Framing von Migration in Reden des Deutschen Parlaments

Thao Van Liane Nguyen, Veronika Hentze

1 Einleitung

Sánchez-Junquera et. al. stellen in ihrer Studie einen neuen Ansatz vor, um Stereotype zu identifizieren. Sie nutzen dafür Frames, das heißt die narrativen Szenarien in denen eine untersuchte Gruppe in öffentlichen Reden platziert wird. Wir stellen diese Experimente mit einem SVM Modell nach und erreichen eine Accuracy von bis zu 0.7.

2 Vorbetrachtung

2.1 Auswahl der Daten

Wir nutzen in dieser Arbeit den Korpus der Plenarprotokolle des Deutschen Bundestages (CPP-BT). Dabei haben wir uns per Aufgabenstellung auf die Periode von 2015 bis 2017 beschränkt.

Sánchez-Junquera et. al. nutzen den Spanischen Teil des ParlSpeech Korpus (Rauh, Schwalbach 2020). Dabei fokusierten sie sich auf die Perioden von 1996 bis 1998, 2006 bis 2008, und 2016 bis 2018. Laut Sánchez-Junquera et. al. ist eine Besonderheit von ParlSpeech, dass es sich dabei um die Transkription von echten Debatten zwischen revelanten sozialen Akteuren handelt. Zusätzlich mache die Dialog-artige Natur des Korpus ein Herangehen aus der Perspektive der Computerlinguistik schwierig. Diese Dialog-artige Natur besteht zu einem gewissen Grad auch in dem CPP-BT Korpus. Das zeigt sich vor allem bei Zurufen, die des Öfteren passieren. Durch die Zurufe kann es zu Problemen bei der Extrahierung der Sätze kommen, da sie gelegentlich den Redefluss unterbrechen und der Satz vorzeitig beendet wird. Auch die Referenz auf vorherige Redner*innen kann potentiell zu Problemen führen.

2.2 Preselection und Annotation

Wir verwenden eine Stichwortliste um den CPP-BT Korpus nach relevanten Reden zu durchsuchen. Wenn eine Rede mehr als 50 Vorkommen von Stichwörtern hat, werden die Sätze extrahiert und für die Annotation verwendet. Es wurden vorerst 500 Sätze von 3 Annotator*innen annotiert.

Unsere Stichwortliste basiert auf der von Sánchez-Junquera et. al. erstellten Liste an Stichwörtern, allerdings umfasst sie mit 130 Stichwörtern knapp mehr als das doppelte der Stichwörter aus Sánchez-Junquera et. al. Wir haben unsere Stichwortliste mit Stichwörtern aus einzelnen, zufällig ausgewählten Reden welche die 50 Stichwort-Grenze erreicht hatten erweitert. Dabei wurde je eine zufällige Rede aus den Jahren in der Zeitspanne von 2015 bis 2017 genuzt.

Bei der Erstellung der Annotationsrichtline haben wir uns an der aus Sánchez-Junquera 2021 orientiert. Bei einer erstmaligen Untersuchung wurde klar, dass einige Frames und Kategorien in dem Korpus gar nicht vorkommen und wurden deshalb weggelassen. Das ist spezifisch Kategorie 6: Entmenschlichung. Des Weiteren wurden die Kategorien 4: Gefahr für die Gesellschaft und Kategorie 5: Gefahr für Einzelne zu einer Kategorie zusammengefasst, da es für beide Kategorien bei der ersten Untersuchung weniger Daten gab. Dadurch gibt es 4 verschiedene Kategorien mit insgesamt 22 Frames:

1. Opfer von Xenophoie
2. Leidende Opfer
3. Ökonomische Resource
4. Gefahr

Das sind insgesamt zwei Frames weniger als bei Sánchez-Junquera et. al. Des weiteren wurden Frames auch an den Deutschen Bundestag und Deutschland angepasst. So wird zum Beispiel bei Frame 1.2 vom Vergleich mit Deutschen Auswander*innen und ehemaligen DDR Bürger*innen

gesprochen, statt vom Vergleich mit Spanischen Personen, die ausgewandert sind.

