Melanie Siegel

Inter-*nett?*

Extreme Meinungen im Netz erkennen und filtern

In: Abecedarium der Sprache / Constanze Fröhlich, Martin Grötschel, Wolfgang Klein (Hg.). – ISBN: 978-3-86599-416-5. – Berlin: Kulturverlag Kadmos, 2019. S. 95-100

Persistent Identifier: urn:nbn:de:kobv:b4-opus4-30248

I – *Stiller, Berlin Charlottenburg*

Inter-*nett?*
Extreme Meinungen im Netz erkennen und filtern

Melanie Siegel

Die sozialen Netzwerke werden in der letzten Zeit überflutet von »Meinungsäußerungen«, die schwere Beleidigungen, Verleumdungen und Diskriminierungen enthalten. Zum Teil werden sie automatisiert geteilt und verbreitet, um den Anschein zu erwecken, hier handele es sich um »Volkes Stimme«.

Wissenschaftlerinnen und Wissenschaftler suchen nun nach Verfahren, um diese automatisch zu erkennen und dann auch Filterungsmöglichkeiten anzubieten. In einem Programmierwettbewerb – einer sogenannten »Shared Task« – der Interest Group on German Sentiment Analysis wurden im September 2018 Methoden dafür getestet.

Um solche Methoden überhaupt entwickeln zu können, muss man zunächst ein sogenanntes annotiertes Textkorpus aufbauen – also extreme Meinungsäußerungen erst einmal sammeln. Die Plattform Twitter eignet sich besonders gut, um solche Texte zusammenzustellen, weil sie einen automatisierten Zugriff erlaubt. Es gibt schon bei der Sammlung verschiedene Möglichkeiten: Man kann eine Reihe von Stichwörtern aufstellen, die vermutlich in extremen Meinungsäußerungen stehen, beispielsweise »kriminell«, und Tweets suchen, die sie enthalten. Man kann auch gezielt nach Hashtags suchen, die auf extreme Meinungsäußerungen hindeuten, wie etwa »#rapefugees«. Eine andere Möglichkeit besteht darin, Accounts zu identifizieren, die besonders häufig extreme Meinungsäußerungen posten, und die Tweets von diesen Accounts zu sammeln. Dabei besteht aber immer die Gefahr, dass man Themenbereiche übersieht. Eine genaue Beobach-

tung der Diskussionen auf Twitter ist daher unumgänglich, bevor eine größere Sammlung beginnt.

Um herauszufinden, welche Eigenschaften Tweets mit extremen Meinungsäußerungen von anderen Tweets unterscheiden, und um das Material für automatische Lernverfahren zu vervollständigen, muss man im zweiten Schritt Tweets sammeln, die mit den extremen Meinungsäußerungen vergleichbar sind, aber keine Beleidigungen, Verleumdungen und Diskriminierungen enthalten. Hier ist es sinnvoll, Tweets aus denselben Themenbereichen zu sammeln, denn sonst würde im Ergebnis alles, was beispielsweise zum Thema »Flüchtlinge« (zu dem es besonders viele extreme Meinungsäußerungen auf Twitter gibt) gepostet wird, als extreme Meinungsäußerung markiert. Eine Möglichkeit ist es, aus den Accounts, die häufig extreme Meinungsäußerungen posten, auch jene anderen Meinungsäußerungen zu nehmen, um eine gute Vergleichbarkeit herzustellen. Insgesamt müssen dabei aber so viele unterschiedliche Twitter-Accounts betrachtet werden wie möglich.

Die gesammelten Tweets müssen anschließend klassifiziert werden. Es stellt sich schnell heraus, dass schon bei der Unterscheidung in extreme (Klasse OFFENSIVE) und andere Tweets (Klasse OTHER) mehrere Personen denselben Text unterschiedlich beurteilen. Es wird deshalb zunächst ein Teil des Datensatzes von mehreren Testpersonen annotiert, also als extrem oder nicht extrem bewertet, dann wird die Übereinstimmung zwischen ihnen (das »Inter-Annotator-Agreement«) gemessen. Im ersten Durchgang gibt es typischerweise nicht genügend Übereinstimmung bei der Einschätzung der Tweets. Daher werden strittige Beispiele diskutiert und daraus eine Annotationsrichtlinie abgeleitet. Danach werden weitere Beispiele annotiert, es wird die Übereinstimmung gemessen, weiter diskutiert und so weiter, bis eine gute Übereinstimmungsquote erreicht ist. Noch etwas komplexer wird es, wenn man die Tweets der Klasse OFFENSIVE weiter aufteilen möchte, in unserem Fall in INSULT (Beleidigungen), ABUSE (Diskriminierungen) und PROFANITY (Beschimpfungen).

