Supervised Machine Learning mit Nutzergenerierten Inhalten: Oversampling für nicht balancierte Trainingsdaten

Anke Stoll

Online publiziert: 21. April 2020
© Der/die Autor(en) 2020

Zusammenfassung Viele der aktuell im Forschungsbereich Onlinekommunikation untersuchten Phänomene wie Hate Speech, Inzivilität oder Offensive Language kommen in einer Stichprobe aus Nutzergenerierten Inhalten (User Generated Content, UGC) vergleichsweise selten vor. Sind die Kategorien in einer Stichprobe nicht gleich verteilt, spricht man von unbalancierten Daten. Für die Textklassifikation mit Überwachtem Maschinellen Lernen (Supervised Machine Learning) sind solche nicht balancierten Stichproben häufig problematisch, da sie die automatisierte Identifikation der Kategorien erschweren und Klassifikationsmodelle (Classifier) oft ungenau und unzuverlässig werden lassen. Kommt eine Kategorie in den Daten nur selten vor, kann sie durch ein statistisches Klassifikationsmodell nur schwer erlernt werden. Zudem tendieren viele ML-Algorithmen dazu, bei Unsicherheit die vorherrschende Kategorie in den Daten vorherzusagen, und die Klassifikation wird zugunsten der überrepräsentierten Kategorie verzerrt.

Die vorliegende Studie untersucht, inwieweit die Methode des Oversampling die Klassifikation von UGC verbessern kann, wenn eine Kategorie in der Stichprobe deutlich unterrepräsentiert ist. Hierfür wurden anhand von verschiedenen nicht balancierten Stichproben aus deutsch- und englischsprachigen Tweets und Nutzerkommentaren Klassifikationsmodelle für die Identifikation von Offensive Language, Inzivilität und Sentiment trainiert und getestet. Verglichen wurden die Ergebnisse bevor und nachdem die Oversampling-Strategien ROS (Random Over Sampling) und SMOTE (Synthetic Minority Over-sampling Technique) auf den Trainingsdaten angewendet wurden. Die Ergebnisse zeigen, dass sowohl ROS als auch SMOTE die Klassifikation von UGC in allen Stichproben deutlich verbessert, vor allem die Identifikation der unterrepräsentierten Kategorie. Die Anwendung von Oversamp-

A. Stoll (✉)

Institut für Sozialwissenschaften, Kommunikations- und Medienwissenschaft,
Heinrich-Heine-Universität Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Deutschland
E-Mail: anke.stoll@hhu.de

ling führt zudem dazu, dass die Verzerrung der Schätzung zu Gunsten der vorhersehenden Kategorie deutlich reduziert wird. Ziel der Studie ist es, Forschenden aus der Kommunikationswissenschaft Erkenntnisse darüber liefern, wie sich die Problematik von nicht balancierten Stichproben auf die automatisierte Inhaltsanalyse mit Supervised Machine Learning auswirkt und bis zu welchem Punkt diesem Problem mit Oversampling begegnet werden kann.

Schlüsselwörter Textklassifikation · Nutzergenerierte Inhalte · Nicht balancierte Daten · Maschinelles Lernen · Automatisierte Inhaltsanalyse

Supervised machine learning with user generated content: oversampling for imbalanced training data

Abstract In online communication research, Document Classification with Supervised Machine Learning (SML) is increasingly being used to automatically identify the content and meaning of User Generated Content (UGC). Detecting categories such as Incivility or Offensive Language is currently of high relevance in this context, but in comparison to “harmless” content, they occur rarely in many samples of UGC. If categories (e.g. UNCIVIL vs. CIVIL) are unequally distributed in a data set, one speaks of imbalanced (or unbalanced) data. For document classification with SML, imbalanced data can lead to multiple problems. First, if a category is infrequent in the training data, it can hardly be learned by a statistical model. Second, most Machine Learning (ML) algorithms easily become biased towards the overrepresented category, and predictions become inaccurate and unreliable. Such biases often remain undiscovered because they are not immediately apparent from the overall performance measures of a model. For example, an incivility classifier could achieve an accuracy of 90% by only predicting the category CIVIL if there are 90% civil and 10% uncivil instances in the sample. Such a classifier would not have learned anything about detecting incivility in UGC at all. Furthermore, its predictions will be biased towards the major category CIVIL in the data. However, since a classifier will be used to label new data and to answer further research questions, it is particularly important that its predictions are as reliable as possible.

One important strategy to overcome the problem of rare categories and imbalanced data is oversampling. Oversampling means that instances of the underrepresented category are weighted, or that new, synthetic cases of the minor category are generated to balance the sample. In ML research, oversampling is an established technique, which is successfully applied in the field of biotechnology and medical technology, e.g. for the diagnosis of diseases and the identification of genes, or in financial management, e.g. in the detection of credit card fraud. However, there is a lack of studies that investigates, whether oversampling also affects and possibly improves document classification with UGC and whether oversampling can be applied to reduce prediction biases from imbalanced samples of text documents such as tweets or user comments.

The present study investigates how oversampling can improve the identification of the outcome categories Incivility, Offensive language, and Sentiment on three imbalanced samples of UGC, including English and German Tweets and Facebook

user comments ($n = 55,400$). To predict the outcome categories, a logistic regression classifier based on Bag-of-Words n -grams is applied, which is a common baseline approach to document classification. The results are compared before and after the application of two different oversampling strategies on the training data. Random Over Sampling (ROS) randomly selects cases of the underrepresented category and simply weights them for classification to balance the case numbers of all categories. The more complex algorithm Synthetic Minority Over-sampling Technique (SMOTE) uses the k nearest neighbors algorithm to generate new, synthetic cases. The new instances are supposed to be similar to the other cases of the minor category, meaning they should have a small distance to each other in the vector space. Here, SMOTE generated new text documents with a similar frequency distribution of words and word combinations.

Results show that both ROS and SMOTE significantly improve the classification of the outcome categories Offensive language, Incivility and Sentiment. After applying oversampling on the training data, the overall classification results improve by up to 15 percent points (after oversampling: Macro-F1_{Offensive_ROS} = 0.71, Macro-F1_{Incivility_SMOTE} = 0.62, Macro-F1_{Sentiment_ROS} = 0.72, before oversampling: Macro-F1_{Offensive} = 0.56, Macro-F1_{Sentiment} = 0.62, Macro-F1_{Incivility} = 0.46). Before applying oversampling, the classifiers failed to identify many instances of the minor categories (Recall_{UNCIVIL} = 0.01, Recall_{NEGATIVE_SENTI} = 0.26, Recall_{OFFENSIVE} = 0.18) and overestimated the major categories. The findings are stable for all categories and forms of UGC. The results show that by using oversampling, instances with relevant information for a category (e.g., certain words to detect offensive language) can be considered, even if they rarely occur in the original, imbalanced distribution. This way, the bias towards the major categories can be reduced and the overall model performance increases. At the same time, however, all classifier lose precision for the minor categories at the same time, which is because both ROS and SMOTE also consider misleading instances from the training data for weighting or generating new text documents.

In sum, the study shows that oversampling on UGC can lead to significant improvement of the model performance and reduction of the estimation bias caused by imbalanced training data. Oversampling works where relevant features (here: words and word combinations) already exist in the sample, since none of the oversampling algorithms can generate data with new information (here: new words or word combinations to identify a category such as incivility). Therefore, oversampling should be considered as a useful and important tool for classification with UGC. It is likely that these findings can also be transferred to other use cases of document classification in communication science that should be investigated by following studies.

