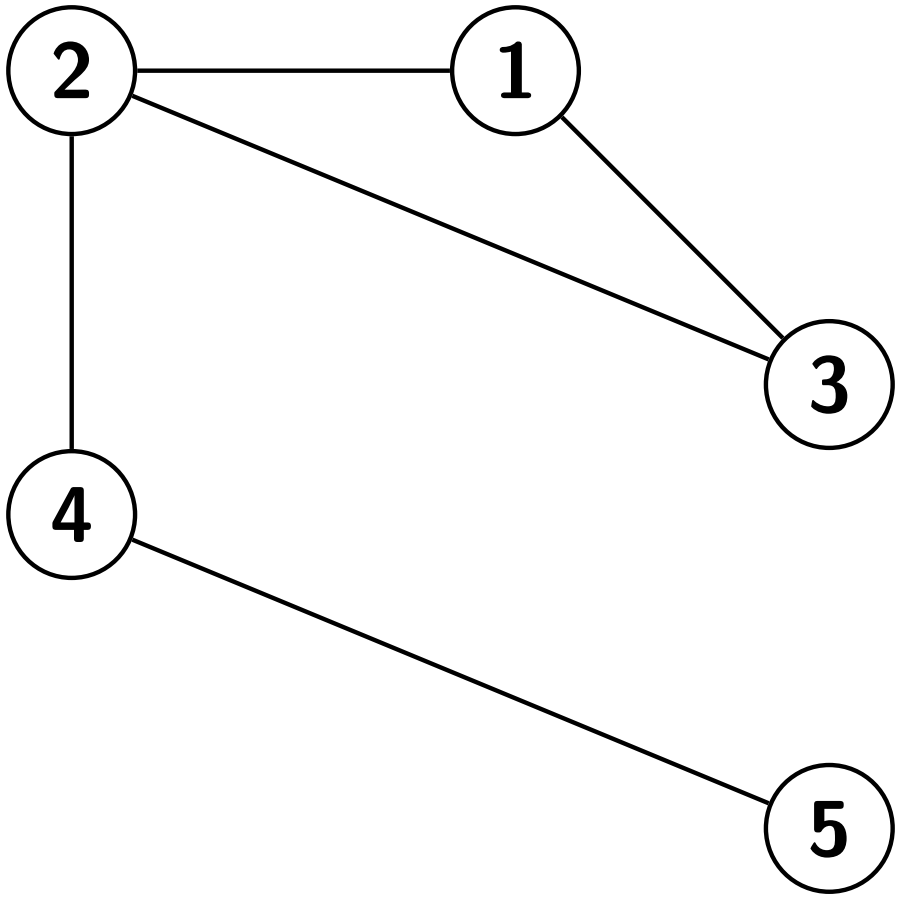
Abbildung: Definition von Hassrede von Cambridge Dictionary.

- "Öffentliche Äußerungen, die Hass zum Ausdruck bringen oder zu Gewalt gegenüber einer Person oder Gruppe aufgrund von Rasse, Religion, Geschlecht oder sexueller Orientierung (= der Tatsache, homosexuell zu sein usw.) ermutigen."

