

Tagger für Intensivierer
Ein Versuch der automatischen Identifikation von
Intensivierern von Adjektiven

Schriftliche Hausarbeit
für die Bachelorprüfung der Fakultät für Philologie
an der Ruhr-Universität Bochum
(Gemeinsame Prüfungsordnung für das Bachelor-Studium im
Rahmen des 2-Fach-Modells an der RUB vom 03. November 2016)

vorgelegt von

Yüzüncüoğlu, Imge

22.08.2022

Jun.-Prof. Dr. Tatjana Scheffler
Prof. Dr. Stefanie Dipper

Inhaltsverzeichnis

1. Einleitung	3
2. Forschungsstand	4
2.1 Theoretischer Forschungsstand	4
2.2 Computerlinguistischer Forschungsstand	5
3. Intensivierer	6
3.1 Der Terminus	6
3.2 Definition	9
3.2.1 Abgrenzung der Intensivierer von Fokus- und Gradpartikeln	11
3.2.2 Verstärkende und abschwächende Intensivierer	12
3.2.3 Morphologische und syntaktische Intensivierer	13
3.3 Intensivierer – eine offene Wortklasse	15
4. Das Korpus	17
4.1 Informationen zum Korpus	18
4.2 Vorbereitung der Korpusdaten	18
5. Tagger für Intensivierer (Ta f In)	20
5.1 Conditional Random Field	22
5.2 Aufbau des CRFs	23
5.2.1 Allgemeines	23
5.2.2 Merkmale	25
5.2.3 Inferenz und Training	26
5.3 Testphase	29
6. Evaluation und Diskussion	29
6.1 Evaluation	29
6.2 Diskussion und Optimierungsmöglichkeiten	31
7. Fazit	32

Abkürzungsverzeichnis

Literaturverzeichnis

Anhang

Eigenständigkeitserklärung

1. Einleitung

Sehr vielfältig, *supercool* und *sauinteressant* sind Beispiele eines Phänomens, das unter anderem als *Intensivierung* bekannt ist. Dabei handelt es sich um einen Prozess, bei dem ein Ausdruck modifiziert beziehungsweise intensiviert wird (Kirschbaum, 2002; Os, 1989) wie die Beispiele (1) und (2) veranschaulichen. Das jeweils intensivierende Element wird hierbei (und im Folgenden) unterstrichen.

- (1) Die Geschichte ist verwirrend.
- (2) Die Geschichte ist total verwirrend.

Die Aussage in Beispiel (1) teilt mit, dass eine bestimmte Geschichte verwirrend sei. Innerhalb von Beispiel (2) wird das Adjektiv *verwirrt* durch den Ausdruck *total* positiv verstärkt, sodass die Eigenschaft, dass die Geschichte verwirrend sei, stärker betont und somit intensiviert wird. Auch Kirschbaum (2002:6) versteht unter der Intensivierung ein „semantisch-funktionale[s] Phänomen der Gradspezifikation“, das aus zwei Teilen besteht. Zum einen aus dem zu intensivierenden Ausdruck *verwirrend* und zum anderen aus dem Intensivierungsmittel *total*. Das Phänomen der Intensivierung ist eine vor allem in der *computer-mediated communication* (CMC), also der ‚computervermittelten Kommunikation‘ (Blogs, Social Media Texte und so weiter), verbreitete Methode, die in verschiedenen und immer wieder neuen Formen erscheint und verschiedene Elemente modifizieren kann, wodurch es eine komplexe Konstruktion ist (Laitenberger, 2016; Tagliamonte, 2016). Daher sind Intensivierer häufig Thema linguistischer Arbeiten, die sich mit ihnen im Allgemeinen befassen oder gezielte Aspekte und Eigenschaften näher analysieren. Um Intensivierer jedoch untersuchen zu können, müssen diese vorerst innerhalb der jeweiligen Korpora manuell identifiziert und als solche markiert werden. Je größer der Datensatz ist, desto aufwendiger gestaltet sich dieser Prozess. Um diesen effizienter zu gestalten, ist es das Ziel dieser Arbeit, ein Modell zu implementieren, das Intensivierer innerhalb eines Datensatzes automatisch erkennt und als solche annotiert. Aufgrund einiger Faktoren, die im Laufe der Arbeit weiter ausgeführt werden, beschränkt sich diese Identifikation jedoch vorerst auf Intensivierern von Adjektiven. Um sich dieser automatischen Annotation annähern zu können, muss vorher eindeutig abgegrenzt werden, wonach das Modell zu suchen hat. Demnach verfolgt diese Arbeit zwei Ziele und gliedert sich wie folgt: zunächst folgt innerhalb von Kapitel

2 ein Einblick in den Forschungsstand. Nach der Erläuterung des Forschungsstandes wird in Kapitel 3 auf das erste Ziel der Arbeit, auf die Definition von Intensivierern, eingegangen. Dabei wird unter anderem die Wahl der Terminologie erläutert, die zusätzlich dabei hilft, den Begriff eindeutig von anderen abzugrenzen. Anschließend folgen in Kapitel 4 Informationen zum für diese Arbeit gewählten Korpus und wie dieser für die Implementierung vorbereitet wurde. In Kapitel 5 wird das zweite Ziel der Arbeit verfolgt, bei dem Informationen zur Implementierung erfolgen. Präsentiert wird eine Erläuterung, weshalb der *Conditional Random Field* Algorithmus gewählt wurde, eine Definition wie dieser aufgebaut ist sowie Details zur Implementierung selbst, die anschließend innerhalb von Kapitel 6 im Vergleich mit zwei weiteren Baselines evaluiert und anschließend diskutiert wird. Zum Schluss erfolgt innerhalb von Kapitel 8 eine Zusammenfassung zur gesamten Arbeit.

2. Forschungsstand

Da diese Arbeit neben dem theoretischen Aspekt ebenso einen computerlinguistischen Teil beinhaltet, ist die Erläuterung bisheriger Untersuchungen zu Intensivierern in zwei Unterkapitel aufgeteilt. Im ersten Teil werden bisherige theoretische Arbeiten zu Intensivierern erörtert. Im zweiten Teil erfolgt ein Überblick über angewandte computerlinguistische Ansätze.

2.1 Theoretischer Forschungsstand

Wie bereits erwähnt, stehen Intensivierer im Fokus einiger Arbeiten, von denen im Folgenden ein paar erörtert werden. Zum Beispiel gibt es Untersuchungen, die sich darauf fokussieren, wie Intensivierer zu definieren sind. Für Definitionen von Intensivierern werden im Deutschen häufig auf Os (1989) und im Englischen auf Quirk et al. (2000) zurückgegriffen, die jeweils auf verschiedene Aspekte von Intensivierern wie beispielsweise welche Funktion sie in einem Satz erfüllen, wie sie in einem Satz integriert werden können und welche Arten von Intensivierern es geben kann, ausführlich eingehen.

Soziolinguistische Studien wie von Ito und Tagliamonte (2003), Tagliamonte (2016), Fuchs (2017) und Stratton (2022) fokussieren sich darauf herauszufinden, welche Faktoren, wie Alter, Geschlecht und Herkunft, Einfluss auf die

Verwendung von Intensivierern haben. Während dabei nicht nur erläutert wird, dass Gruppen von Personen mit einem gemeinsamen Faktor andere Intensivierer bevorzugen als andere Gruppen, können ebenfalls Hypothesen aufgestellt werden, weshalb es sich bei Intensivierern um eine offene Wortklasse handelt.

Forschungen über die semantische Entwicklung von Intensivierern wie von Laitenberger (2016) untersuchen den diachronen Prozess ausgewählter Intensivierer. Dabei kann beobachtet werden, wie sich die Bedeutungen von Intensivierern, die ursprünglich per Definition eine negative Konnotation haben, nun aber als positive Verstärkungen verwendet werden, entwickelten. Laitenberger (2016) kam dabei zu dem Entschluss, dass die Wandlung vom negativ konnotierten Begriff zu einer positiven Verstärkung aus fünf aufeinander aufbauenden Stufen bestünde, die ebenfalls bei russischen Intensivierern zu beobachten seien. Sie präsentiert somit nicht nur einen diachronen Überblick über den Wandel ausgewählter Intensivierer, sondern auch einen Vergleich mit Intensivierern aus einer anderen Sprache. Eine weitere Arbeit, die sich mit dem Vergleich von Intensivierern in verschiedenen Sprachen auseinandersetzt, ist von Baños (2013). Baños (2013) untersucht konkret anhand einer Sitcom, wie Intensivierer vom Englischen ins Spanische übertragen wurden.