Die so gewonnenen Daten teilt man in eine Trainings- und eine etwas kleinere Testmenge auf. Die Testmenge legt man zunächst beiseite und die Forschergruppen beschäftigen sich mit der Trainingsmenge.

Zunächst geht es darum, die Werkzeuge der Sprachverarbeitung an die Sprache in den Tweets anzupassen. Sie sind nämlich meistens für Sprache in Zeitungstexten entwickelt worden und passen nicht zur Groß- und Kleinschreibung, zur Wortwahl oder zur Zeichensetzung in Tweets.

Im nächsten Schritt muss man Wörterlisten aufstellen, die über die Stichwörter hinausgehen, die bei der Suche nach Tweets verwendet wurden. Es gibt im Internet Listen beispielsweise von beleidigenden Ausdrücken oder von Schimpfwörtern, mit denen man hier arbeiten kann. Es ist aber notwendig zu prüfen, ob diese Listen auch für die Daten passend sind, mit denen man arbeiten möchte, also ob die Hasswörter in den Daten auch in den Listen vorkommen und umgekehrt, ob die Listen nicht zu viel enthalten. Daher gehen viele Forschungsgruppen so vor, dass sie zunächst Hasswörter aus den Trainingsdaten extrahieren. Man kann zum Beispiel alle Wörter, die in als OFFENSIVE markierten Tweets vorkommen, mit den Wörtern, die in als OTHER markierten Tweets vorkommen, vergleichen und die extremen Wörter in eine Liste aufnehmen. Der Vorteil davon ist, dass das Ergebnis gut an die Trainingsdaten angepasst ist. Der Nachteil ist, dass auch ganz unproblematische Wörter in die Liste gelangen können, die eben zufälligerweise nur in den offensiven Tweets vorkommen. Für ein gutes Ergebnis braucht man daher sehr viele annotierte Daten. Wenn man sich von der Idee der Wörter als fundamentaler Einheit der Analyse löst, dann kann man auch mit Bi- oder Trigrammen arbeiten. Das können einerseits Ketten von zwei oder drei Wörtern sein, andererseits aber auch einfach Ketten von Buchstaben und Zeichen. Nehmen wir folgendes Beispiel aus den Trainingsdaten der schon genannten GermEval Shared Task 2018:

»Naja, dein Name sagt schon alles! Dumm und dümmer!!!«

Die Bigramme (Wortebene und Satzzeichen) sind:

{(Naja, ,), (, , dein), (dein, Name), (Name, sagt), (sagt, schon), (schon, alles), (alles, !), (!, Dumm), (Dumm, und), (und, dümmer), (dümmer, !), (!, !), (!, !)}

Die Trigramme sind:

{(Naja, , , dein), (, , dein, Name), (dein, Name, sagt), (Name, sagt, schon), (sagt, schon, alles), (schon, alles, !), (alles, !, Dumm), (!, Dumm, und), (Dumm, und, dümmer), (und, dümmer, !), (dümmer, !, !), (!, !, !)}

Auf Zeichenebene (Trigramm):

{›Naj‹, ›aja‹, ›ja,‹, ›a, ‹, ›, d‹, › de‹, ›dei‹, ›ein‹, ›in ‹, ›n N‹, › Na‹, ›Nam‹, ›ame‹, ›me ‹, ›e s‹, › sa‹, ›sag‹, ›agt‹, ›gt ‹, ›t s‹, › sc‹, ›sch‹, ›cho‹, ›hon‹, ›on ‹, ›n a‹, › al‹, ›all‹, ›lle‹, ›les‹, ›es!‹, ›s! ‹, ›! D‹, › Du‹, ›Dum‹, ›umm‹, ›mm ‹, ›m u‹, › un‹, ›und‹, ›nd ‹, ›d d‹, › dü‹, ›düm‹, ›ümm‹, ›mme‹, ›mer‹, ›er!‹, ›r!!‹, ›!!!‹}

Bei N-Grammen auf Zeichenebene löst man sich komplett von der Idee, dass das Wort die fundamentale Einheit ist, mit der Bedeutung transportiert wird. Experimente haben gezeigt, dass dennoch Systeme zur Klassifikation von Texten damit sehr gute Ergebnisse erzielen können, vor allem wenn wie bei Twitter viele Schreibfehler und Schreibvarianten, aber auch ungewöhnliche Wörter verwendet werden.