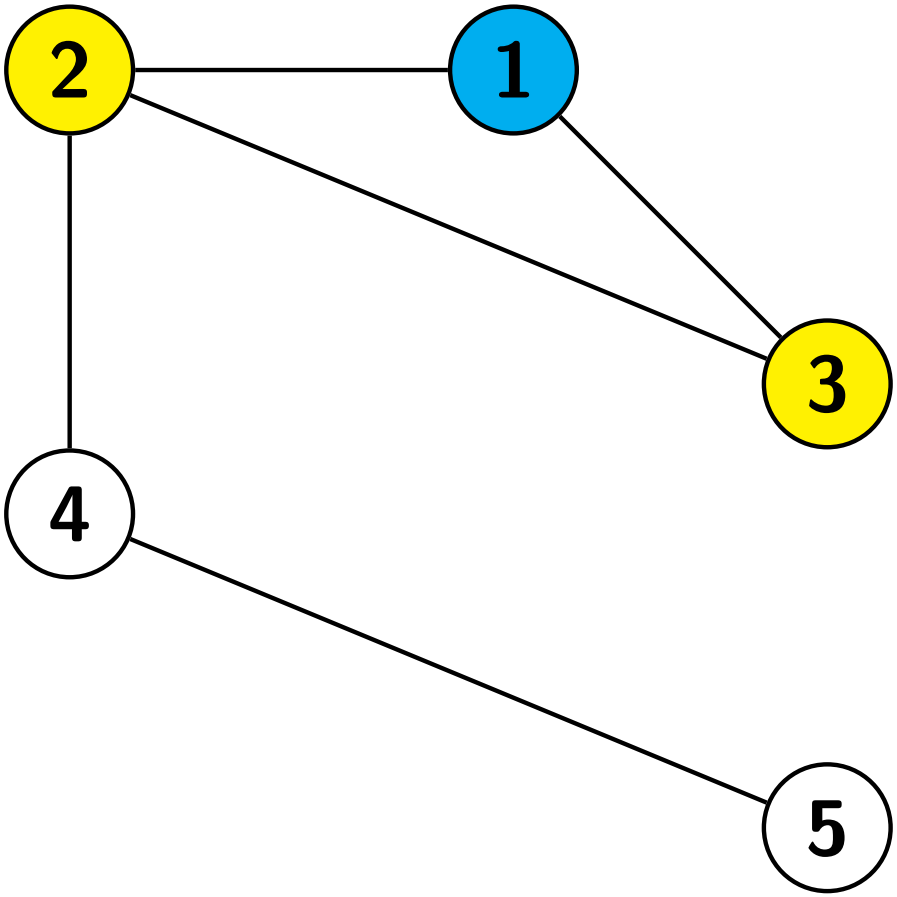
Gliederung

Graph Convolutional Network (GCN)

- Möglichkeit, lokale bzw. räumliche Informationen zu nutzen, um z. B. einen Klassifikator zu bauen.
- Rechnet ein Embedding für jeden Knoten, in Abhängigkeit von den Nachbarn.