2.2 Computerlinguistischer Forschungsstand

Einen maschinell entwickelten Tagger spezifisch für Intensivierer von Adjektiven oder auch Intensivierern im Allgemeinen gibt es bisher offiziell nicht. Identifikationen von Intensivierern wurden und werden bisweilen hauptsächlich manuell durchgeführt. Erste Ansätze zur automatischen Identifizierung sind bei einer Gemeinschaftsaufgabe von EmpiriST (2015), Stratton (2020) und Scheffler (unveröffentlicht) vorzufinden. Die Gemeinschaftsaufgabe von EmpiriST (2015) bestand daraus, dass Teilnehmerinnen und Teilnehmer unter anderem ein *Part-of-Speech* (PoS) Tagssystem, ein System, das für jedes Element innerhalb eines Textes, sei es eine Wortform, eine Zahl, ein Satzzeichen, ein Sonderzeichen oder Ähnliches, eine Wortart zuordnet (Schiller et al., 1999), ausgerichtet auf CMC und Social Media, implementieren sollten (EmpiriST, 2015). Der PoS-Tagger besteht dabei aus den Tags des STTS von Schiller et al. (1999) und sollte um 18 weitere Tags erweitert werden, die Phänomene aus der CMC repräsentieren sollten, die nicht durch bereits bestehende Tags des STTS kategorisiert

werden (Beißwenger et al., 2016). Darunter befindet sich ebenfalls der Tag PTKIFG, der Intensitäts-, Fokus- oder Gradpartikel kennzeichnen soll. Somit ist der Tag eine Sammelklasse und identifiziert nicht nur Intensivierer, weshalb diese erneut manuell herausgesucht werden müssten. Stratton (2020) filterte mittels PoS-Tags 5.000 Adjektive aus einem Korpus und markierte anschließend manuell, ob sie intensiviert wurden oder nicht. Scheffler et al. (unveröffentlicht) hingegen extrahierten nicht einfach alle Adjektive, sondern suchten nach allen prädikativen Adjektivphrasen im Korpus. Dafür verwendeten sie einen automatischen PoS-Tagger, um prädikative Adjektive zu identifizieren. Weiter wurden daraus Phrasen selektiert, die aus einem Personalpronomen, gefolgt von einem Auxiliar sowie sechs weiteren Tokens, von denen keine eine Konjunktion war, und final aus einem prädikativen Adjektiv bestanden. Innerhalb der selektierten Sätze wurden anschließend mittels einer angefertigten Wörterliste bestehend aus Intensivierern jene Wörter im Satz als Intensivierer markiert, die in besagter Liste aufgelistet sind. Mit Hilfe dieses Patterns sollten hohe Präzision und Wiedererkennung erzeugt werden. Allerdings blieb auch hier noch die Aufgabe bestehen, aus den Kandidaten für Intensivierer jene zu markieren, die auch wirklich Intensivierer sind.

3. Intensivierer

Dieses Kapitel beinhaltet eine Erläuterung zur gewählten Terminologie sowie eine Definition dessen, worum es sich bei dem Phänomen der Intensivierung handelt, welche Arten von Intensivierungen es gibt und wie diese erzeugt werden.

3.1 Der Terminus

Sichtet man die Literatur zu Intensivierungsprozessen, werden jene Mittel, die eine Intensivierung erzeugen, unter anderem unter folgenden, unterschiedlichen Begriffen gefasst: *Gradpartikel* (Boettcher, 2009), *Intensivpartikel* (Hentschel und Weydt, 2013), *Intensitätspartikel* (Granzow-Emden, 2019; Breindl, 2009), *Intensitätsadverbien* (Stratton, 2020). Bei der Sichtung englischer Literatur wird deutlich, dass diese ebenfalls keinen einheitlichen Terminus verwenden, wodurch eindeutig wird: die Diskussion über Mittel der Intensivierung findet

ohne einheitliche Terminologien statt (Os, 1989; Kirschbaum, 2002; Stratton, 2020). Innerhalb dieser Arbeit wird daher ein Terminus gewählt, der lediglich die Funktionsbedeutung beinhaltet und ebenfalls von unter anderem Os (1989) und Kirschbaum (2002) im Deutschen und von Tagliamonte und Denis (2008), Ito und Tagliamonte (2003) sowie Stratton (2020) im Englischen verwendet wird: *Intensivierer* (, *intensifier*‘ auf Englisch).

Die Begründung, weshalb in dieser Arbeit dieser Terminus gewählt wird, und weshalb es sinnig wäre, diesen im Allgemeinen zu übernehmen, ist auf drei wesentliche Argumente zurückzuführen. Erstens: einige der aufgezählten Begriffe schränken die Elemente, die eine Intensivierung erzeugen, fälschlicherweise ein, wie zum Beispiel Intensivpartikel sowie Gradadverb, die Intensivierer auf Partikel und Adverbien beschränken. Partikel bezeichnen unflektierbare Einheiten innerhalb eines Satzes, die nicht als ein Satzglied oder Attribut erscheinen können und bestehen aus den Subkategorien *Abtönungspartikel* (wie *aber, eben, eigentlich*), *Fokuspartikel* (wie *allein, genau, lediglich*), *Antwortpartikel* (wie *ja, nein, doch*), *Negationspartikel* (*nicht*) und Intensitätspartikel (wie *beinahe, fast, sehr*) (Granzow-Emden, 2019).¹ Adverbien sind zwar ebenfalls nicht flektierbar, können jedoch im Gegensatz zu den anderen Wörtern, die nicht flektiert werden können, allein als Satzglied im Vorfeld auftreten oder als Attribut erscheinen (wie *eigentlich, oft, vielleicht*) (Hentschel und Weydt, 2013; Granzow-Emden, 2019). Anhand dieser beiden Termini wird bereits deutlich, dass die Wortklasse der Intensivierer keine geschlossene ist, die mittels Partikel oder Adverbien beschrieben werden kann, worauf innerhalb von Kapitel 3.3 näher eingegangen wird. Tendiere man jedoch dazu, bei diesen Bezeichnungen zu bleiben, müsste für jeden Intensivierer einer anderen Wortklasse ein eigener Begriff definiert werden, was eine unnötige Praxis sowie eine Ansammlung vieler Termini wären. Weiter gäbe es dann noch immer das Problem, dass kein gesammelter Oberbegriff existiert. Zweitens behauptet Os (1989:1), dass es „nicht sehr elegant“ sei, einen Terminus zu wählen, der „Grad-“ im Namen beinhaltet. Dies begründen er sowie Laitenberger (2016) damit, dass die Graduierbarkeit oder auch Steigerung eines Lexems nach Lutzeier (1981, zitiert aus Os (1989)) auf die morphologische Kategorie eines Adjektivs zurückzuführen sei und Auskunft darüber gäbe, ob ein Lexem den Komparativ und Superlativ zulasse. Gerade in der

¹ Die Bezeichnungen und Definitionen der einzelnen Subkategorien können im Diskurs variieren (Hentschel und Weydt, 2003).

Wissenschaft ist es wichtig, einheitliche Terminologien mit eindeutigen Definitionen zu verwenden, um ein hohes Maß der Eindeutigkeit erreichen zu können. Die Verwendung des neutraleren Ausdrucks Intensivierung und der damit einhergehenden Intensivierer ermöglicht die Unterscheidung zwischen Graduierbarkeit und Intensivierbarkeit, wodurch einige Missverständnisse bei der Diskussion von Intensivierern – und vor allem für Intensivierer von Adjektiven, weil es sich bei Adjektiven um graduierbare Lexeme handelt – vorgebeugt werden, da sich keine Terminologien schneiden. Dass die Verwendung dieser uneinheitlichen Terminologien auch in anderen Kontexten in die Irre führen kann, wird durch das folgende Beispiel gestützt: innerhalb verschiedener Grammatiken werden in den Definitionen zu Partikeln und der Auflistung ihrer Subkategorien Fokuspartikel, die mit einem Bezugselement erscheinen, das den Fokus des Satzes bildet, selbst jedoch unbetont bleiben, eindeutig von Intensitätspartikeln beziehungsweise Gradpartikeln, wie sie ebenfalls genannt werden, die zur Modifikation eines Ausdrucks verwendet und betont werden, abgegrenzt (Boettcher, 2009; Hentschel und Weydt, 2013; Granzow-Emden, 2019). Auf der Internetseite des Grammatischen Infosystems „grammis“ werden Intensitätspartikel und Fokuspartikel ebenfalls voneinander abgegrenzt. Das Problem ist aber, dass unter dem Artikel zu Fokuspartikeln Gradpartikel als „[a]ndere Bezeichnung“ (Breindl, 2018) aufgelistet sind. Gradpartikel jedoch werden ebenfalls als alternativer Begriff für Intensivierer verwendet. Der Ausdruck Gradpartikel wird demnach mit zwei verschiedenen Funktionen assoziiert, die eindeutig voneinander abzugrenzen sind. Das dritte Argument für diesen Terminus ist, dass Intensivierungen nicht nur aus Lexemen oder Morphemen bestehen können, sondern auch aus ganzen Phrasen, wie aus den Beispielen (3) und (4) zu entnehmen ist. Demnach werden Intensivierungen nicht nur durch einzelne Wortarten erzeugt, sondern auch mit Elementen, die „quer durch die grammatischen Kategorien verlaufen“ (Os, 1989:1), worauf innerhalb von Kapitel 3.3 weiter eingegangen wird.