2.3 Preprocessing

Für den Schritt des Preprocessing haben wir die Sätze tokenisiert, die Groß- und Kleinschreibung einheitlich gemacht indem alle tokens als lowercase gespeichert wurden, und die Punktuation und Stoppwörter entfernt. Beim Vorgehen haben wir uns an Sánchez-Junquera et.al. orientiert.

2.4 Vorgehen und Ziele

Sánchez-Junquera et al. untersuchen in ihrem Paper, wie effektiv klassische ML-Modelle und SOA-Modelle mithilfe der erstellten Taxonomie und des darauf basierenden annotierten Korpus das Framing von Migrant*innen klassifizieren können. Dabei werden zwei binäre Klassifikationsprobleme betrachtet:

- Enthält eine Aussage Stereotype über Migrant*innen oder nicht? (”Stereotype vs. Non-stereotype”)
- Angenommen, die Aussage enthält Stereotype: Werden Migrant*innen als Opfer oder Bedrohung dargestellt? (”Victim vs. Threat”)

Wir führen diese Aufgaben mit einem klassischen ML-Modell für überwachtes Lernen, der Support Vector Machine, durch. Wir nutzen, wie die Autor*innen des Papers, die Implementation von Scikit-learn. Sánchez-Junquera et al. trainieren die Modelle mit der tfidf-gewichteten Bag-of-words Repräsentation der Sätze. Es werden Unigramme, Bigramme und Trigramme verwendet, wobei Unigramme die besten Ergebnisse erzielen. Im Rahmen dieses Projektes nutzen wir ebenfalls die Bag-of-words Repräsentation, wobei wir das Modell sowohl mit tfidf-Gewichtung als auch ohne tfidf-Gewichtung trainieren. Außerdem testen wir Sentiment als Feature.

Ziel des Projektes ist es nicht, eine möglichst hohe Performanz der Modelle zu erreichen. Es geht eher darum, die Methode von Sánchez-Junquera et al. explorativ für einen deutschen Korpus nachzubilden. Ist bei einem klassischen ML-Modell, wie der Support-Vector-Machine ein Lerneffekt erkennbar oder basiert die Klassifizierung mehr auf Zufall und geht nicht tatsächlich signifikant über die Accuracy einer Majority Baseline hinaus? Wie gut eignen sich triviale lexikalische Features für das Training? Ist eine Sentimentanalyse hilfreich? Was sagen uns Konfusionsmatrix etc.?

3 Aufgabe 1: Stereotype vs. Keine Stereotype

Die Arbeit von Sánchez-Junquera et al. hebt sich von vorangegangenen Arbeiten zur automatischen Identifizierung von Stereotypen (über Migrant*innen) dadurch ab, dass sie das Framing von von Migrant*innen ins Zentrum rückt. Dies soll ermöglichen auch subtile Stereotypisierungen zu berücksichtigen, die nicht durch die Kategorie ”Migrant*innen werden wörtlich negative Attribute zugeschrieben” erfasst werden. Das Augenmerkt der ersten Klassifikationsaufgabe liegt also darin, die Modelle in Hinblick auf ihre Fähigkeit, subtile Stereotype zu erkennen, zu bewerten.

Die Annotator*innen wurden gebeten, die Aussagen über Migrant*innen Frames zuzuordnen. Sánchez-Junquera et al. erheben den Anspruch, dass die Taxonomie von Frames, vollständig die Dimensionen der Stereotypisierung von Migrant*innen abbildet, sei sie positiv oder negativ. Alle Sätze, denen ein Frame oder eine Kategorie zugeordnet wurde, enthalten also laut Sánchez-Junquera et al. Stereotype. Sätze, welche keinem Frame oder keiner Kategorie zugeordnet werden konnten, sollten mit ”x” annotiert werden.