Keywords Document Classification · User Generated Content · Imbalanced Data · Machine Learning · Automated Content Analysis

1 Einleitung

Viele Phänomene, die in der Kommunikationswissenschaft aktuell im Forschungsbereich Onlinediskussionen und Nutzergenerierte Inhalte (User Generated Content, UGC) untersucht werden, kommen in einer Stichprobe vergleichsweise selten vor. Dies gilt z.B. für Hate Speech und Offensive Language auf Twitter (vgl. z.B. Coe et al. 2014; Davidson et al. 2017; Waseem 2016; Waseem und Hovy 2016; Kwok und Wang 2013) oder Inzivilität in Nutzerkommentaren auf Facebook oder Nachrichtenseiten (z.B. Rowe 2015; Stoll et al. 2020; Ziegele et al. 2018). Sind die Kategorien (z.B. inzivil vs. nicht inzivil) in einer Stichprobe ungleich verteilt, spricht man von nicht balancierten (oder unbalancierten) Daten (imbalanced/unbalanced data). Für die Textklassifikation mit Überwachtem Maschinellen Lernen (Document Classification with Supervised Machine Learning) können nicht balancierte Stichproben ein essenzielles Problem darstellen, da sie statistischen Klassifikationsmodellen (Classifier) das Erlernen der unterrepräsentierten Kategorie erschweren können und die Klassifikation dadurch unsicher und unzuverlässig wird (vgl. Haixiang et al. 2017, S. 220). Darüber hinaus kann es leicht passieren, dass ein Klassifikationsmodell bei unzureichenden Informationen Fälle bevorzugt der überrepräsentierten Kategorie zuordnet. Eine solche Verzerrung (Bias) zugunsten der vorherrschenden Kategorie geht aus den allgemeinen Gütekriterien des Schätzmodells nicht gleich hervor. Enthielte eine Stichprobe aus Nutzerkommentaren beispielsweise 10 % inzivile und 90 % nicht inzivile Kommentare, könnte ein Classifier auch dann eine Schätzgenauigkeit (Accuracy) von 90 % erreichen, wenn er ausschließlich die Kategorie „nicht inzivil“ klassifizieren würde. Ein solches augenscheinlich sehr genaues Schätzmodell wäre nicht nur kaum oder gar nicht in der Lage, inzivile Kommentare zu erkennen, sondern würde das Vorkommen inziviler Kommentare in Stichproben auch systematisch unterschätzen.

Das Forschungsfeld des Maschinellen Lernens (Machine Learning, ML) hat sich mit dem Problem von nicht balancierten Stichproben bereits eingehender befasst. Ein bewährter Lösungsansatz ist hier das Resampling der Stichprobe vor dem „Lernen“ durch Oversampling oder Undersampling: Beim Oversampling werden Fälle der unterrepräsentierten Kategorie entweder durch Gewichtung oder durch die Generierung künstlicher Beispiele „aufgestockt“. Beim Undersampling wird die vorherrschende Kategorie an die Fallzahl der unterrepräsentierten Kategorie angepasst, indem (zufällig) Fälle entfernt werden (vgl. Haixiang et al. 2017; Loyola-González et al. 2016; Branco et al. 2016). Da die manuelle Codierung von großen Textmengen aufwendig und kostenintensiv ist und eine größere Stichprobe in der Regel eine validere Messung ermöglicht, ist besonders das Potenzial von Oversampling für die Textklassifikation interessant. Im ML existieren Studien zu Auswirkungen von nicht balancierten Daten und der Anwendung von Oversampling in erster Linie im Bereich Bio- und Medizintechnik, zum Beispiel für die Diagnose von Krankheiten oder die Identifikation von Genen (vgl. z.B. Dubey et al. 2014; Herndon und Caragea 2016) und im Bereich Finanzmanagement, etwa bei der Erkennung von Kreditkartenbetrug (vgl. Zakaryazad und Duman 2016). Die Anwendbarkeit und die Auswirkungen von Oversampling-Techniken auf die Textklassifikation, speziell

von UGC, sind hingegen bisher kaum untersucht worden (vgl. den Überblick bei Haixiang et al. 2017).

Die vorliegende Studie untersucht daher, inwieweit sich Oversampling-Strategien auf die Klassifikation von UGC auswirken, wenn eine Kategorie in der Stichprobe unterrepräsentiert ist. Verglichen werden die zufällige Gewichtung (Vervielfältigung) von Fällen der unterrepräsentierten Kategorie (Random Over Sampling, ROS) und die Ergänzung von Fällen der unterrepräsentierten Kategorie durch synthetische Datengenerierung mit dem Oversampling-Algorithmus SMOTE (Synthetic Minority Over-Sampling Technique). Datengrundlage bilden drei Stichproben aus englisch- und deutschsprachigen Tweets bzw. Facebook-Nutzerkommentaren, die nach den Kategorien Sentiment, Offensive Language bzw. Inzivilität manuell codiert wurden. Auf diese Weise können Auswirkungen der Oversampling-Methoden untereinander wie auch hinsichtlich der verschiedenen Kategorien verglichen werden.

Obwohl sich SML als Ansatz für die Automatisierte Inhaltsanalyse in der Kommunikationswissenschaft etabliert, sind Anwendungsstudien im Fach noch rar. Zwar wird Textklassifikation aktuell auch von Disziplinen wie Informatik, Computerlinguistik oder Data Science untersucht, der Transfer in die Kommunikationswissenschaft ist jedoch oft schwierig, da sich Forschungsinteresse und -schwerpunkt teilweise stark unterscheiden. Ziel der vorliegenden Studie ist es daher, Forschenden aus der Kommunikationswissenschaft Erkenntnisse darüber zu liefern, wie sich das Problem von nicht balancierten Stichproben auf die automatisierte Inhaltsanalyse mit SML allgemein und auf die Klassifikation von UGC im Speziellen auswirken kann und inwiefern diesem Problem mit Oversampling beigegeben werden kann.

Im folgenden Kapitel geht es zunächst um die Methode der Textklassifikation mit SML sowie um das Problem von nicht balancierten Stichproben allgemein und hinsichtlich der besonderen Datengrundlage Text.

2 Supervised Machine Learning mit Textdaten

Allen Ansätzen der automatisierten Analyse von Sprache liegt die Annahme zugrunde, dass sich die Bedeutung einer Aussage letztendlich in maschinenlesbaren Mustern manifestiert. Ziel der Textklassifikation mit Supervised Machine Learning (SML) ist es, einen statistischen Zusammenhang zwischen solchen Mustern in einem Text (unabhängigen Variablen) und der Textkategorie (abhängiger Variable) zu modellieren, um die Textkategorie auf Basis der Textmuster vorherzusagen zu können. Die basale Standardrepräsentation für ein Testdokument ist die Häufigkeitsverteilung von Wörtern (Bag-of-Words-Ansatz). Jedes Wort, das in der Stichprobe vorkommt, ist hierbei eine unabhängige Variable, im ML auch Feature genannt. Die Ausprägungen der unabhängigen Variablen sind die (absoluten oder gewichteten) Häufigkeiten, mit denen ein Wort in einem Textdokument vorkommt (vgl. Jurafsky und Martin 2009, S. 58). Die Grundvoraussetzung für eine zuverlässige Schätzung ist, dass es einen statistischen Zusammenhang zwischen Features und Kategorie gibt, den ein Modell erlernen kann, im Falle der Textklassifikation z.B. einen Zusammenhang zwischen Häufigkeitsverteilungen von Wörtern und der Textkategorie. Die Modellgüte wird darüber evaluiert, inwieweit die vom Modell geschätzten Werte

Abb. 1 Beispielverteilungen von **a** trennscharfen und **b** nicht trennscharfen Kategorien

mit den wahren (manuell codierten) Werten der abhängigen Variable (Textkategorie) übereinstimmen. Hierfür wird die Stichprobe in Trainingsdaten und Testdaten aufgeteilt. An den Trainingsdaten wird das Modell trainiert (gefittet) und an den Testdaten evaluiert (vgl. ausführlich z.B. Aggarwal 2018; Aggarwal und Zhai 2012; Géron 2017; Müller und Guido 2016).

