Predicting the Type and Target of Offensive Posts in Social Media

Marcos Zampieri, Shervin Malmasi, Preslav Nakov et al. (2019)

Herausforderungen des Natural Language Processing: Hate Speech, diskriminierende Sprache und populistische Rhetorik | Sommersemester 2024 | Uni Köln | Präsentiert von: Mariia Galevskaia

Agenda

- Einführung
- Hierarchisches Annotationsschema
- Datensammlung (Der OLID-Datensatz)
- Annotationsprozess (I, II, III)
- Experimente und Evaluation (I, II)
- Experimentergebnisse
    - Offensive Language Detection
    - Categorization of Offensive Language
    - Offensive Language Target Identification
- Zusammenfassung
- Diskussion

Einführung

- Die Arbeit befasst sich mit der Erkennung und Kategorisierung von beleidigenden Inhalten in sozialen Medien.

- Im Gegensatz zu früheren Arbeiten, die sich auf bestimmte Arten von beleidigenden Inhalten (z. B. Hassrede oder Cybermobbing) konzentrieren, präsentiert diese Forschung einen hierarchischen Ansatz zur Kategorisierung von beleidigender Sprache und zur Identifizierung ihres Ziels.

Hierarchisches Annotationsschema

Die Autoren schlagen ein dreistufiges Annotationsschema vor:

- **Ebene A**: Ermittelt, ob ein Text beleidigend ist oder nicht (**OFF/NOT**).
- **Ebene B**: Kategorisiert die Art des beleidigenden Inhalts, ob es sich um eine gezielte Beleidigung oder ungezielte Vulgarität handelt (**TIN/UNT**).
- **Ebene C**: Bestimmt das Ziel des beleidigenden Inhalts, ob es eine Einzelperson, eine Gruppe oder etwas Anderes ist (**IND/GRP/OTH**).

Tweet	A	B	C
@USER He is so generous with his offers.	NOT	—	—
IM FREEEEE!!!! WORST EXPERIENCE OF MY FUCKING LIFE	OFF	UNT	—
@USER Fuk this fat cock sucker	OFF	TIN	IND
@USER Figures! What is wrong with these idiots? Thank God for @USER	OFF	TIN	GRP

Datensammlung (Der OLID Datensatz)

- Sammlung von **Tweets** mit der Hilfe von **API** und **Schlüsselwörtern**.
- Nur die Schlüsselwörter und Konstruktionen, die mit hoher Wahrscheinlichkeit in beleidigenden Inhalten vorkommen (z. B " ***she is***" und "***to:BreitBartNews***").
- Während der Annotationsprozess wurden verschiedene Schlüsselwörter **eingeschlossen** und **ausgeschlossen** *(um die Verteilung anstößiger Tweets bei etwa 30 % des Datensatzes zu halten)*.
- Der kompletter Datensatz beträgt **50%** von politischen Inhalten und **50%** von keinen politischen Inhalten.

Keyword	Offensive %
medical marijuana	0.0
they are	5.9
to:NewYorker	8.3
you are	21.0
she is	26.6
to:BreitBartNews	31.6
he is	32.4
gun control	34.7
-filter:safe	58.9
conservatives	23.2
antifa	26.7
MAGA	27.7
liberals	38.0

Annotationsprozess I

- Probe-Annotation mit **300** Instanzen, **6** Experten und **9** Schlüsselwörtern:
  i. für die Evaluierung des Tagsets (Schlüsselwörter);
  ii. für die Evaluierung der Datenabrufmethode;
  iii. für die Erstellung eines Goldstandards mit Instanzen, die als Testfragen verwendet werden könnten, um die Qualität der Annotatoren für den Rest der Daten sicherzustellen (aufgrund des Crowdsourcings)

Annotationsprozess II

- Fleiss' *kappa* (Training-Datensatz, **5** annotators, **21** tweets) = **0.83** für Ebene A

- Annotation mit der Hilfe von **Crowdsourcing** (Annotatoren mit der Erfahrung + Testfragen). Jede Instanz im Datensatz wurde von mehreren Annotatoren annotiert und die Zustimmung zwischen den Annotatoren wurde am Ende berechnet.

Annotationsprozess III

- Zwei Annotationen pro Tweet
- Bei Uneinigkeit zwischen den Annotatoren wurde eine dritte Annotation gefordert und dann wurde die Mehrheitsentscheidung getroffen.
- Das endgültige Label für jeden Tweet wurde auf Basis der Mehrheitsentscheidung bestimmt.
- **60%** der gesamten Zeit reichte die Zustimmung von zwei Annotatoren. Es wurde niemals mehr als drei Annotationen für eine Instanz gebraucht.

Experimente und Evaluation I

- Modelle:
  - **SVM** (Support Vector Machine): Eine klassische Methode des maschinellen Lernens, die sich bei Textklassifizierungsaufgaben bewährt hat.
  - **BiLSTM** (Bidirectional Long Short-Term Memory): Ein rekurrentes neuronales Netzwerk, das den Kontext sowohl links als auch rechts vom aktuellen Wort berücksichtigt, was das Textverständnis verbessert.
  - **CNN** (Convolutional Neural Network): Ein konvolutionales neuronales Netzwerk, das durch die Erkennung bedeutender Merkmale auf verschiedenen Ebenen des Textes gut mit Textdaten arbeitet.
- Alle Modelle wurden auf Training-Datensatz trainiert (**13, 240 Tweets**) und wurden dann auf der Basis der vorhergesagten Labels vom Test-Datensatz (**860 Tweets**) evaluiert.