- (3) Er ist schnell wie der Wind.
- (4) Mir ist ein bisschen übel.

3.2 Definition

Da Intensivierer in der Literatur nicht nur einen uneinheitlichen Terminus haben, sondern auch „die in der Literatur getroffenen Charakterisierungen und zugeordneten Einheiten [...] Schwankungen [aufweisen]“ (Breindl, 2009:399), folgt im Weiteren eine Definition von Intensivierern, die innerhalb dieser Arbeit verwendet wird. Wie bereits angeführt, kann mit Intensivierern die Intensität eines Ausdrucks modifiziert werden, wobei diese Modifikation im Zusammenhang mit der semantischen Kategorie des Grades steht (Quirk et al., 2000; Breindl, 2009; Hentschel und Weydt, 2013; Laitenberger, 2016). Gemeint ist damit eine imaginäre Skala, mittels derer die Bedeutungen graduierbarer Lexeme wie beispielsweise Adjektive modifiziert werden können (Quirk et al., 2000). Den Mittelpunkt dieser Skala stellt der Ausgangspunkt einer bestimmten Norm dar, von der aus ein Ausdruck verstärkt oder abgeschwächt werden kann (Quirk et al., 2000). Je nachdem, welcher Intensivierer für einen Ausdruck gewählt wurde, wird die Bedeutung des Ausdrucks auf einen bestimmten Abschnitt auf der Skala festgelegt (Breindl, 2009; Laitenberger, 2016) wie innerhalb folgender Beispiele zu sehen ist:

- (5) Der Film ist sehr sehenswert!
- (6) Der Film ist kaum sehenswert.

Mit dem Intensivierer *sehr* in Beispiel (5) wird das Adjektiv *sehenswert* verstärkt, wodurch betont wird, dass es sich um einen guten Film handeln muss. Mit dem Intensivierer *kaum* in Beispiel (6) wird jedoch dasselbe Adjektiv abgeschwächt, wodurch vermittelt wird, dass es sich um keinen guten Film gehandelt haben muss. Mehr dazu folgt innerhalb von Kapitel 3.2.2.

Elemente, die im Wirkungsbereich von Intensivierern stehen, können Adjektive, sowohl attributive (Beispiel (7)) als auch prädikative (Beispiel (8)), Adverbien (Beispiel (9)) sowie Verben (Beispiel (10)) sein (Breindl, 2009; Hentschel und Weydt, 2013). Zusammengefasst jene Elemente, die semantisch steigerbar sind.

- (7) Das Kind trägt ein sehr farbenfrohes Kleid.
- (8) Das Zimmer ist unglaublich aufgeräumt.
- (9) Unheimlich gerne komme ich mit.
- (10) Er hat enorm darum gekämpft!

Der Fokus dieser Arbeit beschränkt sich auf Intensivieren von Adjektiven wie in den Beispielen (7) und (8), was innerhalb von Kapitel 4.2 näher begründet wird. Die weiteren folgenden Definitionen gelten hingegen nicht exklusiv für Intensivieren von Adjektiven.

Adjektive sind semantisch steigerbar und erfüllen demnach die Voraussetzung, um intensiviert zu werden. Bei Adjektiven wird unterschieden zwischen *absoluten* und *relativen* Adjektiven (Quirk et al., 2000; Breindl, 2009). Absolute Adjektive gewähren nur zwei zueinander komplementäre Eigenschaften auf der Skala der Modifikation wie zum Beispiel *positiv* vs. *negativ*, *schwanger* vs. *nicht schwanger* oder *tot* vs. *lebendig*. Relative Adjektive hingegen gewähren einen Zwischenraum, einen Übergangsbereich zwischen den zueinander antonymen Polen (Breindl, 2009; Laitenberger, 2016): *klein* vs. *groß*, *dünn* vs. *dick*, *hell* vs. *dunkel*, und so weiter. Demnach sollten absolute Adjektive nicht intensivierbar sein. Allerdings gibt es sowohl innerhalb von CMC als auch in der Alltagssprache vereinzelt Beispiele dafür, dass diese Annahme nicht allgemein gültig ist und solche Adjektive intensiviert werden. Beispiel (11) ist der Titel eines Artikels in ZEIT ONLINE vom 13. Juni 2021 zum Thema Fußball. Beispiel (12) wurde das aus dem privaten Alltag entnommen. Beide demonstrieren die Intensivierung eines absoluten Adjektivs. Ähnliche Beispiele sind ebenfalls bei Breindl (2009) vorzufinden.

(11) „Neuer BVB-Coach Rose: ‚Erster Eindruck mega-positiv‘“²

(12) Die Frau sieht voll schwanger aus!

Vor allem im Kontext der Schwangerschaft wird diese Art der Intensivierung verwendet, um den Zeitpunkt beziehungsweise Zustand einer Schwangerschaft zu beschreiben. Der Ausdruck *voll schwanger* soll in diesem Kontext kenntlich machen, dass ein Schwangerschaftsbauch deutlich zu sehen ist oder andere, für eine Schwangerschaft typische Merkmale, stark ausgeprägt sind. Die Intensivierung absoluter Adjektive wird nach Breindl (2009:411) dadurch ermöglicht, dass

Prädikate, die in fachsprachlichem Kontext exakt definiert sind und komplementär ohne intermediären Bereich sind, [...] in Alltagssprachlichen Kontexten auch als skalare Prädikate uminterpretiert und „entschärft“ werden [können].

² <https://www.zeit.de/news/2021-07/13/neuer-bvb-coach-rose-erster-eindruck-mega-positiv>

Demnach wird für die Intensivierung lediglich vorausgesetzt, dass eine Skala mindestens „konzeptuell zugänglich ist“ (Breindl, 2009:411), damit eine Umin-terpretation stattfinden kann. Somit ist die Intensivierung absoluter Adjektive nicht mehr ausgeschlossen, wodurch sie ebenfalls im Skopus eines Intensivierers stehen können.

3.2.1 Abgrenzung der Intensivierer von Fokus- und Gradpartikeln

Wie bereits innerhalb von Kapitel 3.1 angerissen wurde, werden die Terminologien für Intensivierer, Fokus- und Gradpartikel alternativ verwendet. Dies ist damit zu begründen, dass sie sich in einigen Aspekten ihrer Definitionen überschneiden. Allerdings sind vor allem Intensivierer eindeutig von Fokus- und Gradpartikeln abzugrenzen. Inwieweit die Abgrenzung erfolgt, wurde bereits erläutert. Kurz zusammengefasst: Fokuspartikel erscheinen mit einem Bezugselement, das den Fokus eines Satzes bildet, und bleiben dabei selbst unbetont. Intensivierer hingegen modifizieren einen Ausdruck und werden dabei betont (Kapitel 3.1). Weiter müssen Fokuspartikel nicht wie Intensivierer unmittelbar vor ihrem Bezugswort stehen, sondern können ebenfalls nach ihrem Bezugsausdruck vorkommen oder gar in Distanzstellung dazu stehen (Kapitel 3.2.3; Breindl, 2018). Außerdem muss das Bezugselement des Fokuspartikels nicht semantisch gesteigert werden können, sondern kann wie in Beispiel (13) auch ein Eigenname sein. Wie auch bei der Darstellung der Intensivierer, ist hier das Fokuspartikel unterstrichen.

- (13) Nur Lilly hat Blumen bekommen.