Für die Kategorisierung von Textdokumenten ist der Klassifikationsansatz dort besonders vielversprechend, wo sich die Dokumente aus Kategorie A relativ trennscharf von den Dokumenten aus Kategorie B unterscheiden lassen. Abb. 1a zeigt eine solche ideale Ausgangssituation. In vielen Anwendungsbereichen von SML sind die Kategorien jedoch nicht eindeutig separierbar, sei es durch Rauschen oder weil sich die Kategorien nicht trennscharf hinsichtlich der gewählten Features (z.B. der

Abb. 2 Beispiele für nicht balancierte Daten bei trennscharfen **a** und nicht trennscharfen **b** Kategorien

Häufigkeitsverteilung von Wörtern) unterscheiden lassen. Abb. 1b zeigt ein solches Szenario.

Sind die Kategorien in einer Stichprobe annähernd gleich verteilt (balanciert), wie es in Abb. 1a, b der Fall ist, können Klassifikationsmodelle auch dann zu zufriedenstellenden Ergebnissen gelangen, wenn die Kategorien nicht trennscharf sind. In Abb. 1b z. B. könnte ein Modell noch 80 % Schätzgenauigkeit erreichen, da nur ein geringer Teil der Fälle in den Bereich fällt, wo die Kategorien überlappen. Anders sieht es bei Stichproben aus, in denen die Kategorien *nicht* gleich verteilt, also nicht balanciert, sind. Abb. 2a, b zeigen Beispiele für nicht balancierte Daten. In Abb. 2a lassen sich die Fälle der Kategorien eindeutig separieren, in Abb. 2b überlappen die Fälle.

Die Ausgangslage für die Klassifikation von UGC entspricht in den meisten Fällen dem Szenario aus Abb. 2b, da Phänomene wie Hate Speech oder Inzivilität zum einen vergleichsweise selten vorkommen und zum anderen durch die Textebene allein nicht eindeutig definierbar sind.

Viele gängige Klassifikationsalgorithmen wie Support Vector Machines, Logistische Regression oder Entscheidungsbäume werden bei einer solchen Ausgangslage wie in Abb. 2b unzuverlässig und schätzen bei Unsicherheit tendenziell die vorherrschende Kategorie in den Trainingsdaten (vgl. Haixiang et al. 2017, S. 221; López et al. 2013, S. 121–137). Waseem und Hovy (2016) fanden in ihrer Stichprobe aus Tweets lediglich 11 bzw. 22 % relevanter Fälle für die Kategorien Rassismus bzw. Sexismus. In einer Erweiterung der Studie durch Wasseem (2016), der die Codierung von Experten und Amateuren verglich, wurden sogar nur 1 bis 6 % der Fälle im Datensatz der Kategorie Rassismus und 13 bis 19 % der Kategorie Sexismus zugeordnet. Dies führte dazu, dass viele Modelle vor allem oder ausschließlich die vorherrschende Kategorie „weder noch“ (neither) klassifizierten. Auf diesem Wege kann ein Classifier selbst dann hohe Schätzgenauigkeiten erzielen, wenn die Identifikation der unterrepräsentierten Kategorie fehlerhaft oder gar nicht gelungen ist. Auch Davidson et al. (2017) zeigten in ihrer Studie zur Klassifikation von Offensive Language und Hate Speech auf Twitter, dass die automatisierte Erkennung der seltenen Kategorie Hate Speech deutlich ungenauer ausfiel als die der vorherrschenden Kategorien.

In der vorliegenden Studie soll untersucht werden, ob sich durch die Anwendung von Oversampling die Klassifikation von unterrepräsentierten Kategorien für den Forschungsgegenstand UGC verbessern lässt. Im nächsten Kapitel wird auf die im SML etablierten Oversampling-Strategien ROS und SMOTE näher eingegangen, die in dieser Studie verwendet werden.

3 Oversampling-Strategien im Supervised Machine Learning

Im ML stellen nicht balancierte Trainingsdaten ein essenzielles Problem bei der Entwicklung von Classifiern dar. Dass bestimmte Ereignisse und Phänomene nur selten vorkommen, erschwert deren Identifikation mit statistischen Modellen (vgl. Denil und Trappenberg 2010, S. 220–221; Haixiang et al. 2017, S. 220–221). Ein wichtiger Ansatz ist hier das Resampling der Trainingsdaten durch Undersampling oder Oversampling. Beim Undersampling werden Fälle der vorherrschenden Kategorie (zufällig) entfernt, um sie an die Fallzahl der unterrepräsentierten Kategorie anzugleichen (vgl. López et al. 2013, S. 117–118; Chawla et al. 2002, S. 321–322). Undersampling ist für viele Fragestellungen nachteilig, da auf diese Weise Informationen über die vorherrschende Kategorie verloren gehen. Beim Oversampling hingegen wird die Fallzahl der unterrepräsentierten an die der überrepräsentierten Kategorie angeglichen. Die simpelste Herangehensweise ist hier die Gewichtung einzelner Fälle der unterrepräsentierten Kategorie (Radom Over Sampling, ROS). Hier werden einfach Fälle der unterrepräsentierten Kategorie zufällig gezogen und vervielfacht. Komplexere Algorithmen hingegen generieren neue, synthetische Daten für die unterrepräsentierte Kategorie. Ein im ML populärer Algorithmus ist

SMOTE (Synthetic Minority Over-Sampling Technique), der in unterschiedlichen Anwendungsbereichen bereits erfolgreich eingesetzt wird (vgl. Chawla et al. 2002; Haixiang et al. 2017; Chawla 2009). SMOTE erzeugt neue, künstliche Trainingsbeispiele für die unterrepräsentierte Kategorie auf Basis des k -Nearest-Neighbors-Algorithmus (k -nächste-Nachbarn-Algorithmus). Der nächste Nachbar x_{zi} eines Falles x_i aus der unterrepräsentierten Kategorie ist der nächstgelegene Fall aus dem Trainingsdaten, also der mit der geringsten Distanz zu x_i . Die Distanz zwischen den Datenpunkten wird hier als Ähnlichkeit verstanden. Der nächste Nachbar eines Kommentars, der durch Wortverteilungen repräsentiert wird, wäre demnach der Kommentar in dem die gleichen Wörter möglichst gleich oft vorkommen. Der neue Fall, den SMOTE generiert, befindet sich auf der Strecke zwischen dem jeweiligen Kommentar x_i und seinem nächsten Nachbarn x_{zi} . Sowohl der Nachbar als auch der Punkt auf der Strecke werden zufällig gewählt. Der Parameter k ist vorab festzulegen und definiert die Anzahl der Nachbarn, die für die Generierung des neuen Falls berücksichtigt werden sollen (vgl. Chawla et al. 2002, S. 328–330; Lemaître et al. 2017, S. 2–3). Für die Generierung der neuen Fälle kann SMOTE nur die bereits vorhandenen Informationen aus den Trainingsdaten nutzen. Neue Informationen, z.B. Kommentare mit anderen Wörtern, die nicht bereits in den Daten vorhanden sind, können auf diese Weise nicht erzeugt werden.