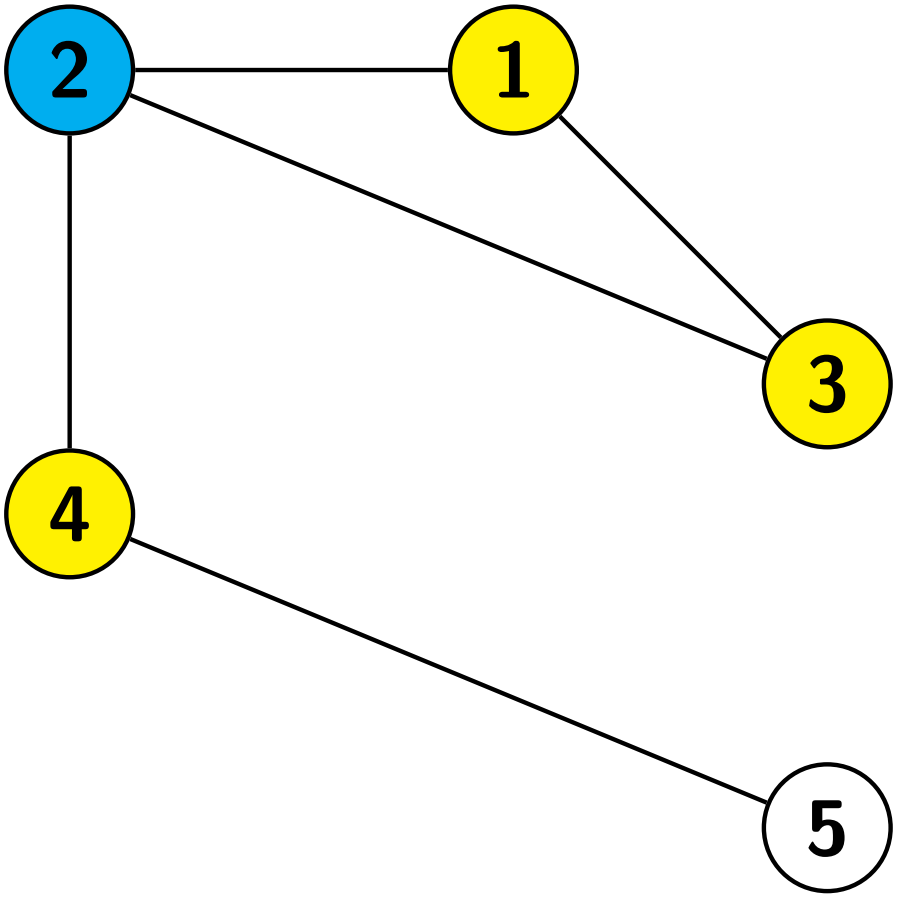


GCN: Pooling



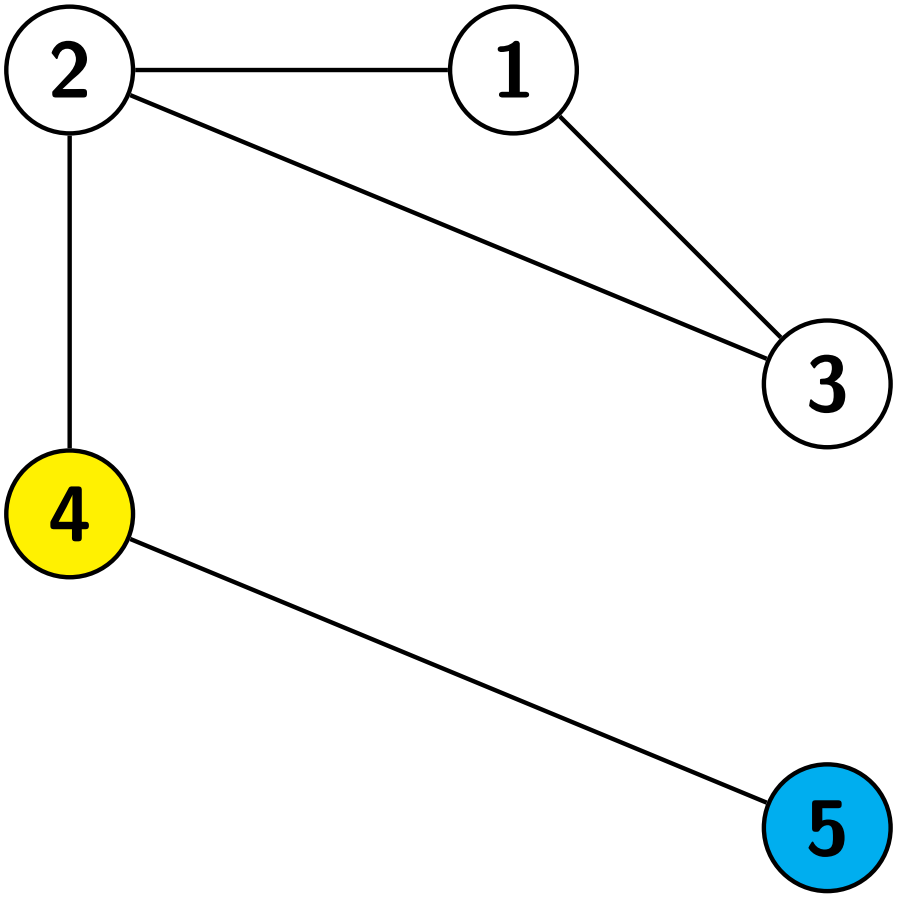
Knote 1: $AvgPool(1, 2, 3) = 2$

GCN: Pooling



Knote 1: $AvgPool(1, 2, 3) = 2$
Knote 2: $AvgPool(2, 1, 3, 4) = 2.5$

GCN: Pooling



Knote 1: $AvgPool(1, 2, 3) = 2$
Knote 2: $AvgPool(2, 1, 3, 4) = 2.5$
Knote 3: $AvgPool(3, 1, 2) = 2$
Knote 4: $AvgPool(4, 2, 5) = 3.667$
Knote 5: $AvgPool(5, 4) = 4.5$

GCN: Linear Layer

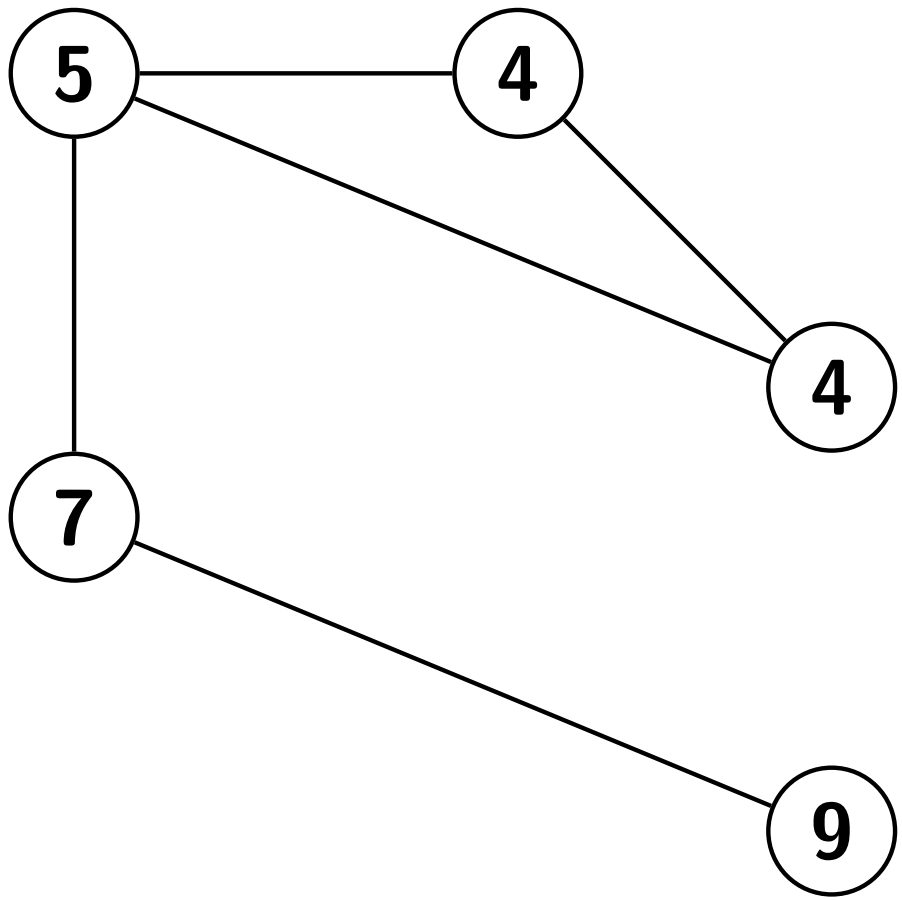
Knote 1:
$LinearLayer(AvgPool(1, 2, 3)) = 4$