Dem Datensatz wird eine Spalte ”Contains stereotypes” mit den Goldlabels für die Klassifikation hinzugefügt. Sätze, welche mit ”x” annotiert wurden, erhalten das Label ”0” für ”enthält keine Stereotype”. Alle anderen Sätze erhalten ”1” als Label.

3.1 Klassenverteilung

Wir haben aus den insgesamt 500 Sätzen 302 Sätze ohne Stereotyp und 198 Sätze mit einem beliebigen Stereotyp erhalten. Das heißt, dass es im ungefähren Verhältnis 3:2 mehr Sätze ohne Stereotyp gibt als mit.

3.2 Extraktion von Features

1. Bag-of-Ngrams ohne Tfidf-Gewichtung

Zunächst wandeln wir den Datensatz in seine Bag-of-words bzw. ”Bag-of-ngrams” Repräsentation um. Jeder Satz wird also als Vektor dargestellt, dessen Werte die Counts aller im Datensatz vorkommenden Ngramme darstellen.

2. Bag-of-ngrams mit Tfidf-Gewichtung

Nun ziehen wir das Tfidf-Maß hinzu, um die Counts der Ngramme zu gewichten.

3. Sentiment Scores

Als letztes berechnen wir für jeden Satz noch den Sentiment score. Dafür nutzen wir das Package TextBlobDE.

3.3 Trainieren und Evaluieren der Modelle

Majority Baseline:

Als Baseline verwenden wir eine Majority Baseline, d.h. es wird immer die häufigste Kategorie vorhergesagt. Sie hat eine Accuracy von 0.604.

3.4 Ergebnisse

1. Bag-of-ngrams ohne Tfidf-Gewichtung:

Die Accuracy ist bei allen Ergebnissen außer bei Bigrammen höher als die Baseline. Am besten hat das Unigram Modell mit einer Accuracy von 0.62 abgeschnitten. Allerdings ist fraglich ob die Verbesserung um lediglich 0.02 im Falle des Unigramm Modells im Vergleich zur Baseline statistisch signifikant ist.

Der F1 score zeigt, dass Kategorie 0 bei allen Modellen bevorzugt vorheresagt wird.

Mit den neuen annotierten Daten ist Verwendung der Bi- und Trigramme als Features erkennbar, dass das Modell keinen Lerneffekt erzielt. Der Recall für das Label 1 beträgt bei beiden Versuchen 0.0, das Modell sagt also nie Label 1 voraus. Bei den Unigrammen zeigt sich auch kein signifikanter Unterschied mit den neuen annotierten Daten. Unigramme weisen jedoch immer noch die beste Performanz auf. Allerdings ist immer noch fraglich, ob es einen Lerneffekt gegeben hat oder es sich um puren Zufall handelte.

2. Bag-of-ngrams mit Tfidf-Gewichtung:

Das Unigram Modell hat wieder die besten Ergebnisse, allerdings ist der macro F1 score mit 0.43 niedriger als bei dem Modell ohne Tfidf-Gewichtung. Mit einem F1 score von 0.00 für Kategorie 1 sind die Bi- und Trigramm Modelle schlechter als die Modelle ohne Tfidf-Gewichtung. Des Weiteren schaffen sie es nicht, die Accuracy der Baseline zu überbieten. Auch hier ist fraglich, ob die Verbesserung vom Unigramm Modell statistisch signifikant ist. Insgesamt schneiden alle Modelle schlechter ab als die Modelle ohne Tfidf-Gewichtung.

Mit den neuen annotierten Daten weist die Verwendung des TF-IDF-Maßes als Feature leicht bessere Ergebnisse auf, allerdings sind diese ebenfalls als nicht besonders positiv zu bewerten. Erst mit dieser "Verbesserung" erreicht das Modell für Bi- und Trigramme die Baseline (vorher waren die Ergebnisse sogar unter der Baseline). Interessanterweise bewirken Unigramme immer noch die beste Performanz. Auch wenn die statistische Signifikanz zu hinterfragen ist, zeigt sich hier also eine Kontinuität in den Ergebnissen.