Darüber, ob Oversampling-Strategien wie ROS und SMOTE auch die Klassifikation von UGC beeinflussen oder verbessern können, existiert bisher kaum Forschung. Wo entscheidende Informationen (z.B. Wörter oder Wortkombinationen) für die Vorhersage der Textkategorien in den Daten fehlen, würde auch Oversampling die Ergebnisse nicht verbessern, da weder die zufällige Gewichtung noch die Generierung von Fällen dazu in der Lage ist, Features hinzuzufügen. Darüber hinaus könnte Oversampling die Ergebnisse sogar verschlechtern, sollten Fälle gewichtet bzw. neu generiert werden, die zu Fehlklassifikation führen. Dies ist vor allem dort denkbar, wo sich die zu klassifizierenden Kategorien sehr stark überlappen, also sich z.B. durch Wortverteilungen nicht trennscharf separieren lassen. Die vorliegende Studie untersucht daher, inwieweit sich die Anwendung von Oversampling auf die Klassifikation von nicht balancierten Stichproben aus UGC auswirkt.

4 Methode und Untersuchungsdesign

Für die Analyse werden Klassifikationsmodelle an verschiedenen Stichproben aus UGC trainiert und evaluiert und die Klassifikationsergebnisse auf Basis der ursprünglichen, nicht balancierten Verteilung mit den Ergebnissen nach dem Oversampling der Trainingsdaten verglichen.

4.1 Samples

Es werden verschiedene nicht balancierte Stichproben aus UGC verwendet, die nach verschiedenen Kategorien manuell codiert worden sind: ein Datensatz aus englischsprachigen Tweets, der nach Sentiment codiert ist, ein Datensatz aus deutschspra-

chigen Tweets, der nach Offensive Language codiert ist, und ein Datensatz aus deutschsprachigen Facebook-Nutzerkommentaren, der nach Inzivilität codiert ist.

Sentiment-Datensatz: Der größte Datensatz (im Folgenden Sentiment-DF) enthält ca. 49.000 englischsprachige Tweets, die nach Sentiment (NEUTRAL, POSITIVE, NEGATIVE) manuell codiert wurden. Der Datensatz wurde im Rahmen der internationalen Shared Task „SemEval-2017 Sentiment Analysis in Twitter“ (Rosenthal et al. 2017) der Association for Computational Linguistics erhoben und zur Verfügung gestellt. Die Codierung erfolgte über die Plattform CrowdFlower mit simplen Codieranweisungen und Beispielen für jede Kategorie.¹ Jeder Tweet wurde von fünf Personen codiert. Details über Reliabilität und Datenqualität sind nicht bekannt, jedoch berichten die Autoren von mehreren Feedback-Schritten aus manueller Qualitätskontrolle und Anpassungen der Codieranweisungen, bis die Datenqualität überzeugte (vgl. Rosenthal et al. 2017, S. 504). Für die Vergleichbarkeit der Datensätze untereinander und für die bessere Verständlichkeit werden nur die zwei Kategorien NEUTRAL und NEGATIVE berücksichtigt, die im Datensatz sehr schief verteilt sind. Insgesamt enthält die Stichprobe 22.182 Tweets der Kategorie NEUTRAL und 7713 der Kategorie NEGATIVE.

Offensive-Datensatz: Der zweite Datensatz (Offensive-DF) enthält 15.418 deutschsprachige Tweets, die nach Offensive Language manuell codiert wurden. Der Datensatz wurde ebenso wie der Sentiment-DF im Rahmen einer Shared Task, der „Germeval Task 2, 2019 – Shared Task on the Identification of Offensive Language“,² erstellt und zur Verfügung gestellt (vgl. Struß et al. 2019). Die manuelle Codierung erfolgte durch die Organisatorinnen und Organisatoren der Shared Task über ein simples Codierschema.³ Für die Studien werden die Kategorien OFFENSIVE und OTHER (nicht offensive) berücksichtigt. Insgesamt enthält der Datensatz 10.328 Tweets der Kategorie OTHER und 5090 Tweets der Kategorie OFFENSIVE.

Inzivilität-Datensatz: Der dritte Datensatz (Inzivilität-DF) enthält ca. 10.000 manuell codierte Facebook-Nutzerkommentare, die im Rahmen einer standardisierten manuellen Inhaltsanalyse erhoben und von sechs geschulten Personen codiert wurden (vgl. Stoll et al. 2020, S. 116–117). Kommentare wurden als inzivil codiert, wenn sie negative Stereotypen oder politischen Extremismus enthielten oder demokratische Rechte, Normen und Werte bedrohten (vgl. Papacharissi 2004). Die Inter-coder-Reliabilität liegt bei einem akzeptablen Niveau von Krippendorffs $\alpha = 0,73$ auf 100 Kommentaren. Insgesamt wurden 8096 Kommentare der Kategorie NICHT INCIVIL und 2068 Kommentare der Kategorie INZIVIL zugeordnet.

Für alle drei beschriebenen Stichproben werden zunächst jeweils Klassifikationsmodelle auf der ursprünglichen, nicht balancierten Verteilung trainiert und getestet.

¹ Zum vollständigen Codierschema vgl. Nakov et al. (2019). Datensätze abrufbar unter: <http://alt.qcri.org/semeval2017/task4/index.php?id=results>.

² Die Shared Task wird unterstützt von der Interest Group on German Sentiment Analysis (IGGSA) und der Interest Group on Social Media Analysis (GSCL) der German Society for Computational Linguistics and Language Technology.

³ Zum vollständigen Codierschema vgl. Struß et al. (2019). Datensätze abrufbar unter: <https://projects.fzai.h-da.de/iggss/data-2019/>.

Anschließend werden die Oversampling-Strategien ROS und SMOTE angewendet und die Modelle jeweils erneut trainiert und evaluiert. Auf diese Weise lassen sich die Auswirkungen der Oversampling-Strategien sowohl für jeden Datensatz vergleichen als auch für die Datensätze untereinander.

4.2 Modell und Evaluation

Im SML erfolgt das Training (oder „Lernen“) des Klassifikationsmodells nur an einem Teil der Stichprobe, den Trainingsdaten. Wie gut das Modell die Kategorien tatsächlich vorherzusagen gelernt hat, zeigt sich auf dem separaten Testdatensatz. Die Modellgüte bemisst sich daraus, wie gut die geschätzten Werte des Modells mit den wahren (manuell codierten) Werten auf den Testdaten übereinstimmen. Als Features (unabhängige Variablen) werden für diese Studie die (gewichteten) Häufigkeiten von Einzelwörtern und Kombinationen aus zwei und drei Wörtern (Unigramme, Bigramme, Trigramme) verwendet (vgl. z. B. Gaydhani et al. 2018; Schonlau und Guenther 2016; Stoll et al. 2020). Eine Verarbeitung der Textdokumente vorab (Preprocessing) kann für die Vorhersage irrelevante Varianz in den Daten verringern. Hierzu gehört das Entfernen nicht informativer Wörter, sogenannte Stop Words,⁴ die in einer Sprache besonders häufig vorkommen und daher als eher „bedeutungsarm“ gelten (vgl. Müller und Guido 2016, S. 341–342). Oft ist es sinnvoll, Wörter und Zeichen auf eine gemeinsame Zeichenkette zu normalisieren, von denen im Einzelnen kein entscheidender Beitrag für die Vorhersage der Textkategorien zu erwarten ist. Für die hier verwendete Datengrundlage aus UGC werden alle Nutzerverlinkungen zu der Zeichenkette `@USER`, Web-Verlinkungen zu `URL`, freistehende Zahlen zu „Nummer“ sowie sehr lange Folgen von Buchstaben oder Satzzeichen zu maximal drei aufeinanderfolgenden Zeichen (z. B. `waaaaaas` `????` und `waaaaaaaaaaaaaaaaaaaaaas` `?????` zu `waaaas` `????`)⁵ normalisiert. Für die Vorhersage der Textkategorien ist es außerdem sinnvoll, die Smileys aus Satzzeichen in die entsprechende Emotion umzu codieren (z. B. `:(` und `:-)` und `:` zu `sad`), damit diese als Features erhalten bleiben.