Knote 2:
$LinearLayer(AvgPool(2, 1, 3, 4)) = 5$

Knote 3:
$LinearLayer(AvgPool(3, 1, 2)) = 4$

Knote 4:
$LinearLayer(AvgPool(4, 2, 5)) = 7$

Knote 5:
$LinearLayer(AvgPool(5, 4)) = 9$



GCN: Weitere Details

- Wie werden die Informationen zur Nachbarschaft gespeichert?

GCN: Weitere Details

- Wie werden die Informationen zur Nachbarschaft gespeichert?

  In eine Adjazenzmatrix:

	1	2	3	4	5
1	1	1	1	0	0
2	1	1	1	1	0
3	1	1	1	0	0
4	0	1	0	1	1
5	0	0	0	1	1

GCN: Weitere Details

- Wie werden die Informationen zur Nachbarschaft gespeichert?

  In eine Adjazenzmatrix:

	1	2	3	4	5
1	1	1	1	0	0
2	1	1	1	1	0
3	1	1	1	0	0
4	0	1	0	1	1
5	0	0	0	1	1

- Dependenzgrafen sind gerichtet, kann man das bilden?

GCN: Weitere Details

- Wie werden die Informationen zur Nachbarschaft gespeichert?

  In eine Adjazenzmatrix:

	1	2	3	4	5
1	1	1	1	0	0
2	1	1	1	1	0
3	1	1	1	0	0
4	0	1	0	1	1
5	0	0	0	1	1

- Dependenzgrafen sind gerichtet, kann man das bilden?

  Nein. Pfeile werden als ungerichtete Kanten interpretiert.

GCN: Weitere Details

- Wie werden die Informationen zur Nachbarschaft gespeichert?

  In eine Adjazenzmatrix:

	1	2	3	4	5
1	1	1	1	0	0
2	1	1	1	1	0
3	1	1	1	0	0
4	0	1	0	1	1
5	0	0	0	1	1

- Dependenzgrafen sind gerichtet, kann man das bilden?

  Nein. Pfeile werden als ungerichtete Kanten interpretiert.
- Was ist der initiale Wert von Knoten?

GCN: Weitere Details

- Wie werden die Informationen zur Nachbarschaft gespeichert?

  In eine Adjazenzmatrix:

	1	2	3	4	5
1	1	1	1	0	0
2	1	1	1	1	0
3	1	1	1	0	0
4	0	1	0	1	1
5	0	0	0	1	1

- Dependenzgrafen sind gerichtet, kann man das bilden?

  Nein. Pfeile werden als ungerichtete Kanten interpretiert.
- Was ist der initiale Wert von Knoten?

  Implementationsabhängig: Zufällige Werte, vortrainierte Embeddings (GloVe, BERT, einige LSMT, ...)

Gliederung

Dependency Graphical Convolutional Networks

- "Abusive language detection using syntactic dependency graphs" (Narang and Brew, 2020).

- "Abusive language detection using syntactic dependency graphs" (Narang and Brew, 2020).
- Datensatz:

Datensatz	**Kategorien**		
Davidson et al. (2017)	Hass	Beleidigend	Harmlos
	1,430	19,190	4,163
Waseem and Hovy (2016)	Rassismus	Sexismus	Harmlos
	1,939	3,148	11,115
Davidson erweitert	Hass	Beleidigend	Harmlos
	1,430	19,190	15,278

- "Abusive language detection using syntactic dependency graphs" (Narang and Brew, 2020).
- Datensatz:

Datensatz	**Kategorien**		
Davidson et al. (2017)	Hass	Beleidigend	Harmlos
	1,430	19,190	4,163
Waseem and Hovy (2016)	Rassismus	Sexismus	Harmlos
	1,939	3,148	11,115
Davidson erweitert	Hass	Beleidigend	Harmlos
	1,430	19,190	15,278