3. Sentiment:

Das Modell mit nur Sentiment ist genau so gut wie die Bi- und Trigram Modelle mit tfidf-Gewichtung.

Mit den neuen annotierten Daten beträgt der Recall für das Label 1 nicht mehr 0.0. und die F1-Scores (Macro und gewichtet) verbessert sich dadurch sichtlich. Trotzdem bleibt die Accuracy unter der Baseline.

4. Kombinationen von Features:

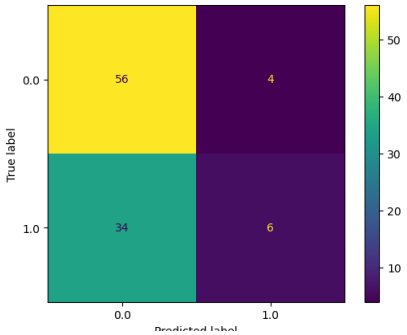
Wir testen außerdem, ob eine Kombinationen von unterschiedlichen Features eine Verbesserung der Performance bewirkt.

Bag-of-ngrams + Sentiment:

Das kombinierte Unigram Modell erzielt die insgesamt besten Ergebnisse. Allerdings ist die Verbesserung nur gering. Durch die Kombination mit dem Sentiment Modell hat sich der Marcro F1-score bei allen Modellen erhöht.

Mit den neuen annotierten Daten verschlechtert sich die Performanz für die Kombinationen Bigramme + Sentiment sowie Trigramme + Sentiment. Für die Kombination Unigramme + Sentiment verbessert sich das Modell bei der Vorhersage von Label 0 (unstereotypisch), verschlechtert sich allerdings bei der Vorhersage von Label 1. Auch mit den neuen Daten lässt sich jedoch feststellen, dass diese Feature-Kombination insgesamt die beste Performanz erzielt.

Konfusionsmatrix:



Wie auch bei Sánchez-Junquera et. al. gibt es mehr Konfusion für Nicht-Stereotype. Sie wurden häufiger falsch zugeordnet, d.h. Stereotype wurden öfter nicht als solche erkannt und falsch zugeordnet. Allerdings ist der Unterschied bei Sánchez-

Junquera et. al. nicht so extrem.

Laut Sánchez-Junquera et. al. lassen sich Stereotype oft nicht einfach an gewissen Wörtern ausmachen. Insgesamt 90% wurden als kein Stereotyp vorhergesagt, die eigentliche Verteilung müsste bei 60% liegen. Ein Grund für diese Verteilung könnte die Unausgeglichenheit der Daten sein. Da es mehr Daten zu nicht Sterotypen gibt, erkennt das Modell diese leichter und ordnet auch fälschlicherweise Sätze anderer Kategorien zu. Dieses Problem wird durch die geringe Menge an Daten noch verstärkt.

Auch mit mehr Daten ändert sich nicht viel an der Konfusion. Das Modell hat immer noch deutlich mehr Probleme damit, stereotypische Aussagen zu erkennen. Das Modell erkennt in fast 90% der Fälle nicht, dass es sich um eine stereotypische Aussage handelt. Mit den verwendeten Features (Unigramm-Counts + Sentiment) konnte also kein Lerneffekt erzielt werden.

3.5 Interpretation: Warum scheitert das Modell in den falschen Beispielen?

Ein möglicher Grund dafür, dass so oft fälschlicherweise Nicht-Stereotype voraus gesagt werden, ist dass es im Verhältnis 3:2 mehr Daten für Nicht-Stereotype gibt als für Stereotype. Durch diese Unausgeglichenheit

Laut Sánchez-Junquera et. al. lassen sich Stereotype nicht nur durch die Präsenz spezifischer Wörter erkennen. Das könnte erklären, warum oft ein Nicht-Stereotyp vorhergesagt wird, da die Erkennung von Stereotypen schwieriger ist.