Als Schätzmodell wird die Logistische Regression⁶ verwendet, die im SML ein etablierter Baseline-Algorithmus für Textklassifikationsprobleme und auch in der Inferenzstatistik etabliert ist (vgl. Wiegand et al. 2018; Risch et al. 2018). Für das Training und das Testen der Modelle wurden die Stichproben jeweils einmal in 75 % Trainingsdaten und 25 % Testdaten gesplittet (vgl. ausführlich Raschka und Mirjalili 2017). Die Evaluation der Modelle erfolgt durch den Abgleich der geschätzten mit den wahren Werten im Testset. Ein einfaches Übereinstimmungsmaß, das viele Studien als Standardmaß mit angeben, ist die Accuracy (Genauigkeit). Die Accuracy kann Werte zwischen 0 und 100 (bzw. 0 bis 100 %) annehmen und gibt an, wie oft das Modell bei der Vorhersage richtigliegt. Gängige Maße für die Modellleistung für jeweils eine der Kategorien sind Recall, Precision und F1-Score. Der Recall gibt an,

⁴ Hier verwendet: Stop Word Corpus ISO (Deutsch) ist eine Sammlung vieler verschiedener Stop Word Corpora für das Deutsche. Dokumentation unter <https://github.com/stopwords-iso/stopwords-iso/blob/master/CREDITS.md>.

⁵ Hier verwendet: `class nltk.tokenize.casual.TweetTokenizer`; vgl. Bird et al. (2009), Python 3.6.

⁶ Hier verwendet: `class sklearn.linear_model.LogisticRegression`; vgl. Pedregosa et al. (2011), Python 3.6.

Tab. 1 Klassifikationsergebnisse für Sentiment

	Kategorie	Precision	Recall	F1-Score	Macro-F1-Score	Accuracy
Nicht balanciert	<i>NEGATIVE</i>	0,69	0,26	0,38	0,62	0,78
	<i>NEUTRAL</i>	0,79	0,96	0,87		
Random Over Sampling	<i>NEGATIVE</i>	0,54	0,65	0,59	0,72	0,77
	<i>NEUTRAL</i>	0,87	0,81	0,84		
SMOTE $k=5$	<i>NEGATIVE</i>	0,54	0,66	0,59	0,71	0,77
	<i>NEUTRAL</i>	0,87	0,80	0,84		

Anmerkung: $N_{\text{Test_NEUTRAL}} = 1914$, $N_{\text{Test_NEGATIVE}} = 556$; Logistic-Regression-Classifer: *penalty* = „l2“, *solver* = „warn“

wie viele Fälle aus einer Kategorie das Modell erkannt hat (von 0 bis 100 %). Die Precision gibt an, wie viele Fälle einer Kategorie das Modell *richtig* erkannt hat (von 0 bis 100 %). Aus Recall und Precision lässt sich der F1-Score berechnen, der eine Art balanciertes Maß aus beiden Metriken darstellt und daher ein guter Indikator für die Schätzleistung des Modells für die jeweilige Kategorie darstellt. Neben der Accuracy geben viele Studien den Macro-F1-Score an, der ein Durchschnittswert aus den F1-Scores aller Kategorien ist und ein aussagekräftigeres Gesamtmaß für die allgemeine Schätzleistung eines Klassifikationsmodells ist (vgl. ausführlich z. B. Powers 2011; Gerón 2017).

5 Ergebnisse

Die folgenden Tabellen zeigen für jeden der drei Datensätze (Sentiment-DF, Offensive-DF und Inzivilität-DF) die Ergebnisse auf Basis der nicht balancierten Ursprungsverteilung sowie die Ergebnisse nach der Anwendung von ROS bzw. SMOTE auf den Trainingsdaten. Alle Werte beziehen sich auf die Schätzung auf dem Testset.

Tab. 1 zeigt die Klassifikationsergebnisse für den Sentiment-Datensatz aus englischsprachigen Tweets.

Es ist zu sehen, dass die unterrepräsentierte Kategorie *NEGATIVE* im Schnitt schlechter klassifiziert wird, jedoch nach dem Oversampling der Trainingsdaten sowohl mit ROS als auch mit SMOTE deutlich besser erkannt wird. Das Modell, das auf den nicht balancierten Daten trainiert wurde, findet lediglich 26 % der negativen, aber 96 % der neutralen Tweets in den Testdaten ($\text{Recall}_{\text{NEGATIVE}} = 0,26$, $\text{Recall}_{\text{NEUTRAL}} = 0,96$). Sowohl ROS als auch SMOTE verbessern den Recall für die unterrepräsentierte Kategorie um fast 40 % ($\text{Recall}_{\text{NEGATIVE_ROS}} = 0,65$, $\text{Recall}_{\text{NEGATIVE_SMOTE}} = 0,66$). Die Schätzleistung für die überrepräsentierte Kategorie *NEUTRAL* wird durch das Oversampling ebenfalls beeinflusst. Bei der Vorhersage von neutralen Tweets liegt das Modell öfter richtig ($\text{Precision}_{\text{NEUTRAL_ROS}}$, $\text{Precision}_{\text{NEUTRAL_SMOTE}} = 0,87$). Jedoch werden nur noch ca. 80 % anstelle von 96 % der Fälle der Kategorie *NEUTRAL* in den Testdaten erkannt ($\text{Recall}_{\text{NEUTRAL_ROS}} = 0,81$, $\text{Recall}_{\text{NEUTRAL_SMOTE}} = 0,80$, $\text{Recall}_{\text{NEUTRAL}} = 0,96$). Insgesamt hat das Oversampling die Klassifikation der unterrepräsentierten Kategorie *NEGATIVE* deutlich verbessert. Die Vorhersage der vorherrschenden Kategorie *NEUTRAL* wurde genauer, jedoch

Tab. 2 Klassifikationsergebnisse für Offensive Language in deutschsprachigen Tweets

	Kategorie	Precision	Recall	F1-Score	Macro-F1-Score	Accuracy
Nicht balanciert	<i>OFFENSIVE</i>	0,86	0,18	0,29	0,56	0,72
	<i>OTHER</i>	0,71	0,99	0,82		
Random Over Sampling	<i>OFFENSIVE</i>	0,64	0,57	0,60	0,71	0,75
	<i>OTHER</i>	0,80	0,84	0,82		
SMOTE $k=5$	<i>OFFENSIVE</i>	0,54	0,65	0,59	0,68	0,70
	<i>OTHER</i>	0,81	0,73	0,76		

$N_{\text{Test_OTHER}} = 1745$, $N_{\text{Test_OFFENSIVE}} = 858$; Logistic-Regression-Classifer: *penalty* = „l2“, *solver* = „warn“.

konnte das Modell weniger neutrale Tweets entdecken. Die Ergebnisse von ROS und SMOTE entscheiden sich hier nur unwesentlich.

Tab. 2 zeigt die Klassifikationsergebnisse für den Offensive-Datensatz aus deutschsprachigen Tweets mit den Kategorien *OFFENSIVE* und *OTHER* (nicht offensive).