- Dependenzsyntax Parser

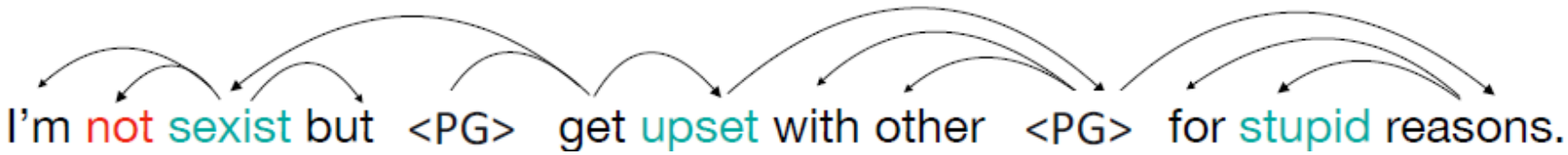


Abbildung: Syntaxbaum von einem Tweet von Waseem and Hovy (2016) Datensatz.

DepGCN: Modell

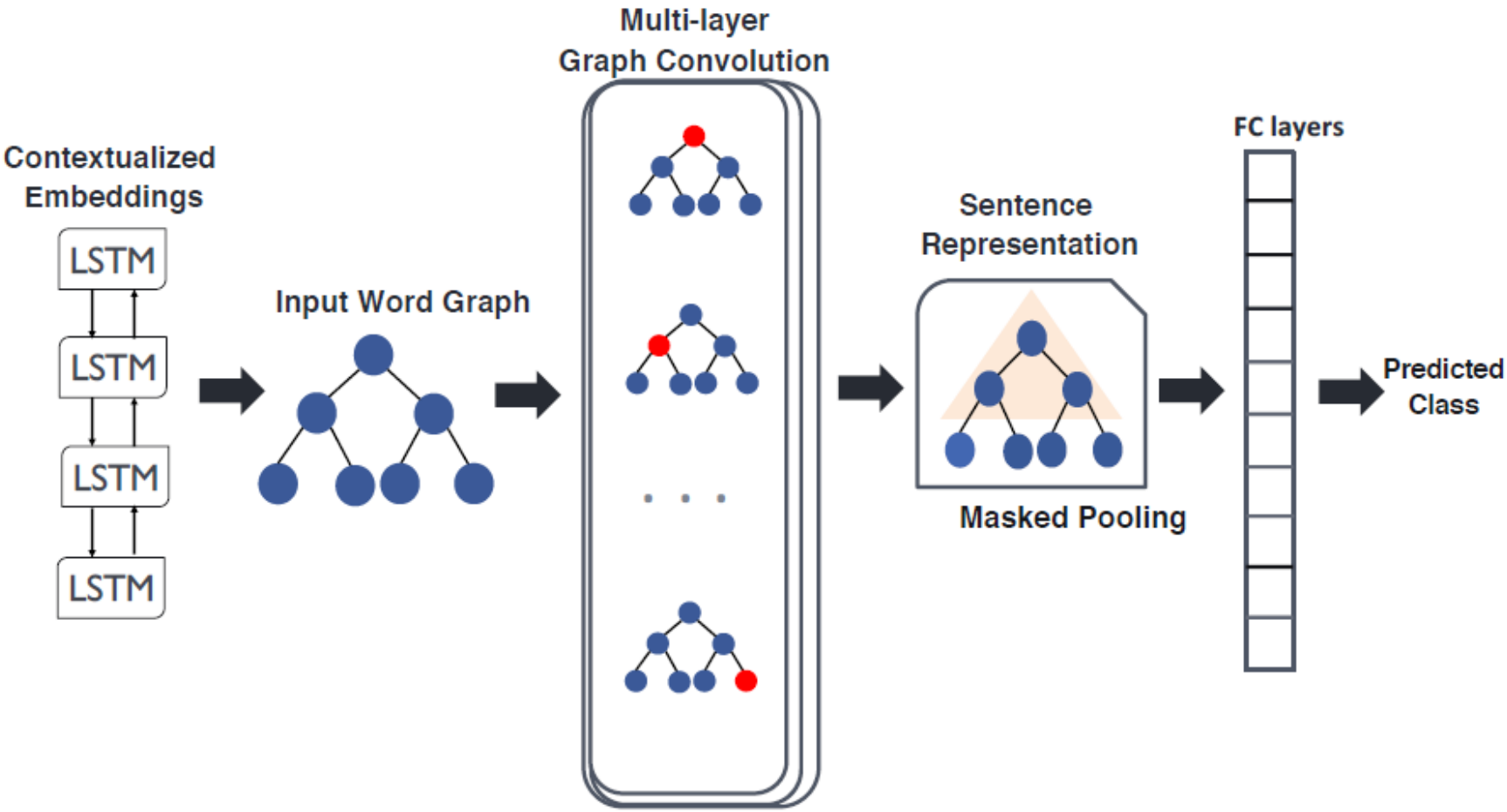


Abbildung: DepGCN Modell (Narang and Brew, 2020).