Die falsche Klassifiszierung könnte auch daran liegen, dass Maschinen Schwierigkeiten mit Rhetorischen Mitteln haben. Beispiele hierfür sind unter anderem die rhetorische Frage bei 7[1], oder auch der Sarkasmus in 48. Die rhetorischen Mittel werden nicht erkannt bzw. sind der Maschine unbekannt und werden nicht richtig klassifiziert. Das scheint uns vor allem beim Sarkasmus wahrscheinlich.

Interessanterweise sind die Sätze, die fälschlicherweise als Stereotyp vorhergesagt wurden, oft deutlich länger als der Durchschnitt. Eine mögliche Begründung könnte sein, dass in den längeren Sätzen mehr Features gefunden werden, die dann eher zu Stereotypen passen als zu Nicht-Stereotypen.

[1] Die hier referenzierten Zahlen sind der Index von Beispielen falscher Klassifikationen. Diese sind im verlinkten Jupyter Notebook auf GitHub zu finden.

3.6 Assoziation zwischen Ngrammen und Labels

Sánchez-Junquera et al. argumentieren, dass Stereotype nicht nur in Form von Zuschreibungen negativer Attribute einer Personengruppe auftreten, sondern oft deutlich subtiler sind. Vielmehr zeichnen sich Stereotype durch das Framing, also durch das Narrativ und den Kontext, in welches die jeweilige Personengruppe platziert wird, aus. Indem das Framing von Migration/Migrant*innen annotiert wird, sollen auch diejenigen Aussagen über Migrant*innen, welche nicht auf den ersten Blick als stereotypisch erkennbar sind, identifiziert werden.

Die Autor*innen des Papers kommen auf einen interessanten Ansatz, in dem die Bereicherung der Sozialwissenschaften durch komputationelle Verfahren sichtbar wird. Sie berechnen für alle Ngramme (Bi- und Trigramme) den PMI (Pointwise Mutual Information) zu den jeweiligen Labels (also ”Stereotype vs. Nonstereotype” bzw. ”Victims vs. Threats”). PMI ist ein statistisches Maß, welches die Assoziation zwischen Ereignissen misst. Mit anderen Worten: Zum einen untersuchen die Autor*innen, welche Ngramme am stärksten mit stereotypischen bzw. nicht-stereotypischen Aussagen assoziert sind. Zum anderen untersuchen sie welche Ngramme am stärksten mit pro-migrantischen Aussagen, welche Migrant*innen als Opfer framen, assoziiert sind und welche Ngramme am stärksten mit anti-migrantischen Aussagen, in denen Migrant*innen als Bedrohung dargestellt werden, assoziiert sind. Ziel dabei ist, subtilere, nicht-triviale linguistische Muster automatisch zu ermitteln, welche zwar stark mit Stereotypen bzw. einer bestimmten Ausprägung von Stereotypen verbunden (also oft gleichzeitig mit ihnen auftreten), jedoch nicht für uns als solche ersichtlich sind.

Das wollen wir ebenfalls nachbilden.

Interpretation

Aus der Liste von Bigrammen, die am stärksten mit Aussagen ohne Stereotype assoziiert sind, kann man eher wenig Information herausziehen. Bigramme wie ”abgegrenzt flüchtlinge” oder ”abschieben abschrecken” sind überraschend. Rein intuitiv und vielleicht ein wenig mit Weltwissen würden wir vermuten solche Bigramme in einem Kontext vorzufinden, in dem zum Beispiel die Asylpolitik kritisiert wird. Zumindest sind diese Bigramme nicht neutral. Betrachtet man die Bi-

gramme, die am stärksten mit stereotypischen Aussagen assoziiert sind, zeigen sich ebenfalls Bigramme, die man nicht ganz klar nachvollziehen kann (z.B. ”abgeordneten bündnisses”).

Einige Bigramme nehmen wir allerdings auch als recht plausibel wahr. Ein Beispiel ist das Bigramm ”abgestattet islamistischen”. Das häufige Aufbringen von Islamismus im Zusammenhang mit Geflüchteten oder Migrant*innen in rechtspopulistischen, anti-migrantischen Diskursen, die sich vor allem stereotypischen Bildern bedienen, ist bekannt. Auch, dass ”ablehnen parallelgesellschaften” in der Liste auftaucht, erscheint plausibel. Ein stereotypisches Bild, welches oft gezeichnet wird, ist das der Migrant*innen, die sich in Deutschland eine ”Parallelgesellschaft” aufbauen, ohne den Willen sich in in die deutsche Gesellschaft zu integrieren.