Wie auch für den Sentiment-DF fällt die Klassifikationsleistung auf den nicht balancierten Daten insgesamt deutlich schlechter aus als nach dem Oversampling, besonders die Vorhersage der unterrepräsentierten Kategorie *OFFENSIVE*. Ohne Oversampling werden nur 18 % der Tweets der Kategorie *OFFENSIVE* erkannt, jedoch fast alle Tweets der Kategorie *OTHER* ($\text{Recall}_{\text{OFFENSIVE}} = 0,18$, $\text{Recall}_{\text{OTHER}} = 0,99$). Anders als beim Sentiment-DF unterscheiden sich die Ergebnisse der Oversampling-Strategien ROS und SMOTE. Mit ROS werden 57 anstelle von nur 18 % der Fälle der Kategorie *OFFENSIVE* erkannt, mit SMOTE sind es 65 % ($\text{Recall}_{\text{OFFENSIVE_ROS}} = 0,57$, $\text{Recall}_{\text{OFFENSIVE_SMOTE}} = 0,65$). Allerdings wird mit ROS die Schätzung genauer als mit SMOTE ($\text{Precision}_{\text{OFFENSIVE_ROS}} = 0,65$, $\text{Precision}_{\text{OFFENSIVE_SMOTE}} = 0,54$). Zudem beeinflussen SMOTE und ROS die Schätzung der überrepräsentierten Kategorie *OTHER* auf unterschiedliche Weise. SMOTE verschlechtert den Recall für die Kategorie *OTHER* um ca. 26 % ($\text{Recall}_{\text{OTHER_SMOTE}} = 0,73$), ROS hingegen um ca. 15 % ($\text{Recall}_{\text{OTHER_ROS}} = 0,84$). Ähnlich wie auf dem Sentiment-DF verbessert Oversampling die Klassifikation von Offensive Language

Tab. 3 Klassifikationsergebnisse für Inzivilität für deutschsprachige Facebookkommentare

	Kategorie	Precision	Recall	F1-Score	Macro-F1-Score	Accuracy
Nicht balanciert	<i>INZIVIL</i>	0,80	0,01	0,03	0,46	0,79
	<i>NICHT INZIVIL</i>	0,79	1,00	0,88		
Random Over Sampling	<i>INZIVIL</i>	0,45	0,32	0,38	0,62	0,77
	<i>NICHT INZIVIL</i>	0,83	0,89	0,86		
SMOTE $k=5$	<i>INZIVIL</i>	0,46	0,33	0,38	0,62	0,77
	<i>NICHT INZIVIL</i>	0,83	0,89	0,86		

$N_{\text{Test_INZIVIL}} = 537$, $N_{\text{Test_NICHT-INZIVIL}} = 1985$; Logistic-Regression-Classifer: *penalty* = „l2“, *solver* = „warn“

insgesamt jedoch deutlich, besonders für die unterrepräsentierte Kategorie. Die beste Klassifikation wird mit ROS erreicht ($\text{Macro-F1}_{\text{ROS}} = 0,71$).

Tab. 3 zeigt die Klassifikationsergebnisse für den Inzivil-Datensatz aus deutschsprachigen Facebook-Nutzerkommentaren mit den Kategorien *INZIVIL* und *NICHT INZIVIL*.

Am Inzivilität-DF zeigt sich am deutlichsten das unerwünschte Klassifikationsmuster auf nicht balancierten Daten: Das Modell findet alle Fälle der überrepräsentierten ($\text{Recall}_{\text{NICHT-INZIVIL}} = 1,0$) und praktisch keinen der unterrepräsentierten Kategorie ($\text{Recall}_{\text{INZIVIL}} = 0,01$). Dennoch wird so eine Schätzgenauigkeit von 79 % erreicht ($\text{Accuracy}_{\text{Inzivilität}} = 0,79$), was auf den ersten Blick auf ein zuverlässiges Modell schließen lassen könnte. Wie auch in den Datensätzen Sentiment-DF und Offensive-DF wird mit Oversampling der Trainingsdaten eine bessere Schätzung der unterrepräsentierten Kategorie erreicht. Dennoch fällt die Klassifikation der Kategorie *INZIVIL* auch mit Oversampling noch deutlich schlechter aus als die der Kategorie *NICHT INZIVIL* ($\text{F1}_{\text{INZIVIL_ROS}}, \text{F1}_{\text{INZIVIL_SMOTE}} = 0,38$, $\text{F1}_{\text{NICHT-INZIVIL_ROS}}, \text{F1}_{\text{INZIVIL_SMOTE}} = 0,86$). Die Ergebnisse für ROS und SMOTE unterscheiden sich insgesamt kaum. Nach dem Oversampling entdeckt das Modell ca. 33 % der inzivilen Kommentare ($\text{Recall}_{\text{INZIVIL_ROS}} = 0,32$, $\text{Recall}_{\text{INZIVIL_SMOTE}} = 0,33$). Auch fällt auf, dass durch das Oversampling die Precision der Kategorie *INZIVIL* um ca. 35 % abfällt ($\text{Precision}_{\text{INZIVIL}} = 0,80$, $\text{Precision}_{\text{INZIVIL_ROS}} = 0,45$, $\text{Precision}_{\text{INZIVIL_SMOTE}} = 0,46$). Das zeigt, dass Oversampling das Modell insgesamt zwar deutlich verbessert ($\text{Macro-F1}_{\text{ROS}} = 0,62$, $\text{Macro-F1}_{\text{SMOTE}} = 0,62$ im Gegensatz zu $\text{Macro-F1} = 0,46$), die unterrepräsentierte Kategorie jedoch nach wie vor sehr ungenau vorhergesagt wird ($\text{F1}_{\text{INZIVIL_ROS}} = 0,38$, $\text{F1}_{\text{INZIVIL_SMOTE}} = 0,38$).

Die Ergebnisse der Studie zeigen, dass Oversampling die Klassifikation von UGC in allen drei Datensätzen deutlich verbessert, vor allem die Klassifikation der unterrepräsentierten Kategorien. Zudem konnte Oversampling den Bias zu Gunsten der vorherrschenden Kategorien in den Trainingsdaten reduzieren. Dies zeigte sich am stärksten auf dem Sentiment-Datensatz. Für die Klassifikation von Sentiment und Inzivilität gibt es kaum Unterschiede zwischen ROS und SMOTE. Für die Klassifikation von Offensive Language führt ROS insgesamt zu besseren Ergebnissen. Insgesamt wurden Sentiment und Offensive Language in Tweets deutlich besser erkannt als Inzivilität in Facebook-Nutzerkommentaren ($\text{Macro-F1}_{\text{Sentiment_ROS}} = 0,72$; $\text{Macro-F1}_{\text{OffensiveLanguage_ROS}} = 0,71$, $\text{Macro-F1}_{\text{Inzivilität_SMOTE}} = 0,62$).