Die Abgrenzung der Intensivierer von Gradpartikeln gestaltet sich als etwas schwieriger. Sie teilen sich die Eigenschaft, eine Modifikation auf einer Skala vorzunehmen (Breindl, 2009), und sind im Gegensatz zum Fokuspartikel betonte Elemente. Allerdings unterscheiden sie sich darin, welche Elemente sie modifizieren können. Im Skopus eines Gradpartikels kann nämlich ein Substantiv stehen. Würde jedoch ein Substantiv im Skopus von Intensivierern stehen, läge keine Intensivierung mehr vor, sondern lediglich ein „attributives Modifikationsverhältnis“ (Breindl, 2009:402).

3.2.2 Verstärkende und abschwächende Intensivierer

Intensivierer, die einen Ausdruck amplifizieren, werden *verstärkende* Intensivierer genannt (Beispiel (5)) und jene, die einen Ausdruck diminuieren *abschwächende* Intensivierer (Beispiel (6)) (Os, 1989). Das Adjektiv *sehenswert* in Beispiel (5) ist ein positiv konnotiertes und trägt in diesem Kontext die Bedeutung, dass es sich lohnt, sich den besagten Film anzusehen. Durch den verstärkenden Intensivierer *sehr* wird die gute Qualität des Films weiter hervorgehoben und stärker betont. Abschwächende Intensivierer hingegen regulieren die Bedeutung eines Ausdrucks wie in Beispiel (6). Beispiele (5) und (6) ähneln sich und unterscheiden sich lediglich darin, dass in Beispiel (6) anstelle des verstärkenden Intensivierers *sehr* nun *kaum* steht. Damit wird die positive Eigenschaft des Films, nämlich dass es sich lohnen würde, ihn anzusehen, abgeschwächt, wodurch interpretiert werden kann, dass der Film nicht gut ist. Beispiel (5) verstärkt und (6) schwächt somit ein positiv konnotiertes Adjektiv. Dies funktioniert jedoch auch für negativ konnotierte Adjektive, bei denen die jeweils negative Eigenschaft verstärkt (Beispiel (14)) beziehungsweise abgeschwächt (Beispiel (15)) wird:

(14) Der Film ist sehr schlecht.

(15) Der Film ist kaum schlecht.

Ob ein Intensivierer jedoch eine verstärkende oder abschwächende Funktion übernimmt, hängt nicht nur von dem gewählten Intensivierer ab, sondern kann auch durch den semantischen Kontext beeinflusst werden, wie zum Beispiel bei dem Lexem *ganz* (Hentschel und Weydt, 2013). *Ganz* zählt nämlich sowohl zu den verstärkenden als auch zu den abschwächenden Intensivierern. Welche Funktion es übernimmt, sei abhängig von dem Wort, der Konnotation des Wortes, dem Kontext und der Betonung mit dem es zusammen auftaucht (Pusch, 1981; Hentschel und Weydt, 2013).

(16) Das Spiel ist ganz einfach.

(17) Das Spiel ist ganz unterhaltsam.

In Beispiel (16) wird mittels *ganz* die Einfachheit eines Spiels betont. In Beispiel (17) jedoch wird *ganz* verwendet, um die eigentlich positive Eigenschaft, dass das Spiel unterhaltsam ist, abzuschwächen.

Beide Arten von Intensivierern werden von Forscherinnen und Forschern in weitere Subkategorien eingeteilt, die darstellen sollen, welcher Kategorie der Intensivierung der Intensivierer angehört (Quirk et al., 2000; Breindl, 2009). Dazu gehören unter anderem *maximierende Intensivierer* (Quirk et al., 2000), auch bekannt als *Grenzwert-Intensivierer* (Breindl, 2009), die einen Ausdruck insofern verstärken, dass ein Extremum der Skala (fast) erreicht ist, und *verstärkende Intensivierer* (Quirk et al., 2000), auch bekannt als *Bereichs-Intensifikatoren* (Breindl, 2009), die einen Ausdruck lediglich zwischen den beiden Extrema auf der Skala verschieben und weiter in untere, mittlere und obere Skalenabschnitte unterteilt werden können. Diese werden entwickelt, um eine Orientierung auf der Skala auf Basis des verwendeten Intensivierers zu ermöglichen. Die genannten Subkategorien von Intensivierern sind jedoch nur grobe semantische Richtungen, da die Verwendung von Intensivierern je nach gewünschtem Effekt der Sprecherin oder des Sprechers variieren können, wie innerhalb von Kapitel 3.3 weiter ausgeführt wird. Da es, ähnlich wie bei dem verwendeten Terminus, keine einheitlichen Einteilungen der Subkategorien der Intensivierer gibt und diese vorerst keinen Einfluss auf das Klassifikationsmodell haben werden, wird auf diese nicht weiter eingegangen.

Ein weiteres Phänomen bei der Intensivierung ist, dass negativ konnotierte Lexeme, die in der Regel mit gleichermaßen negativ konnotierten Bezugswörtern zusammen erscheinen, als verstärkende Intensivierer verwendet werden können, um ein positiv konnotiertes Adjektiv stärker zu betonen (Werner, 1960, zitiert aus Os (1989); Hentschel und Weydt, 2013; Laitenberger, 2016) wie in Beispiel (18) veranschaulicht:

(18) Das Lied ist schrecklich gut.

Schrecklich beschreibt ursprünglich etwas Furchtbares, Entsetzliches oder Grauensvolles, wird hier jedoch als Verstärkung des positiven Attributs *gut* verwendet.

3.2.3 Morphologische und syntaktische Intensivierer

Intensivierung kann auf zweierlei Weisen realisiert werden: syntaktisch und morphologisch. Bei der morphologischen Intensivierung wird die Modifikation

des Adjektivs durch die Bildung von Komposita erzeugt, wie anhand der folgenden Beispiele dargestellt wird.

- (19) Der Kuchen ist zuckersüß.
- (20) Dieser Winter ist superkalt.
- (21) Die Tasche ist schweineteuer.

Bei der Bildung des Kompositums wird die Präfigierung verwendet. Werden die Komposita in ihre einzelnen Lexeme zerlegt wie zum Beispiel *zuckersüß* aus Beispiel (19) in *zucker* und *süß*, steht der Intensivierer stets an vorderer und der zu intensivierende Ausdruck an nachfolgender Stelle. Kirschbaum (2002) und Renz-Gabriel (2021) merken an, dass die Intensivierungen durch die Präfigierungen überwiegend aus den Präfixen *ur-*, *erz-*, *hyper-*, *super-* und *ultra-* bestehen. Anhand der Beispiele wird jedoch ebenfalls deutlich, dass Nomen wie *Zucker* und *Schweine* ebenfalls als Intensivierer vorkommen können. Weiter gibt es umgangssprachliche Beispiele innerhalb der CMC, in denen die morphologische Intensivierung nicht nur durch Präfigierung stattfindet, sondern mittels Interfixe wie in Beispiel (22).

- (22) The food tastes fan-fucking-tastic!
,Das Essen schmeckt fan-verdammt-tastisch!‘

Hier wurde der verstärkende Intensivierer *fucking* ‚verdammt‘ zwischen der ersten und zweiten Silbe des Adjektivs *fantastic* ‚fantastisch‘ eingefügt. Weiter weist Beispiel (22) ebenfalls das in Kapitel 3.2.2 erwähnte Phänomen auf, dass negativ konnotierte Begriffe als verstärkende Intensivierer für positive Eigenschaften verwendet werden. Der Terminus *fucking* wird nämlich normalerweise dafür verwendet, um eine Genervtheit oder Frustration ausgelöst durch jemanden oder etwas, auszudrücken. Er dient hier jedoch der Funktion, das positive Adjektiv *fantastic* stärker zu betonen. Selbst wenn Beispiele wie diese vergleichsweise selten vorkommen, sollten sie dennoch nicht unkommentiert bleiben.

Syntaktische Intensivierungen können aus Vergleichsphrasen (Beispiel (3)), deiktischen Formen (Beispiel (23)), dass-Konsekutivsätzen (Beispiel (24)), entnommen aus Kirschbaum (2002)), Adjektiven (Beispiel (24)) und weiterem bestehen (Os, 1989; Kirschbaum, 2002; Stratton, 2022).

- (23) Das Wetter war so warm.

- (24) Er lügt, dass sich die Balken biegen.
(25) Die Katze kommt ziemlich häufig vorbei.