Insgesamt findet man jedoch nur punktuell Bigramme, welche sich klar mit unserer menschlichen Intuition vereinbaren lassen. Sánchez-Junquera et al. sehen das in ihrem Paper als Indiz dafür, dass Stereotype sich nicht trivialerweise und intuitiv sprachlich charakterisieren lassen und argumentieren, dass komputationelle Verfahren deshalb interessante Erkenntnisse liefern können. Für unsere Studie wäre diese Schlussfolgerung deutlich zu weit hergeholt. Die sehr begrenzte Datenmenge lässt keine überzeugende Verallgemeinerung zu. Viele der Bigramme tauchen vermutlich alleinig aufgrund ihrer schieren Frequenz in den Listen auf und nicht, weil sie semantisch mit Stereotypen bzw. nicht mit Stereotypen assoziiert sind.

Die Listen haben sich mit den neuen annotierten Daten nicht großartig verändert. In der Liste der mit stereotypischen Aussagen assoziierten Bigramme ist nun (’abschaffung asylbewerberleistungsgesetzes’) verschwunden. Das ist möglicherweise dadurch erklärbar, dass durch die größere Datenmenge nun sehr spezifische Bigramme, welche vorher zufällig viel Gewicht hatten nun an Bedeutung verloren haben. Ansonsten bleibt es jedoch bei der obigen Auswertung.

4 Aufgabe 2: Opfer vs. Bedrohung

Bei diesem Experiment geht es darum herauszufinden, ob das Modell es schafft, die Dimensionen der Stereotype von Aussagen richtig zu klassfizieren. Das Augenmerk von Sánchez-Junquera et al. liegt hierbei insbesondere auf Aussagen, welche zwar als anti-migrantisch zu definieren sind, jedoch durch rhetorische Mittel von Politiker*innen als solche unkenntlich gemacht werden. Menschliche Annotator*innen schaffen es (in den meisten Fällen), solche Aussagen richtig einzuordnen. Die Frage ist, ob die getesteten Modelle auch dazu in der Lage sind.

Für diese Aufgabe erstellen wir einen Teil-Datensatz bestehend aus allen Aussagen, denen von den Annotator*innen ein Frame oder eine Kategorie zugeordnet wurde. Sánchez-Junquera et al. definieren die Suprakategorie ”Opfer” als die Vereinigung von Kategorien 1 (”Opfer von Xenophobie”) und Kategorie 2 (”Leidende Opfer”) sowie die Suprakategorie ”Bedrohung” als die Vereinigung von Kategorie 4 (”Gefahr für das Kollektiv”), Kategorie 5 (”Gefahr für Einzelne”) und Kategorie 6 (”Entmenschlichung”). Kategorien 4 und 5 haben wir zusammengefasst, Kategorie 6 aufgrund fehlender Daten ausgeschlossen. Kategorie 3 (”Ökonomische Ressource”) wurde für dieses Experiment von Sánchez-Junquera et al. exkludiert, da zu wenig Daten vorlagen. Aus demselben Grund entschieden auch wir uns dazu, diese Kategorie nicht einzubeziehen. Letzendlich erhielten also alle Aussagen mit einem Frame ¡ 3 das Label ”1” für ”Opfer” und alle Aussagen mit einem Frame ¿ 3 das Label ”0” für ”Bedrohung.