6 Diskussion

Viele Fragestellungen im Bereich UGC beschäftigen sich mit Phänomenen, die in einer Stichprobe vergleichsweise selten vorkommen. Für die Textklassifikation mit Supervised Machine Learning (SML) können solche nicht balancierte Stichproben ein Problem darstellen, da sie die automatisierte Identifikation der Kategorie erschweren und Klassifikationsmodelle somit ungenau und unzuverlässig werden können. Die vorliegende Studie zeigt, dass die Methode des Oversampling die Klassifikation von UGC auf Basis von Wortverteilungen deutlich verbessern kann, wenn eine Kategorie in der Stichprobe unterrepräsentiert ist. Für unterschiedliche Stich-

proben aus Tweets und Nutzerkommentaren wurden die Klassifikationsergebnisse für Sentiment, Offensive Language und Inzivilität vor und nach der Anwendung von Oversampling verglichen. Sowohl die zufällige Vervielfältigung von Fällen (Random Over Sampling, ROS) als auch die Generierung synthetischer Fälle durch den Oversampling-Algorithmus SMOTE (Synthetic Minority Over-Sampling Technique) hat die Identifikation der unterrepräsentierten Textkategorien und die Modellleistung insgesamt verbessern und der systematischen Überschätzung der vorherrschenden Kategorie in den Daten entgegenwirken können. Die Klassifikationsergebnisse für die vorherrschende Kategorie sind nach dem Oversampling jedoch ungenauer ausgefallen.

Für die Studie wurden zwei unterschiedliche Oversampling-Techniken angewendet, die sich in anderen Anwendungs- und Forschungsbereichen von Machine Learning (ML) bereits bewährt haben, deren Anwendbarkeit und Auswirkungen auf die Textklassifikation von UGC jedoch noch weitestgehend unerforscht ist. ROS ist ein vergleichsweise simples Verfahren, um die Verteilung der Kategorien in den Trainingsdaten anzugleichen. Hierfür werden Fälle der unterrepräsentierten Kategorie einfach zufällig ausgewählt und vervielfältigt. SMOTE generiert hingegen künstliche Fälle, die den Fällen der unterrepräsentierten Kategorie möglichst ähnlich sein sollen. Hierfür nutzt SMOTE den k-Nearest-Neighbors-Algorithmus, der unter Ähnlichkeit eine geringe Distanz zwischen Datenpunkten im Vektorraum versteht. Zwei Kommentare sind sich demnach ähnlich, wenn sie die gleichen Wörter und Wortkombinationen ähnlich oft enthalten. Obwohl die Vorgehensweise von SMOTE raffinierter ist als die simple Gewichtung einzelner Fälle, zeigen die Ergebnisse, dass SMOTE im Schnitt keine besseren Ergebnisse erzielt. Dass beide Methoden die Klassifikationsleistung ähnlich gut verbessern, lässt darauf schließen, dass bereits „gute“ Features in den Trainingsdaten vorhanden waren, in der ursprünglichen Verteilung lediglich nicht oft genug vorkamen. Ein menschlicher Codierer oder eine menschliche Codiererin könnte ein Schimpfwort beispielsweise auch dann als guten Indikator für die Kategorie Inzivilität oder Offensive Language erkennen, wenn es nur ein einziges Mal in der Stichprobe vorkäme. Für statistische Classifier trifft das meist nicht zu. Um für die Klassifikation berücksichtigt zu werden, muss ein Wort sowohl in den Trainingsdaten als auch in den Testdaten vorkommen, und zwar möglichst ausschließlich in einer Kategorie. Die Ergebnisse lassen darauf schließen, dass einige entscheidende Wörter und Wortkombinationen in allen Stichproben bereits vorhanden waren, jedoch nicht ausreichend oft vorkamen. Hier kann die Anwendung von Oversampling die Klassifikation verbessern, und zwar offenbar schon durch die simple Vervielfältigung zufälliger Kommentare.

Wo entscheidende Informationen für die Erkennung von Textkategorien fehlen, kann Oversampling die Klassifikation jedoch nicht verbessern, da weder SMOTE noch ROS neue, „bessere“ Features (hier Wörter und Wortkombinationen) generieren kann. Dies spiegelt sich auch im Vergleich der Datensätze untereinander wider, der zeigt, dass der Erfolg von Oversampling je nach Datengrundlage schwanken kann. Die naheliegende Erklärung hierfür ist, dass dort, wo die Fälle besonders stark überlappen, Oversampling zwangsläufig auch mehr „schlechte“ Fälle vervielfältigt. Bereits auf der ursprünglichen nicht balancierten Verteilung ist die Klassifikation von Offensive Language und Sentiment besser gelungen als die Klassifikation

von Inzivilität. Diese Differenz bleibt auch mit der Anwendung von Oversampling bestehen. Auch hat Oversampling in allen Stichproben dazu geführt, dass sich die Precision der Schätzung verschlechtert hat, das Modell bei der Klassifikation der seltenen Kategorie also öfter falsch lag. Diese Ergebnisse lassen ebenfalls den Schluss zu, dass Oversampling bei sich überlappenden Fällen auch zu Fehlklassifikationen beiträgt und dass die Verbesserungspotenzial dort besonders hoch ist, wo bereits „gute“ Features für die jeweiligen Kategorien in den Daten vorhanden sind. Da das Potential von Oversampling im ML bereits ein wichtiger Teilbereich ist, existieren für SMOTE beispielsweise noch komplexere Erweiterungen und Varianten, die Informationen aus den Trainingsdaten nutzen, um gezielter Trainingsbeispiele für die Generierung auszuwählen (z. B. SVM-SMOTE oder Borderline-SMOTE, vgl. Nguyen et al. 2009; Han et al. 2005). Die hier angewendeten Strategien können daher nur exemplarisch für die verschiedenen Oversampling-Algorithmen sein, und es ist sicherlich zu empfehlen, unterschiedliche Methoden für das eigene Forschungsvorhaben auszuprobieren. Wie auch die Ergebnisse der Studie bestätigen, kann der Erfolg der Methode je nach Datengrundlage variieren.

Über das Oversampling hinaus bestätigt die vorliegende Studie, dass sich die hier untersuchten Konzepte Sentiment, Offensive Language und Inzivilität nicht vollständig über die Verteilung von Einzelwörtern, Bigrammen und Trigrammen identifizieren lassen (vgl. z. B. Burnap und Williams 2015; Stoll et al. 2020; Risch et al. 2018). Die Mehrdeutigkeit (Ambiguität) von Wörtern, Wortgruppen oder Phrasen in Abhängig des Kontextes, in dem sie verwendet oder rezipiert werden, ist eine Hauptursache für die Fehlklassifikation von Texten. Viele aktuelle SML-Ansätze bei der Textklassifikation verwenden daher anstelle der Bag-of-Words-Repräsentation für Textdokumente alternative Features, wie z. B. Word Embeddings (vgl. Mikolov et al. 2013; Peters et al. 2018) oder nutzen leistungsstärkere Deep-Learning-Modelle auf sehr viel größeren Datensätzen (z. B. Zhang et al. 2018; Nakov et al. 2019; Ruppenhofer et al. 2018; Struß et al. 2019). Insofern soll die hier vorliegende Studie weniger Ansätze und Modelle für Klassifikation von UGC präsentieren, als die Problematik von nicht balancierten Stichproben für die automatisierte Inhaltsanalyse mit Supervised Machine Learning verdeutlichen. Die hier verwendete Datengrundlage konnte Oversampling für verschiedene Textkategorien, Plattformen, Sprachen und der Art der Codierung vergleichbar machen, jedoch kann die Identifikation der einzelnen Konzepte Sentiment, Offensive Language und Inzivilität mit anspruchsvolleren Klassifikationsansätzen sicherlich besser gelingen.