DepGCN: Ergebnisse

Modell	Hass	Beleidigend	Harmlos
N-grams	0.35	0.88	0.88
BERT	0.45	0.94	0.96
BiLSTM	0.31	0.93	0.94
DepGCN	0.47	0.94	0.96
BiLSTM + DepGCN	**0.49**	**0.95**	**0.97**

Tabelle: F1-Score mit dem erweiterten Davidson et al. (2017) Datensatz. Adaptiert von Narang and Brew (2020).

DepGCN: Ergebnisse

Modell	Hass	Beleidigend	Harmlos	Gesamt
N-grams	0.46	0.94	0.84	0.89
BERT	0.42	**0.95**	**0.88**	**0.91**
BiLSTM	0.52	0.94	0.86	0.90
DepGCN	0.50	0.94	0.86	0.90
BiLSTM + DepGCN	**0.53**	0.94	0.87	**0.91**

Tabelle: F1-Score mit dem originalen Davidson et al. (2017) Datensatz. Adaptiert von Narang and Brew (2020).

DepGCN: Ergebnisse

Modell	Rassismus	Sexismus	Harmlos	Gesamt
N-grams	0.75	0.71	0.88	0.83
BERT	**0.78**	**0.81**	**0.91**	**0.88**
BiLSTM	0.72	0.71	0.89	0.84
DepGCN	0.76	0.72	0.88	0.83
BiLSTM + DepGCN	**0.78**	0.74	0.90	0.85

Tabelle: F1-Score mit dem Waseem and Hovy (2016) Datensatz. Adaptiert von Narang and Brew (2020).

DepGCN: Trainings- und Inferenzzeit

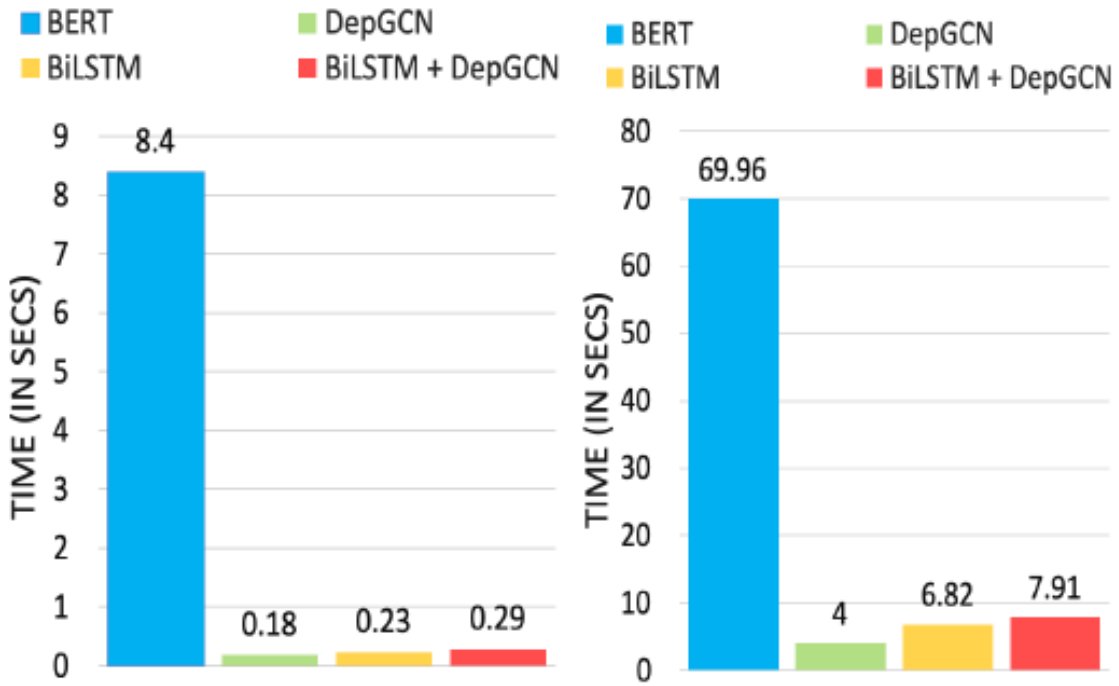


Abbildung: Inferenzzeit pro 1k Tweets (links) und Trainingszeit pro Epoch (rechts) (Narang and Brew, 2020).