4.1 Klassenverteilung

Aus den 198 Sätzen mit Stereotyp erhalten wir 128 Sätze, die mit Kategorien annotiert wurden und welche der Suprakategorie ”Opfer” angehören Davon sind 39 aus Kategorie 1: Opfer von Xenophobie und 89 aus Kategorie 2: Leidende Opfer. 54 Sätze gehören zur Suprakategorie ”Bedrohung”. Diese besteht nur aus den Sätzen von Kategorie 4. Wie bereits erwähnt wurden die 16 Sätze aus Kategorie 3 nicht mit betrachtet, da es zu wenig Daten gibt. Es ist auffällig, dass es mehr als doppelt so viele Sätze in der ”Opfer” Suprakategorie gibt als in der ”Bedrohung” Suprakategorie”.

4.2 Ergebnisse

1. Bag-of-ngrams ohne Tfidf-Gewichtung:

Alle Modelle schneiden mit einer Accuracy von 0.7 und dem Macro F1 score von 0.41 gleich gut ab. Der F1 Score für Kategorie 0 (Bedrohungen) liegt bei 0. Das zeigt dass Kategorie 0 nicht vorhergesagt wurde.

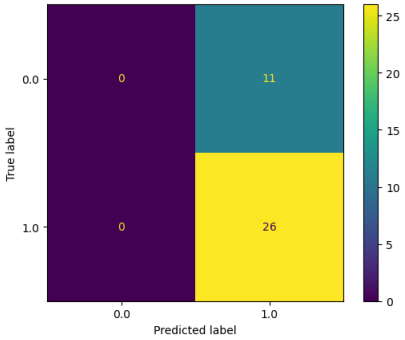
Die Baseline wurde mit jedem Modell geschlagen.

2. Bag-of-ngrams mit Tfidf-Gewichtung:

Bei den Modellen mit tfidf-Gewichtung schneiden alle Modelle genauso gut ab wie bei den Modellen ohne tfidf-Gewichtung.

3. Sentiment:

Das Modell mit Sentiment erreicht ebenfalls eine Accuracy von 0.7 und ein Macro F1 score von 0.41. Dadurch wird auch hier die Baseline geschlagen.



Es ist auffällig, dass gar keine Vorhersagen zur Suprakategorie Bedrohung gemacht werden. Die Daten für Experiment 2 sind unausgeglichener als die für Experiment 1, d.h. es besteht ein ungefähres Verhältnis von 2:1 für Opfer zu Bedrohung bei Experiment 2. Im Vergleich dazu ist das ungefähre Verhältnis 3:2 für Nicht-Stereotype zu Stereotype bei Experiment 1. Deshalb gehen wir davon aus, dass durch die unausgeglichene Daten kombiniert mit der geringeren Datenmenge das Modell nicht lernen konnte, was die Kategorie Bedrohung ausmacht. Wie auch Sánchez-Junquera et. al. kommen wir zu dem Schluss, dass die geringere Größe des Trainingsets einen Einfluss auf die Ergebnisse von Experiment 2 hat.

Sánchez-Junquera et. al. vermuten, dass die Konfusion des Models darauf beruht, dass in der Politik die gleichen Worte für verschiedene Zwecke verwendet werdem, um nicht als xenophob bezeichnet zu werden. Das erscheint uns plausibel, und könnte eine Begründung sein, warum für Bedrohung keine Sätze gefunden wurden.

5 Evaluation

Bei welchem Experiment ist die Performanz besser? Bei Experiment 2 ist die Accuracy mit 0.7 bei jedem Modell deutlich höher als beim besten Ergebnis von Experiment 1, welches eine Accuracy von 0.63 hat. Allerdings ist der Marco F1 score von Experiment 2 mit 0.41 schlechter als der Macro F1 von 0.5 beim besten Model von Experiment 1. Die höhere Accuracy mit geringerem F1 score könnte daran liegen, dass durch die verschiedenen unbalancierten Daten eine majority baseline nachgebildet wird: je größer die Majority Class ist, desto höher ist auch die Accuracy wenn nur oder größenteils die Majority Class vorhergesagt wird.

Insgesamt schätzen wir ein, dass Experiment 1 die bessere Performance hat. Das wird vor allem bei der Betrachtung der F1 scores deutlich.