7 Fazit

Schätzmodelle auf Basis von Textdaten zu erstellen bringt Stichprobenprobleme mit sich, die in der Inferenzstatistik oft für Überraschungen sorgen. Die im besonderen Maße heterogene Datengrundlage Text führt dazu, dass sich Muster oft erst in sehr großen Stichproben manifestieren, die weit über die in der Kommunikationswissenschaft üblichen Größenordnungen hinausgehen. Im Machine Learning (ML) ist es darüber hinaus durchaus üblich, die Ergebnisse eines Schätzmodells durch Gewichtung oder Generierung von künstlichen Fällen zu beeinflussen. Gera-

de in Stichproben aus Textdokumenten fehlen oft nicht nur relevante Informationen für die automatisierte Identifikation eines bestimmten Phänomens, Inhalts oder einer Textkategorie, sondern viele potenziell nützliche Informationen bleiben auch „statistisch unsichtbar“. Die vorliegende Studie konnte zeigen, dass Oversampling diesem Problem bei der Textklassifikation mit Supervised Machine Learning (SML) entgegenwirken kann, wenn eine Textkategorie vergleichsweise selten vorkommt, die Stichprobe also unbalanciert ist. Nicht balancierte Stichproben sind auch deshalb ein Schwachpunkt vieler SML-Ansätze, weil die Verteilung der Kategorien in den Trainingsdaten in Form einer verzerrten Schätzung weitergegeben werden kann. Dies ist vor allem dann der Fall, wenn Informationen für die Identifikation der Kategorien fehlen, z. B. Kontextinformationen über die Textebene hinaus, die für die Erkennung von Kategorien wie Inzivilität oder Hate Speech relevant sein können. Eine mögliche Verzerrung in der Schätzung ist anhand der allgemeinen Performance-Maße Accuracy und Macro-F1 nicht gut zu erkennen und kann besser beurteilt werden, wenn darüber hinaus auch die Maße Recall, Precision und F1 für jede der Kategorien angegeben werden. Dies gilt nicht nur für die Textklassifikation von UGC, sondern für SML-Verfahren allgemein. Die Ergebnisse der Studie konnten zeigen, dass sich durch Oversampling die Schätzleistung des Klassifikationsmodells in den einzelnen Kategorien angleicht, der Bias also verringert werden kann. Selbst wenn ein Classifier durch Oversampling an allgemeiner Schätzgenauigkeit einbüßen sollte, ist es für viele Fragestellungen sicher sinnvoll, im Gegenzug auf den systematischen Schätzfehler zu verzichten.

Zusammenfassend lässt sich festhalten, dass Oversampling die Textklassifikation von UGC deutlich verbessern kann, wenn eines der Konzepte in der Stichprobe unterrepräsentiert ist. Die Studie hat außerdem bestätigt, dass die Klassifikation von UGC besonders schlecht bei nicht balancierten Daten gelingt und wenn die Kategorien hinsichtlich der Verteilung von Wörtern und Wortgruppen nicht trennscharf sind. Oversampling kann das Problem von fehlenden Informationen in der Stichprobe nicht lösen, jedoch in den Daten bereits existieren „gute“ Features (z. B. Wörter oder Wortkombinationen) für die Textkategorien in der Schätzung sichtbar machen und das Modell somit verbessern.

Funding Open Access funding provided by Projekt DEAL.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

Literatur

- Aggarwal, C. C. (2018). *Machine learning for text*. New York: Springer.
- Aggarwal, C. C., & Zhai, C. (Hrsg.). (2012). *Mining text data*. Berlin: Springer.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. New York: O'Reilly Media.
- Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2), 1–31.
- Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2), 223–242.
- Chawla, N. V. (2009). Data mining for imbalanced datasets: an overview. In O. Maimon & L. Rokach (Hrsg.), *Data mining and knowledge discovery handbook* (S. 875–886). Boston: Springer.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321–357.
- Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64, 658–679.
- Davidson, T., Warmesley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th Conference on Web and Social Media*. AAAI. (S. 512–515).
- Denil, M., & Trappenberg, T. (2010). Overlap versus imbalance. In *Canadian conference on artificial intelligence* (S. 220–231). Berlin: Springer.
- Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., Ye, J., & Alzheimer's Disease Neuroimaging Initiative (2014). Analysis of sampling techniques for imbalanced data: An n= 648 ADNI study. *NeuroImage*, 87, 220–241.
- Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. (2018). *Detecting hate speech and offensive language on twitter using machine learning: an n-gram and tfidf based approach*. arXiv preprint arXiv:1809.08651.
- Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. Heidelberg: O'Reilly Media.
- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: review of methods and applications. *Expert Systems with Applications*, 73, 220–239.
- Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing* (S. 878–887). Berlin: Springer.
- Herndon, N., & Caragea, D. (2016). A study of domain adaptation classifiers derived from logistic regression for the task of splice site prediction. *IEEE transactions on nanobioscience*, 15(2), 75–83.
- Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River: Prentice Hall.
- Kwok, I., & Wang, Y. (2013). Locate the hate: detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence* (S. 1621–1622).
- Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1), 559–563.
- Loyola-González, O., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., & García-Borroto, M. (2016). Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing*, 175, 935–947.
- López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Information sciences*, 250, 113–141.
- Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north American Chapter of the Association for computational linguistics: human language technologies* (S. 746–751).
- Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. Heidelberg: O'Reilly Media.
- Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2019). SemEval-2016 task 4: Sentiment analysis in Twitter. In S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, & T. Zesch (Hrsg.), *Proceedings of the 10th international workshop on semantic evaluation* (S. 1–18).

- Nguyen, H. M., Cooper, E. W., & Kamei, K. (2009). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1), 4–21.
- Papacharissi, Z. (2004). Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, 6(2), 259–283.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Vanderplas, J., et al. (2011). Scikit-learn: machine learning in python. *Journal of machine learning research*, 12, 2825–2830.
- Peters, M., Neumann, M., Iyyer, M., Gradner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations*. CoRR 1802.05365.
- Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, Informedness, Markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.
- Raschka, S., & Mirjalili, V. (2017). *Python machine learning*. Birmingham: Packt Publishing Ltd.
- Risch, J., Krebs, E., Loser, A., Riese, A., & Krestel, R. (2018). Fine-grained classification of offensive language. In J. Ruppenhofer, M. Siegel & M. Wiegand (Hrsg.), *Proceedings of the GermEval 2018 Workshop and the 14th Conference on Natural Language Processing (KONVENS 2018)*.
- Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (S)*. 502–518.
- Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, communication & society*, 18(2), 121–138.
- Ruppenhofer, J., Siegel, M., & Wiegand, M. (2018). 14th Conference on Natural Language Processing KONVENS 2018.
- Schonlau, M., & Guenther, N. (2016). Text mining using n-grams. <https://ssrn.com/abstract=2759033>. Zugriffen: 13. Apr. 2020.
- Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting incivility and impoliteness in Online discussions. *Computational Communication Research*, 2(1), 109–134.
- Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., & Klenner, M. (2019). Overview of GermEval Task 2, 2019 shared task on the identification of offensive language. In J. M. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand & M. Klenner (Hrsg.), *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*.
- Waseem, Z. (2016). November). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science (S)*. 138–142).
- Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop (S)*. 88–93).
- Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the germeval 2018 shared task on the identification of offensive language. In J. Ruppenhofer, M. Siegel & M. Wiegand (Hrsg.), *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*.
- Zakaryazad, A., & Duman, E. (2016). A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing. *Neurocomputing*, 175, 121–131.
- Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1253.
- Ziegele, M., Daxenberger, J., Quiring, O., & Gurevych, I. (2018). Developing automated measures to predict incivility in public online discussions on the facebook sites of established news media. In *Paper presented at the 68th annual conference of the international communication association (IICA)*.

Anke Stoll ist wissenschaftliche Mitarbeiterin am Lehrstuhl für Politische Onlinekommunikation am Institut für Sozialwissenschaften der Heinrich-Heine-Universität Düsseldorf.