Bei der syntaktischen Intensivierung ist es, wie in Beispiel (24) zu sehen, nicht zwingend, dass der Intensivierer vor dem zu intensivierenden Ausdruck steht. Der Intensivierer kann demnach auch nach dem zu intensivierenden Ausdruck erscheinen. Diese Art der Intensivierung ist im Deutschen jedoch selten (Kirschbaum, 2002; Stratton, 2020). Aus den vorgeführten Beispielen geht ebenfalls hervor, dass Intensivierer grundsätzlich unflektiert sind. Weiter bleibt die topologische Eigenschaft von Intensivierern festzuhalten, die für den weiteren Verlauf eine wichtige Rolle spielt: Intensivierer von Adjektiven und Adverbien sind lediglich mit ihrem Bezugsausdruck verschiebbar (Altmann und Hahnemann, 2007). Dies lässt darauf rückschließen, dass Intensivierer von Adjektiven stets unmittelbar vor dem Adjektiv stehen und keine weiteren Satzglieder dazwischen erscheinen können. Ein weiteres, wichtig zu erwähnendes Phänomen ist, dass mehrere Intensivierer aufeinanderfolgend auftreten können wie in den Beispielen (26) und (27). Dies wird als *stacking* (Scheffler, unveröffentlicht) bezeichnet.

- (26) Das ist so megacool!
(27) Julian mag Blumen schon ganz gerne.

Weitere Eigenschaften, wie der Satzgliedstatus der Intensivierer, werden vorerst nicht weiter für das zu implementierende Modell benötigt, weswegen bei Interesse an diesen auf die Literaturen von Bierwisch (1987), Engelen (1990), Helbig und Buscha (2007) sowie Altmann und Hahnemann (2007) verwiesen wird.

3.3 Intensivierer – eine offene Wortklasse

Wie bereits an mehreren Stellen vorgegriffen, handelt es sich bei den Intensivierern um eine offene Wortklasse. Viele Intensivierer sind Partikel oder weisen ähnliche Eigenschaften wie Partikel auf, weswegen eine Begriffsbezeichnung, die den Terminus „Partikel“ beinhaltet, nachvollziehbar, jedoch nicht korrekt ist (Kapitel 3.1). Intensivierer können darüber hinaus auch aus Nomen, Adjektiven, Adverbien, Vergleichsphrasen und weiterem bestehen (Kapitel 3.2.3). Letztendlich sind sie eine stark wandelnde und offene Klasse. Einzusehen ist

dies anhand verschiedener Studien, bei denen beobachtet werden kann, dass in untersuchten Datensätzen neue Intensivierer erscheinen und andere bekannte immer weniger bis nicht mehr verwendet werden (Ito und Tagliamonte, 2003; Tagliamonte, 2016; Stratton, 2022). Doch wie kommen diese stark wandelnden und innovativen Formen zustande? Fest steht, dass sich Sprache immer in einem Wandel befindet und vielen Einflüssen ausgesetzt ist. Soziolinguistische Studien wie von Ito und Tagliamonte (2003), Tagliamonte (2016), Fuchs (2017) und Stratton (2022) können empirisch nachweisen, dass vor allem die Verwendung von Intensivieren von verschiedenen Faktoren abhängig ist, zu denen das Alter, das Geschlecht, das verwendete Medium (gesprochene Sprache, Geschriebenes, CMC), die Zielgruppe (privat, öffentlich, institutionell) sowie der geographische Ursprung gehören. Vor allem das Alter hat einen starken Einfluss auf die Verwendung von Intensivierern wie aus der Studie von Tagliamonte (2016) hervorgeht, bei der ein eindeutiger Unterschied bei der Verwendung von Intensivierern über drei verschiedene Generationen hinweg nachgewiesen werden konnte. Viele der von jüngeren Generationen verwendeten Intensivierer seien sogar so neu gewesen, dass sie noch nicht in der geschriebenen Sprache vorzufinden waren und vorerst umgangssprachlich blieben (Tagliamonte, 2016). Die Entstehung dieser neuen Intensivierer kann bisher auf die drei folgenden Motivationen zurückgeführt werden: *Verblassung*: wenn ein Ausdruck unter anderem aufgrund seines häufigen Auftretens an Bedeutung für den Sprecher verliert und sich demnach nicht mehr als ausreichend erweist, um die gewollte Intensität hinter dem gewünschten Ausdruck auszudrücken (siehe auch Ito und Tagliamonte, 2003). *Gruppenzugehörigkeit*: Sprache wird oft verwendet, um sich als Teil einer Gruppe identifizieren zu können. Welche Art von Gruppe ist hier offengehalten und kann Leute des gleichen Alters, der gleichen Tätigkeit oder Ähnliches bezeichnen. Gerade bei Jugendlichen ist die Jugendsprache als Gruppenphänomen bekannt und Thema einiger Forschungen wie von Augenstein (1998) und Bahlo et al. (2019). Demnach liegt es nahe, dass neue Intensivierer etabliert und als Stilmarkierungen verwendet werden, um sich vom Standard der alltäglichen Umgangssprache abzugrenzen (Bahlo et al., 2019). Dies sei nach Ito und Tagliamonte (2003) über Generationen hinweg zu beobachten, da ältere Personen den Intensivierer *very* ‚sehr‘, eine spätere Generation den Intensivierer *really* ‚wirklich‘ und eine weitere Generation später den Intensivierer *so* ‚so‘ frequenter verwendet als die jeweils anderen. *Originalität*: Sprachstile dienen nicht

nur als Identifizierung zu einer Gruppe, sondern auch zur Selbstdarstellung einer Person. Der Gebrauch innovativer Intensivierer kann demnach auch als Methode verwendet werden, sich so originell wie möglich auszudrücken und einen eigenen Charakter zu entwickeln. Somit sind die Hintergründe der rapide wandelnden Intensivierer geklärt. Doch wie werden innovative Intensivierer entwickelt? Neue Intensivierer entstehen zum einen durch Neologismen und zum anderen durch die so genannte *Desemantisierung*. Durch Neologismen entstandene Intensivierer sind sprachlich neu gebildete Begriffe, die aus bereits vorhandenen Begriffen, Entlehnungen oder durch Bedeutungsübertragungen entwickelt werden. Ein im Englischen häufig vorkommendes Beispiel dafür ist:

(28) I'm hella fast! ,Ich bin ganz/total schnell!‘

Viele der neuen und innovativen Formen können auf die Desemantisierung zurückgeführt werden. Bei der Desemantisierung verblasst die ursprüngliche Bedeutung eines Lexems, wodurch die Bedeutung in semantisch gegenläufigen Kontexten und somit als Funktionselement verwendet werden kann (Breindl, 2009; Stratton, 2020), wie beispielsweise der häufig vorkommende Intensivierer *sehr*. Dieser stammt ursprünglich aus dem Althochdeutschen (*sêr*) und wurde dort sowohl als Adjektiv für *verwundet* oder *schmerzvoll*, als auch als Nomen mit der Bedeutung *Schmerz* verwendet (Stratton, 2020). Im Mittelhochdeutschen wandelte sich *sêr* zu *sêre* um, einem Adverb, das *schmerzlich* bedeutet. Aus *sêre* wurde *sehr* mit der Bedeutung *im hohen Maß*, das heutzutage im Neuhochdeutschen als Intensivierer verwendet wird und in keinem Bezug mehr zur ursprünglichen Bedeutung steht. Mehr Details zu diesem Prozess sind bei Ito und Tagliamonte (2003) vorzufinden.

4. Das Korpus

Innerhalb dieses Kapitels wird das Korpus, das für diese Arbeit verwendet wurde, vorgestellt sowie die einzelnen Schritte der Vorverarbeitung der Daten für das Modell erläutert.