Gliederung

Syntax-based LSTM (SyLSTM)

- "Leveraging Dependency Grammar for Fine-Grained Offensive Language Detection using Graph Convolutional Networks" (Goel and Sharma, 2022).

Syntax-based LSTM (SyLSTM)

- "Leveraging Dependency Grammar for Fine-Grained Offensive Language Detection using Graph Convolutional Networks" (Goel and Sharma, 2022).
- Datensatz:
  - Offensive Language Identification Dataset (Datensatz zur Identifizierung beleidigender Sprache), von Zampieri et al. (2019): 14000 Tweets.
  - Hate Speech and Offensive Language Dataset (Datensatz zu Hassreden und beleidigender Sprache), von Davidson et al. (2017): 25000 Tweets.

Syntax-based LSTM (SyLSTM)

- Vorverarbeitung:

Vorverarbeitungsschritt	**Beschreibung**
Ersetzen von Benutzernamen	Alle Benutzernamen werden durch '@user' ersetzt. Zum Beispiel wird '@india' zu '@user'.
Ersetzen von URLs	URLs in einem Tweet werden durch das Wort 'url' ersetzt.
Hashtag-Segmentierung	Zum Beispiel wird '#banislam' zu '# banislam'.
Emoji-Normalisierung	Normalisierung von Emoji-Vorkommen in Text. Zum Beispiel wird ':)' zu 'smiley face'.
Trennung von zusammengesetzten Wörtern	Trennun von zusammengesetzten Wörtern. Zum Beispiel wird 'putuporshutup' zu 'put up or shut up'.
Reduzierung von Wortlängen	Reduzierung von Wortlängen und Ausrufezeichen. Zum Beispiel wird 'waaaaayyyy' zu 'waayy'.

Tabelle: Vorverarbeitungsschritte. Adaptiert von Goel and Sharma (2022).