Experimente und Evaluation II

- Da die Verteilung der Labels sehr unausgeglichen war, wurde die Leistung verschiedener Modelle anhand des **makro-durchschnittlichen F1-Scores** (macro-averaged F1 score) evaluiert und verglichen.
- Als andere Leistungsindikatoren wurden auch die Genauigkeit (**Precision**), die Vollständigkeit (**Recall**), **F1-Score** und gewichteter Durchschnitt (**weighted average**) für jede Klasse berücksichtigt.
- Zusätzlich wurde die Leistung der Modelle mit einfachen Mehrheits- und Minderheitsklassen-Baselines verglichen.

A	B	C	Training	Test	Total
OFF	TIN	IND	2,407	100	2,507
OFF	TIN	OTH	395	35	430
OFF	TIN	GRP	1,074	78	1,152
OFF	UNT	—	524	27	551
NOT	—	—	8,840	620	9,460
All			13,240	860	14,100

Experimentergebnisse: Offensive Language Detection

	NOT			OFF			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
SVM	0.80	0.92	0.86	0.66	0.43	0.52	0.76	0.78	0.76	0.69
BiLSTM	0.83	0.95	0.89	0.81	0.48	0.60	0.82	0.82	0.81	0.75
CNN	0.87	0.93	0.90	0.78	0.63	0.70	0.82	0.82	0.81	**0.80**
All NOT	-	0.00	0.00	0.72	1.00	0.84	0.52	0.72	0.	0.42
All OFF	0.28	1.00	0.44	-	0.00	0.00	0.08	0.28	0.12	0.22

- Zwei anderen Modelle übertreffen **SVM**. **CNN** (**0.80**) aber übertrifft an seiner Stelle **BiLSTM**.

Experimentergebnisse: Categorization of Offensive Language

	TIN			UNT			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
SVM	0.91	0.99	0.95	0.67	0.22	0.33	0.88	0.90	0.88	0.64
BiLSTM	0.95	0.83	0.88	0.32	0.63	0.42	0.88	0.81	0.83	0.66
CNN	0.94	0.90	0.92	0.32	0.63	0.42	0.88	0.86	0.87	**0.69**
All TIN	0.89	1.00	0.94	-	0.00	0.00	0.79	0.89	0.83	0.47
All UNT	-	0.00	0.00	0.11	1.00	0.20	0.01	0.11	0.02	0.10

- CNN (**0.69**) übertrifft wieder BiLSTM und SVM.
- Alle Modelle erkennen besser **TIN** im Vergleich zu **UNT**.

Experimentergebnisse: Offensive Language Target Identification

	GRP			IND			OTH			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
SVM	0.66	0.50	0.57	0.61	0.92	0.73	0.33	0.03	0.05	0.58	0.62	0.56	0.45
BiLSTM	0.62	0.69	0.65	0.68	0.86	0.76	0.00	0.00	0.00	0.55	0.66	0.60	**0.47**
CNN	0.75	0.60	0.67	0.63	0.94	0.75	0.00	0.00	0.00	0.57	0.66	0.60	**0.47**
All GRP	0.37	1.00	0.54	-	0.00	0.00	-	0.00	0.00	0.13	0.37	0.20	0.18
All IND	-	0.00	0.00	0.47	1.00	0.64	-	0.00	0.00	0.22	0.47	0.30	0.21
All OTH	-	0.00	0.00	-	0.00	0.00	0.16	1.00	0.28	0.03	0.16	0.05	0.09

- Alle drei Modelle zeigen ähnliche Ergebnisse. **SVM** - 0.45, **BiLSTM** - 0.47, **CNN** - 0.47
- **OTH Gruppe (= 0)**: 1. heterogene Sammlung von Zielen; 2. weniger Trainingsdaten (nur **395**)

Zusammenfassung

- Die Erstellung des OLID-Datensatzes, der sich als eine wertvolle Ressource für zukünftige Forschungen darstellt.
- Die Entwicklung des dreistufigen Annotationsschema zur Erkennung und Kategorisierung von beleidigenden Inhalten.
- Die Experimente haben gezeigt, dass neuronale Netze wie CNN und BiLSTM bei der Erkennung und Klassifizierung von beleidigenden Inhalten effektiver sind als traditionelle Methoden wie SVM.
- Ein weiterer Schritt wäre, den OLID-Datensatz mit den anderen Datensätzen, die für die gleichen Aufgaben annotiert wurden, zu vergleichen.
- Eine weitere Entwicklung von Datensätzen in anderen Sprachen gemäß dem dreistufigen Annotationsschema.

Diskussion

- Was könnten die Gründe für die unterschiedlichen Leistungen der Modelle in den verschiedenen Klassifizierungsstufen (**A, B, C**) sein?

- Wie könnte die Implementierung solcher Modelle in sozialen Medien die Meinungsfreiheit beeinflussen?

Quelle

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, Ritesh Kumar. 2019. *Predicting the Type and Target of Offensive Posts in Social Media.*