4.1 Informationen zum Korpus

Bei dem verwendeten Korpus handelt es sich um einen Teil aus dem TwiBloCoP³ (Twitter+Blog Corpus – Parenting), einem multimedialen Textkorpus, generiert von Scheffler et al. (noch unveröffentlicht). TwiBloCoP besteht insgesamt aus 468 anonymisierten Blogposts entnommen aus selbstgeführten Blogs und 81.440 Tweets aus Twitter von insgesamt 44 Personen über den Zeitraum Oktober 2016 bis Februar 2017 und befasst sich inhaltlich mit deren Familienleben oder familienbezogenen Thematiken. Das Korpus steht wissenschaftlichen Arbeiten auf Anfrage zur Verfügung, wodurch Korpusdaten in Form von Rohtext, satzsegmentierte oder tokenisierte XML-Dateien sowie Vorverarbeitungen in Form von Markierungen ausgewählter Modalpartikeln und Intensivierer bereitgestellt werden. Die Annotation der Intensivierer begann zur selben Zeit wie die Anfertigung dieser Arbeit und war zu ihrer Fertigstellung noch nicht vollends abgeschlossen. Insofern können vorerst nur die annotierten Blogposts mit etwa 24.974 Sätzen und 360.000 Tokens verwendet werden, bei denen es noch keinen Austausch unter den annotierenden Personen gab, um Übereinstimmungen und Unstimmigkeiten bei den Annotationen zu besprechen und zu bearbeiten, worauf bei der Diskussion der Evaluationsergebnisse und den Optimierungsmöglichkeiten in Kapitel 6.2 geachtet werden muss. Mittels einer Wörterliste von Intensivierern erfolgte eine automatische Annotation der Korpusdaten, die anschließend manuell von einem Team bestehend aus vier Personen überprüft wurde. Dabei wurden Markierungen, die keine Intensivierer waren, gelöscht und neue hinzugefügt, die nicht über die Wörterliste erkannt wurden. Annotiert wurden alle Intensivierer, syntaktische und morphologische, unabhängig davon, was sie intensivieren, weswegen nicht nur Intensivierer von Adjektiven, sondern auch von Verben, Negationen und anderem gekennzeichnet sind. Für diese Arbeit wurden die tokenisierten und annotierten Dateien direkt von der Annotationsplattform WebAnno extrahiert.

4.2 Vorbereitung der Korpusdaten

Damit die extrahierten Daten maschinell verarbeitet werden können, müssen diese vorher vorbereitet werden. Dies geschieht ebenfalls mittels einer

³ <http://staff.germanistik.rub.de/digitale-forensische-linguistik/forschung/textkorpus-sprachliche-variation-in-sozialen-medien/>

Implementierung, innerhalb der die Inhalte der Dateien eingelesen, einzelne tokenisierte Sätze mittels der bereits in Tokens segmentierten Datei generiert, jedem Token ein PoS-Tag zugewiesen und anschließend jedes Token mit einem Label gemäß der BIO-Annotation ergänzt wird. Dazu im Folgendem mehr. Für das PoS-Tagging wurde das Pythonpaket `SoMeWeTa`⁴ (Social Media und Web Tagger) verwendet, das innerhalb der Gemeinschaftsaufgabe von EmpiriST (2015) von Proisl (2018) entwickelt wurde. `SoMeWeTa` erzielt mittels der möglichen Domänenanpassung an deutsche Web- und Social Media Texte eine Genauigkeit von 91,55 Prozent beim Zuteilen der PoS-Tags und eignet sich somit bestens für das PoS-Tagging des vorliegenden Korpus. Da Intensivierer nicht nur aus einzelnen Tokens, sondern auch aus einer Sequenz an Tokens bestehen können wie in Beispiel (4), reicht eine binäre Annotation von *Intensivierer* und *kein Intensivierer* nicht aus. Stattdessen wird das BIO-Tagging verwendet, bei dem Tokens, die keine Intensivierer sind, mit *O* (outside of intensifier ‚außerhalb eines Intensivierers‘) annotiert werden, Tokens, die den Beginn einer Sequenz als Intensivierer markieren oder als einzelnes Token als Intensivierer fungieren, mit *B-ITSF* (beginning of intensifier ‚Anfang eines Intensivierers‘) und Tokens, die zu einer Sequenz gehören und auf den ersten Token der Sequenz folgen, mit *I-ITSF* (inside of intensifier ‚innerhalb eines Intensivierers‘). Ein Beispiel ist innerhalb der Abbildung 4.1 dargestellt. Weil Intensivierer eine komplexe Klasse sind und es sich hier um einen ersten konkreten Versuch handelt, diese automatisch in einem Text zu markieren, beschränkt sich das in Kapitel 5 präsentierte Modell `TafIn` (Tagger für Intensivierer) vorerst auf Intensivierer von Adjektiven. Da jedoch nicht nur Intensivierer von Adjektiven, sondern auch von beispielsweise Verbphrasen annotiert wurden, wird bei der Erstellung der BIO-Tags ebenfalls darauf geachtet, dass nur Tokens und Sequenzen, die vor einem Adjektiv stehen, die jeweiligen Annotationen *B-ITSF* und *I-ITSF* erhalten. Dadurch, dass es bei der Annotation der Korpusdaten ebenfalls nicht möglich war, das intensivierende Element bei einem morphologischen Intensivierer einzeln zu markieren, sondern das ganze Kompositum als Intensivierer markiert wurde, folgte in einem weiteren Schritt die Selektierung aller Adjektive, die als Intensivierer annotiert wurden. Diese Adjektivkomposita wurden mittels des Moduls `CharSplit`⁵ in ihre einzelnen Lexeme zerlegt, für die ebenfalls PoS-

⁴ <https://github.com/tsproisl/SoMeWeTa>

⁵ <https://github.com/dtuggener/CharSplit>

Tags generiert wurden. Weiter erhielt das erste Lexem des Kompositums die Annotation B-ITSF und das zweite Lexem die Annotation O. Daraus resultiert eine Annotation der Form *Token, PoS-Tag, BIO-Label* für jedes einzelne Token (siehe Tabelle 4.1). Zuletzt werden die Sätze, deren einzelne Tokens eine Annotation beinhalten, mit der Funktion `train_test_split`⁶ von `sklearn` in Trainings- und Testdaten zu jeweils 80 und 20 Prozent aufgeteilt. Die Trainingsdaten dienen dazu, das Modell zu trainieren und Gewichte von Merkmalen zu optimieren. Die Testdaten bleiben dem Modell hingegen unbekannt, sodass Ergebnisse aus der Trainingsphase unbeeinflusst an den Testdaten angewandt werden können.

Token	PoS-Tag	BIO Label
Das	ART	O
Wetter	NN	O
war	VAFIN	O
so	PTKIFG	B-ITSF
warm	ADJD	O
.	\$.	O

Tabelle 4.1. Beispielhafte Darstellung der Annotation der einzelnen Tokens nach der Vorverarbeitung.

5. Tagger für Intensivierer (TafIn)

Bei der Analyse von Intensivierern ist es unausweichlich, diese vorerst herauszusuchen und zu annotieren. Dieser Prozess findet bisher oft mit der Verwendung von Wörterlisten statt (Kapitel 4.1; Carrillo-de-Albornoz und Plaza, 2013). Dies bringt zwei Probleme mit sich: zum einen sind Intensivierer eine offene und wandelnde Klasse, weswegen mit Wörterlisten keine neuen identifiziert werden können. Zum anderen werden einige Tokens wie *so* und *zu* fälschlicherweise als Intensivierer markiert. Eine Wörterliste kann nämlich zum Beispiel nicht unterscheiden, ob es sich bei *zu* um einen Intensivierer handeln soll wie in Beispiel (29) oder lediglich um ein Partikel vor einem Infinitiv wie in (30).

(29) Ich habe zu viel gegessen.

(30) Ich habe noch zu lernen.

Beides muss durch eine zeitaufwendige, manuelle Durchsicht der Daten ausgebaut werden. Demnach ist es für weitere Analysen von Intensivierern von