SyLSTM: Modell

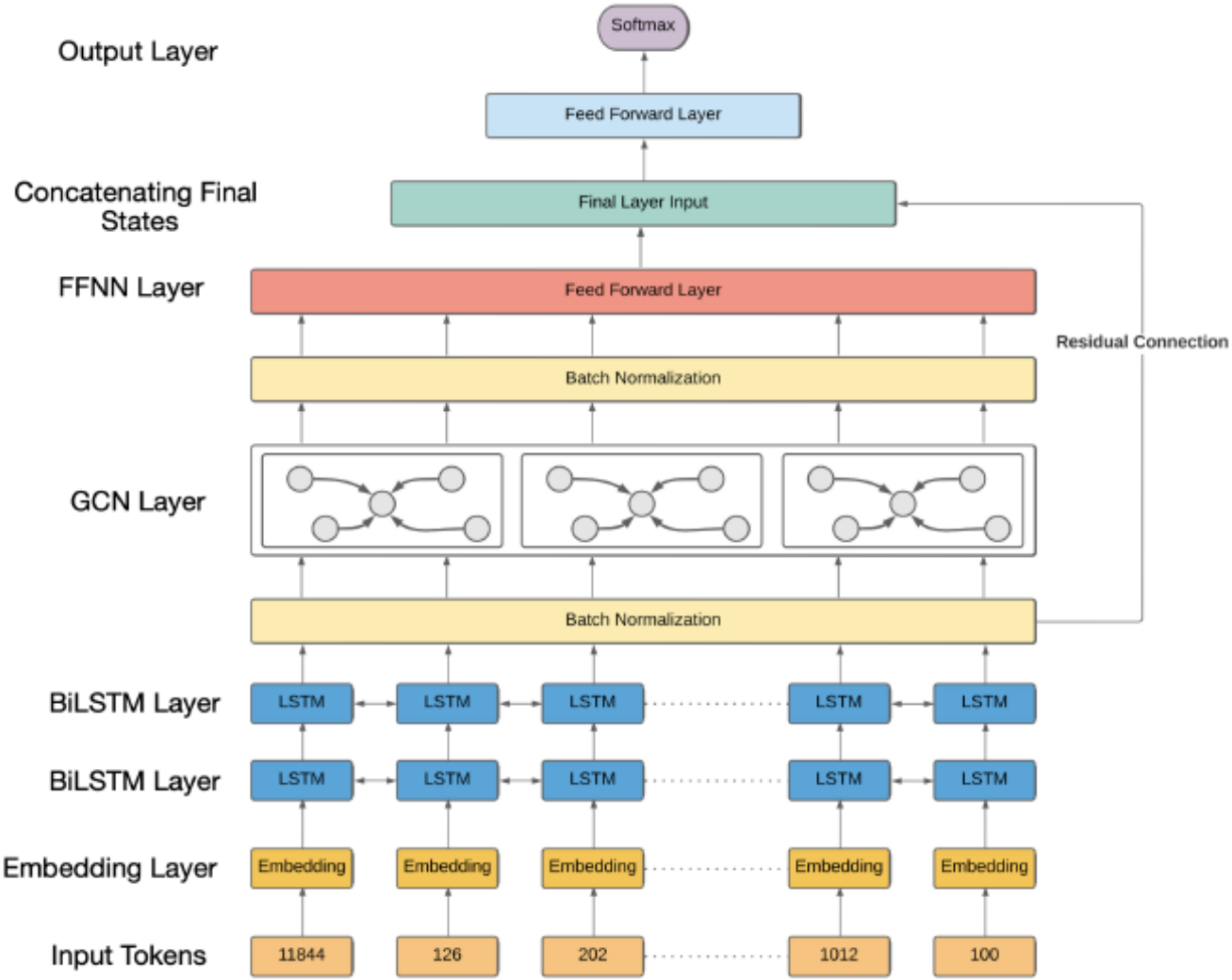


Abbildung: SyLSTM Modell (Goel and Sharma, 2022).

SyLSTM: Ergebnisse

Modell	**Precision**	**Recall**	**F1-score**
All OFF	8.4	28.2	12.1
All NOT	52.4	72.7	60.4
SVM	77.7	80.2	78.6
BiLSTM	81.7	82.8	82.0
BERT	87.3	85.8	85.7
SyLSTM	85.2	88.1	86.4
SyLSTM*	**87.6**	**88.1**	**87.4**

Tabelle: Erkennung beleidigender Sprache mit Zampieri et al. (2019) Datensatz. Trennung zwischen beleidigender Sprache (OFF) und nicht beleidigender Sprache (NOT). Adaptiert von Goel and Sharma (2022).

SyLSTM: Ergebnisse

Modell	Precision	Recall	F1-score
All TIN	78.7	88.6	83.4
All UNT	1.4	11.3	12.1
SVM	81.6	84.1	82.6
BiLSTM	84.8	88.4	85.7
BERT	88.4	92.3	89.6
SyLSTM	90.6	91.6	91.4
SyLSTM*	**94.4**	**92.3**	**93.2**

Tabelle: Kategorisierung beleidigender Sprache mit Zampieri et al. (2019) Datensatz. Trennung zwischen gezielten Beleidigungen und Angriffen (TIN) und nicht gezielten Beleidigungen und Angriffen (UNT). Adaptiert von Goel and Sharma (2022).

SyLSTM: Ergebnisse

Modell	Precision	Recall	F1-score
All GRP	13.6	37.4	19.7
All IND	22.1	47.3	30.3
All OTH	3.4	16.2	5.4
SVM	56.1	62.4	58.3
BiLSTM	56.1	65.8	60.4
BERT	58.4	66.2	60.9
SyLSTM	60.3	**67.4**	63.4
SyLSTM*	**62.4**	66.3	**64.4**

Tabelle: Identifizierung beleidigender Sprachziele mit Zampieri et al. (2019) Datensatz. Trennung zwischen Gruppe (GRP), Individuum (IND), und Anderen (OTH). Adaptiert von Goel and Sharma (2022).

SyLSTM: Ergebnisse

Modell	Precision	Recall	F1-score
All HATE	0.2	6.1	0.4
All OFF	3.1	16.9	5.3
All NONE	58.8	77.2	66.7
SVM	84.9	90.1	88.2
BiLSTM	90.3	90.2	90.3
BERT	91.2	90.4	91.0
DepGCN	-	-	90.0
BiLSTM + DepGCN	-	-	91.0
SyLSTM	90.5	91.4	91.4
SyLSTM*	**92.3**	**92.8**	**92.7**

Tabelle: Erkennung von Hassrede und beleidigender Sprache mit Davidson et al. (2017) Datensatz. Trennung zwischen Hassrede (HATE), beleidigende Sprache (OFF), und harmlose Sprache (NONE). Adaptiert von Goel and Sharma (2022).

Abschluss

- Dependenzsyntax zusammen mit der Nutzung von GCNs kann dabei helfen, Hassrede...
  - ... besser zu erkennen;
  - ... von beleidigender Sprache besser zu trennen;
  - ... schneller zu erkennen.

Danke für die Aufmerksamkeit!

Literatur

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Divyam Goel and Raksha Sharma. 2022. Leveraging dependency grammar for fine-grained offensive language detection using graph convolutional networks. *arXiv preprint arXiv:2205.13164*.

Kanika Narang and Chris Brew. 2020. Abusive language detection using syntactic dependency graphs. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 44–53.

Emily A. Vogels. 2021. The state of online harassment.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.