Anmerkung

Beim ersten Experiment hat sich bereits folgendes deutlich abgezeichnet: Ein solch (nicht nur) linguistisch komplexes Problem lässt sich nicht mit einem einfachen linearen Klassifikationsmodell, z.B. einer Support-Vector-Machine lösen. Auch ein Datensatz von 1000 Sätzen im Vergleich zu 500 Sätzen bewirkt keinen signifikanten Unterschied. Aus diesem Grund haben wir von der weiteren Reevaluation des zweiten Experimentes (Opfer vs. Bedrohung) mit den neuen annotierten Daten abgesehen.

6 Fazit und Diskussion

ánchez-Junquera et al. vergleichen in ihrem Paper klassische ML-Modelle, welche durch relativ triviale lexikalische Features trainiert wurden, mit komplexeren SOA-Transformermodellen. Bei der ersten Klassifikationsaufgabe (”Stereotype vs. Nonstereotype”) sind die klassischen ML-Modelle den komplexeren Transformermodellen weit unterlegen. Beim zweiten Experiment (”Victims vs. Threats”) ist es umgekehrt. Die Autor*innen vermuten, dass komplexere Transformermodelle aufgrund der komplexen Natur der Aufgabe besser geeignet sind. Es herrschen jedoch, insbesondere für die zweite Aufgabe, zu wenig Daten vor, was die bessere Performanz der klassischen ML-Modelle gegenüber den Transformermodellen bei dieser Aufgabe erklären würde.

Da wir nur mit einem klassischen ML-Modell gearbeitet haben, lässt sich kein Vergleich zu komplexeren Transformermodellen wie bei Sánchez-Junquera et al. herstellen. Es lässt sich dennoch feststellen, dass das Trainieren des Modells mit rein lexikalischen Features definitiv nicht ausreicht, um einen tatsächlichen Lerneffekt, was die linguistischen Muster bezüglich Stereotypen angeht, zu erzielen. Auch die Hinzunahme der Sentimentanalyse brachte keine signifikanten Verbesserungen. Schon bei einem Modell, welches die Sentimentanalyse ”perfekt”, zumindest der Bewertung genügend menschlicher Annotator*innen entsprechend, durchführt, wäre es nur spekulativ vermutbar, inwiefern Sentiment und Stereotyp-

isierung eines Satzes zusammenhängen. Bei einem ”fehlerhaften” Sentiment-Modell verstärkt sich die Unsicherheit noch enorm.

Die schlechte Performanz bei beiden Aufgaben hängt sehr wahrscheinlich mit dem Mangel an annotierten Trainingsdaten zusammen. Desweiteren wurden die zu annotierenden Sätze für unser Projekt automatisch extrahiert. Sánchez-Junquera et al. hingegen wählten diese manuell aus. Dies könnte ebenfalls einen Faktor für die Diskrepanz der Ergebnisse darstellen.

In zukünftigen Untersuchungen könnte man eventuell mit der Einheit der Annotationen experimentieren. Beide Projekte sind von der Satzebene ausgegangen. Während Sánchez-Junquera et al. sich fragen, ob es eventuell sogar möglich sein sollte, einem Satz mehrere Frames zuzuordnen, haben wir uns die Frage gestellt, ob ein Satz überhaupt genügend Kontext bietet, um ihn adäquat zu annotieren. Der fehlende Kontext eines einzelnen Satzes machte es oft schwer, ihm einem Frame zuzuordnen, zumal diese sehr spezifisch formuliert sind.

Bemerkung: Aufgrund des deutlich kleineren Rahmens unseres Projektes haben wir eine Berechnung des Inter-Annotator-Agreements ausgelassen. Dies wäre aber (zumindest, um einem wissenschaftlichen Anspruch gerecht zu werden) essentiell gewesen, um unseren Ergebnissen Legitimation zu verleihen.

7 Verlinkungen

Github-Repository mit dem Projekt

Hinweis: Im Repository befindet sich eine Juypter Notebook Datei, die das Vorgehen weiter beschreibt und die Ergebnisse präsentiert