⁶ https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

großem Interesse, diese mindestens zu einem Großteil und bestenfalls komplett automatisch identifizieren zu können. Ein erster Ansatz, der hier verfolgt wurde, ging davon aus, dass es sich bei dieser Identifikation lediglich um die Entscheidungsfrage handelt, ob ein Token ein Intensivierer ist oder nicht. Diese Art der Entscheidungsfrage wird als *Klassifikation* bezeichnet, da ein Element einer vordefinierten Kategorie zugeordnet werden soll (Jurafsky und Martin, 2022). Um ein Modell zu implementieren, das eine Klassifikation durchführt, gibt es mehrere Algorithmen, die verwendet werden können. Zu den bekanntesten und klassischen zählen unter anderem der *Naive Bayes* Algorithmus (NB) und die *logistische Regression* (LR). Bei der Bestimmung von PoS-Tags wird häufig auf das *Hidden Markov Modell* (HMM) zurückgegriffen. Nach ersten Implementierungsversuchen stellte sich jedoch heraus, dass sich keiner dieser drei Algorithmen für die Identifikation von Intensivierern eignet. Die Hauptgründe dafür sind, dass alle drei Modelle mit voneinander unabhängigen Merkmalen arbeiten, keine Sequenzen erkennen und annotieren können, mit einer großen Anzahl an komplex einzubringender Spezifizierungen arbeiten müssten, um so präzise wie möglich sein zu können, sowie nur mit den Informationen eines jeweiligen Tokens x_i arbeiten können, um ein dazu passendes Label y_i generieren zu können (Lafferty et al., 2001; Sutton und McCallum, 2007). Wie innerhalb von Kapitel 3 ersichtlich wurde, ist das Lexem vor einem Intensivierer irrelevant, um zu bestimmen, ob ein Lexem ein Intensivierer ist oder nicht. Es gibt nämlich viele Fälle, in denen ein Intensivierer zum Beispiel am Satzanfang steht oder nicht direkt auf ein Kopulaverb folgt. Wichtiger ist, was nach dem Intensivierer erscheint. Die Hindernisse, weswegen sich keines der genannten Modelle für die Identifikation von Intensivierern eignet, können mit dem *Conditional Random Field* (CRF) überwunden werden. Es kann nämlich voneinander abhängige Merkmale erkennen, Sequenzen annotieren mehr Merkmalsinformationen extrahieren, sowie mehr Informationen als nur von einem x_i verwenden (Lafferty et al., 2001; Sutton und McCallum, 2007; Sikdar und Gambäck, 2016; Chavan, 2019; Jurafsky und Martin, 2022), wie innerhalb der folgenden Kapitel erläutert wird.

Bevor auf die Implementierung eingegangen werden kann, muss darauf hingewiesen werden, dass `TafIn` die erste offizielle Implementierung ist, die sich konkret mit der Identifizierung von Intensivierern von Adjektiven befasst. Wie aus Kapitel 3 hervorgekommen ist, handelt es sich dabei um eine sehr komplexe

Klasse, die in verschiedenen Formen vorkommen kann. `TafIn` fokussiert sich demnach vorerst darauf, ein Grundgerüst für die Identifizierung von Intensivierern zu etablieren, das zu einem späteren Zeitpunkt weiter ausgebaut werden kann. Konkret bedeutet dies, dass nur die Identifizierung von Intensivieren, die unmittelbar vor Adjektiven stehen, angezielt wird. Identifikationen von Interfixen, Stacking, Intensivieren von Adverbien, Negationen oder Ähnlichem fallen vorerst aus. Daher sind im Folgenden, wenn Intensivierer erwähnt werden, konkret Intensivierer von Adjektiven gemeint. Weiter ist die folgende Implementierung nicht von Grund auf programmiert worden, sondern basiert auf der CRF-Implementierung `python-crfsuite`⁷ von Peng und Korobov (2020).

5.1 Conditional Random Field

CRF ist ein Algorithmus, der oft bei der *Named Entity Recognition* ‚Erkennung namentlicher Einheiten‘ verwendet wird (Sutton und McCallum, 2007; Jurafsky und Martin, 2022). Bei der Erkennung von namentlichen Einheiten wie Personennamen, Ortsnamen, Organisationen und so weiter in einem Text handelt es sich um einen Prozess, bei dem innerhalb von einer Sequenz an Wörtern eine Einheit erkannt und mit einem vordefinierten jeweiligen Label versehen wird. Bei der Erkennung von Intensivierern handelt es sich um einen ähnlichen Prozess: innerhalb einer bestimmten Satzsequenz soll erkannt werden, bei welchen Tokens es sich um einen Intensivierer oder eine Intensivierersequenz handelt. Hierbei können verschiedene, allgemeine Merkmale eingebunden werden, wodurch eine Konfrontation mit unbekannten Wörtern in einem Testdatensatz weniger problematisch ist als bei den anderen Modellen oder Wörterlisten (Sutton und McCallum, 2007; Chavan, 2019). Wie bei der NER ist es nämlich auch hier der Fall, dass es innerhalb der Trainingsdaten, an denen ein Modell trainiert werden muss, weniger Intensivierer gibt als andere Tokens. Weiter sind Intensivierer eine offene und innovative Klasse, sodass auch neue Intensivierer innerhalb eines Testkorpus vorkommen können, weswegen es praktischer ist, diese anhand eines Kontextes erkennen zu können. Diese Merkmale können Kriterien sein wie: Wortart, vorheriges Wort, nachfolgendes Wort, Wortlänge, Groß- oder Kleinschreibung und so weiter. Ist durch die Merkmale beispielsweise bekannt,

⁷ Link zum Pythonpaket: <https://pypi.org/project/python-crfsuite/> und Link zum GitHub Repository: <https://github.com/scrapinghub/python-crfsuite>

dass vor dem zu markierenden, unbekannten Token eine Determinante steht, kann ausgeschlossen werden, dass es sich bei dem aktuellen Token um ein Verb handeln wird (Jurafsky und Martin, 2022).

5.2 Aufbau des CRFs

Im Folgenden wird auf die allgemeine Struktur des CRFs, auf die Generierung der Merkmale sowie auf den Vorgang der Inferenz und des Trainings eingegangen. Dargestellte Gleichungen wurden aus Jurafsky und Martin (2022) entnommen oder sind an diesen angelehnt und werden, um auf sie referieren zu können, ebenfalls mit der Nummerierung aus dem Buch selbst in Klammern ergänzt.

5.2.1 Allgemeines

Ziel des CRFs ist es, mit der Berechnung der bedingten Wahrscheinlichkeit einer Sequenz X , bestehend aus einzelnen Tokens $x_i \dots x_n$ ($X = x_i \dots x_n$), eine Sequenz Y , bestehend aus Labels für jedes Token, zu generieren ($Y = y_i \dots y_n$), wobei jedem einzelnen Token x_i ein Label y_i zuzuordnen ist (Sha und Pereira, 2003; Jurafsky und Martin, 2022). Sei Beispiel (23) die gegebene Inputsequenz. Dann sehen die vereinfachte Inputsequenz X und die gewünschte Outputsequenz Y wie folgt aus:

$$X = ("Das", "Wetter", "war", "so", "warm", ".")$$

$$Y = ("O", "O", "O", "B - ITSF", "O", "O")$$

Um die beste Outputsequenz Y für eine Inputsequenz X zu bestimmen, müssen vorerst alle möglichen Outputsequenzen Y ermittelt werden, wovon jene mit der höchsten Wahrscheinlichkeit ausgewählt wird (Jurafsky und Martin, 2022). Um dies umzusetzen, benötigt das CRF Input-/Satzsequenzen X , Output-/Labelsequenzen Y sowie K Merkmale mit einem Gewicht w_k für jedes Merkmal F_{ik} (Jurafsky und Martin, 2022). Die bedingte Wahrscheinlichkeit einer Outputsequenz Y für eine Inputsequenz X wird mit der Gleichung

$$p(Y|X) = \frac{\exp(\sum_{k=1}^K w_k F_k(X, Y))}{\sum_{Y' \in Y} \exp(\sum_{k=1}^K w_k F_k(X, Y'))}$$

Gleichung 5.1 (8.23)

ermittelt. Der Nenner der Gleichung kann hier rausgezogen und mittels $Z(X)$ dargestellt werden, wodurch die Gleichung der bedingten Wahrscheinlichkeit $p(Y|X)$ wie folgt umgeschrieben werden kann:

$$p(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{k=1}^K w_k F_k(X, Y)\right)$$

Gleichung 5.2 (8.24)

mit

$$Z(X) = \sum_{Y' \in Y} \exp\left(\sum_{k=1}^K w_k F_k(X, Y')\right)$$

Gleichung 5.3 (8.25)

Die K Funktionen $F_k(X, Y)$ werden als *globale Merkmale* bezeichnet und stellen die Merkmale einer ganzen Inputsequenz X und Outputsequenz Y dar. Um die globalen Merkmale zu bestimmen, werden vorerst die Summen der *lokalen Merkmale* für jede Position i in Y berechnet mittels:

$$F_k(X, Y) = \sum_{i=1}^n f_k(y_{i-1}, y_i, X, i)$$

Gleichung 5.4 (8.26)

Jedes lokale Merkmal f_k ist in der Lage auf Informationen des Outputs des jeweiligen y_i , des Outputs des vorherigen Tokens y_{i-1} , der gesamten Inputsequenz X sowie Teile davon und ebenso auf die Informationen der aktuellen Position i zuzugreifen (Sha und Pereira, 2003; Sutton und McCallum, 2007). Durch die Eingrenzung auf den lediglich benachbarten Output $y_i - 1$ wird dieses CRF als ein *linear chain CRF*, ‚Linearketten‘ CRF, bezeichnet (Jurafsky und Martin, 2022). Ohne diese Einschränkung könnten Informationen aus einem weiter distanzierten Output wie y_{i-3} verwendet werden, was in diesem Fall jedoch nicht benötigt wird. Allgemein wird das uneingeschränkte CRF Modell nach Jurafsky und Martin (2022) weniger für die Verarbeitung natürlicher Sprachen verwendet.