Eine andere Möglichkeit ist die Suche nach sogenannten »Word-Embeddings« in nicht annotierten (und daher in großen Mengen verfügbaren) Twitter-Daten. Die Idee dabei ist, dass semantisch ähnliche Wörter in ähnlichen Kontexten stehen. Nehmen wir beispielsweise das Wort »scheiss« (die wenigsten auf Twitter Schreibenden benutzen ein ß). Zunächst wird eine Liste von Wörtern generiert, die zusammen mit diesem Wort auftreten. Wir suchen dann (automatisiert) nach Wörtern, die zusammen mit ähnlichen Wörtern auftreten wie die in unserer Liste, und hoffen dann, dass diese eine ähnliche Semantik haben – die in diesem Beispielfall wohl kaum recht freundlich sein dürfte.

Neben dem Abgleich mit Wortlisten wird häufig ein Werkzeug zur Sentiment-Analyse – also zur Klassifikation von

Meinungsäußerungen als positiv, negativ oder neutral – verwendet. Diese Werkzeuge werden normalerweise eingesetzt, um Produktbewertungen in Verbraucherportalen automatisch zu klassifizieren. Damit erkennt man beispielsweise schnell, ob ein neues Produkt bei den Verbrauchern Anklang findet oder ob es eher kritisch beurteilt wird.

In unserem Kontext suchen wir bis auf wenige Ausnahmen nur nach extrem negativen, nicht nach positiven Äußerungen. Daher ist die Ausgabe der Sentiment-Analyse ein Baustein unter mehreren in der Erkennung extremer Meinungsäußerungen. Die Programme dazu müssen aber auf Twitter-Texte angepasst werden, denn sie sind ja für Produktbewertungen optimiert. Auch hier werden die Wörterbücher wieder angepasst. Da die Trainingsdaten nicht nach ihrem Sentiment annotiert sind, ist es sinnvoll, andere mit Sentiment annotierte Twitter-Daten hinzuzuziehen, um die Qualität der Analyse bewerten zu können.

Mit dem Abgleich eines Tweets mit den erstellten Wörterbüchern sowie mit der Sentiment-Analyse bekommt man Ergebniswerte zu dessen Beurteilung. Wie wird nun die Entscheidung getroffen, ob dieser Tweet eine extreme Meinungsäußerung ist und herausgefiltert werden sollte? Man kann einerseits so vorgehen, dass man mit den Trainingsdaten experimentiert und Schwellwerte herausfindet. Das kann zum Beispiel darauf hinauslaufen, dass Tweets mit einem stark positiven Sentiment nicht als extreme Meinungsäußerungen klassifiziert werden oder dass nur Tweets mit mindestens zwei Übereinstimmungen mit den Wörterbüchern als extrem angesehen werden. Andererseits kann man ein System zum maschinellen Lernen mit diesen Werten für die Trainingsdaten »füttern« und die Schwellwerte automatisch generieren lassen.

Die Testdaten kommen ganz zum Schluss zum Einsatz, um das bisher entwickelte System der automatischen Klassifizierung zu überprüfen. Dies geschieht, indem man die Daten mit dem entwickelten System klassifizieren lässt und das Ergebnis mit der eingangs erstellten Annotation durch

die Testpersonen vergleicht. So kann man sehen, wie präzise der Algorithmus funktioniert.

Die automatische Klassifizierung kann aber letztlich nur Anhaltspunkte geben – die Entscheidung, ob ein Tweet oder eine Äußerung auf einer anderen Plattform aus dem Netz genommen wird, muss ein Mensch treffen. Ein Vergleich mit den Spamfiltern für unsere E-Mails ist hier hilfreich: Es kann immer noch passieren (wenn auch inzwischen selten), dass Nachrichten herausgefiltert werden, die kein Spam sind. Andererseits landen immer wieder auch unerwünschte Werbemails in unserem Postfach.

Um die Trefferquote der automatischen Erkennung zu verbessern, müssen möglichst viele Daten möglichst präzise annotiert werden, so dass die Systeme damit weiterentwickelt werden. Da sind wir bei der deutschen Sprache noch ganz am Anfang. Der Weg zu einem »netten« Internet ist dementsprechend lang und es bleibt wohl eine fromme Hoffnung, dass nicht nur automatisierte Filter auf Dauer für einen guten Umgangston sorgen müssen, sondern auch das umsichtige Verhalten der Internetnutzer selbst.