5.2.2 Merkmale

Das Besondere an CRF ist, dass im Gegensatz zum HMM Merkmale mittels der Merkmalsfunktionen simpel eingebaut werden können. Die Auswahl der Merkmale, die bei der Erkennung von Intensivierern helfen und für das Modell übernommen werden, wird manuell bestimmt. Zu den Merkmalen, die innerhalb von `TafIn` für jedes Token generiert werden, wurden Folgende gewählt:

token:	das Token x_i , das gelabelt werden soll,
token.islower():	ob das Token groß oder klein geschrieben ist,
postag:	welchen PoS-Tag das Token hat.

Tokens, die am Satzanfang oder -ende stehen, werden um die jeweiligen Merkmale

bos:	Beginn eines Satzes ‚begin of sentence‘ und
eos:	Ende eines Satzes ‚end of sentence‘

ergänzt. Tokens, die weder ein **bos** noch ein **eos** Merkmal erhalten, werden um folgende Merkmale erweitert:

prev_token:	Angabe des vorangehenden Tokens,
prev_postag:	PoS-Tag des vorangehenden Tokens,
prev_token.islower():	ob das vorangehende Token groß oder klein geschrieben ist,
prev_label:	Klasse des vorangehenden Tokens,
next_token:	Angabe des folgenden Tokens,
next_token.islower():	ob das folgende Token groß oder klein geschrieben ist,
next_postag:	PoS-Tag des nachfolgenden Tokens.

Einige der Merkmale werden allgemein verwendet, um Intensivierer anhand eines abstrakten Kontexts identifizieren zu können. Merkmale wie **token.islower()** in Kombination mit **bos** hingegen helfen konkret dabei, Intensivierer an Satzanfängen zu erkennen, da sie, bei korrekter Rechtschreibung, nur dort groß geschrieben werden. Bei der Erkennung von Sequenzen sind **prev_postag**

sowie **prev_label** nützlich und **next_postag** hilft bei der Erkennung, ob das folgende Token ein Adjektiv ist und dementsprechend überhaupt intensiviert werden kann. Ebenfalls von großem Interesse wäre das Merkmal gewesen, ob ein Token einen Kontext semantisch steigern kann. Eine solche Annotation hätte jedoch den Umfang dieser Arbeit überschritten, weswegen auf diese Eigenschaft vorerst verzichtet werden musste. Ein Beispiel, wie die generierten Merkmale für jedes Token aussehen können, wird anhand des Tokens *so* aus dem Satz „*Das Wetter war so warm.*“ demonstriert:

```
{'bias': 1.0, 'token': 'so', 'token.islower()': True, 'postag':  
'PTKIFG', 'prev_token': 'war', 'prev_postag': 'VAFIN', 'prev_to-  
ken.islower()': True, 'prev_label': 0, 'next_token': 'warm',  
'next_token.islower()': True, 'next_postag': 'ADJD'}
```

Jedes globale Merkmal F_k wird mit dem jeweiligen Gewicht w_k multipliziert (Jurafsky und Martin, 2022). Das Gewicht w_k ist eine Zahl, die widerspiegelt, wie wichtig ein Merkmal bei der Erkennung von Intensivierern ist (Jurafsky und Martin, 2022). Hierbei wird jedoch nicht für jedes lokale Merkmal f_k ein Gewicht bestimmt. Stattdessen werden die Werte jedes einzelnen lokalen Elements innerhalb einer ganzen Inputsequenz X summiert, um daraus das globale Merkmal F_k zu bestimmen.

5.2.3 Inferenz und Training

Mit der Gleichung 5.2 werden alle möglichen Outputsequenzen Y für eine gegebene Inputsequenz X berechnet. Um die beste Outputsequenz \hat{Y} für eine Inputsequenz X zu generieren, wird aus allen möglichen Outputsequenzen Y jene gewählt, die die höchste Wahrscheinlichkeit hat, also:

$$\hat{Y} = \operatorname{argmax}_{Y \in \mathcal{Y}} P(Y|X)$$

Gleichung 5.5 (8.22)

was, sobald Gleichung 5.2 hineinsubstituiert wird, nichts anderes ist als:

$$\hat{Y} = \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \frac{1}{Z(X)} \exp \left(\sum_{k=1}^K w_k F_k(X, Y) \right)$$

Gleichung 5.6 (8.27)

Da die Exponentialfunktion keinen Einfluss auf das *argmax* hat und der Nenner $Z(X)$ konstant für eine Sequenz X ist, können diese in der Gleichung weggelassen werden. Gleichzeitig kann $F_k(X, Y)$ mit der Summe aller lokalen Merkmale ersetzt werden, woraus die folgenden Gleichungen resultieren:

$$\hat{Y} = \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{k=1}^K w_k \sum_{i=1}^n f_k(y_{i-1}, y_i, X, i)$$

Gleichung 5.7 (8.29)

$$= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{i=1}^n \sum_{k=1}^K w_k f_k(y_{i-1}, y_i, X, i)$$

Gleichung 5.8 (8.30)

Um nun die geeignetste Outputsequenz \hat{Y} zu identifizieren, wird eine $N \times T$ dimensionale Matrix der Übergangswahrscheinlichkeiten erstellt (Jurafsky und Martin, 2022). Die wahrscheinlichste Outputfrequenz kann somit ermittelt werden, indem rekursiv die Label zurückverfolgt werden, die in der vorangegangenen Spalte den maximalen Wert haben (Jurafsky und Martin, 2022):

$$v_t(j) = \max_{i=1}^N v_{t-1}(i) \sum_{k=1}^K w_k f_k(y_{t-1}, y_t, X, t) \quad 1 \leq j \leq N, 1 \leq t \leq T$$

Gleichung 5.9 (8.33)

Dieser Prozess ist ein leicht abgeänderter Prozess des HMM *Viterbi*, dessen ausführliche Erläuterung jedoch den Rahmen dieser Arbeit sprengen würde. Stattdessen wird auf das Kapitel 8 von Jurafsky und Martin (2022) verwiesen, bei denen eine ausführliche Erläuterung mittels grafischer Darstellungen vorzufinden ist.

Schlussendlich müssen noch die Gewichte der Merkmalsfunktionen trainiert werden. Dies geschieht anhand der Trainingsdaten unter denselben überwachten Algorithmen wie bereits bei der LR, weswegen auch hier der *stochastic gradient descent* (SGD) ‚stochastische Gradientenabstieg‘ angewandt wird (Jurafsky und Martin, 2022). Um zu bestimmen, ob die Gewichte erhöht oder verringert

werden müssen, die somit Auskunft darüber geben, ob eine Merkmalsfunktion relevant bei der Erkennung von Intensivierern ist, muss zunächst die *cross-entropy loss function* ‚Verlustfunktion‘ L_{CE} ermittelt werden, die vermittelt, inwieweit sich die berechnete Outputsequenz \hat{Y} von der Gold-Outputsequenz Y unterscheidet (Jurafsky und Martin, 2022). In anderen Worten: inwieweit weicht die Labelsequenz \hat{Y} , die das Modell einer Inputsequenz X zugeordnet hat, von der Labelsequenz im Goldwert Y ab? Goldwert, Goldlabels beziehungsweise Gold-Outputsequenzen sind die BIO-Annotationen, die den einzelnen Tokens vorher (Kapitel 4.2) zugeordnet wurden und somit als „richtige“ Klassifikation der einzelnen Tokens/Sequenzen gelten. Bei dem Label \hat{y}_i für einen einzelnen Token x_i einer Inputsequenz X sähe die Berechnung der loss function wie folgt aus:

$$L_{CE}(\hat{Y}, Y) = - \sum_{m=1}^M y_k \log \hat{y}_k = -\log \hat{y}_c$$

Gleichung 5.10 (8.44 und 8.45)

wobei M die Anzahl aller möglichen Labels (bei TafIn zur Erinnerung O, B-ITSF und I-ITSF) und c das korrekte Label darstellt (Jurafsky und Martin, 2022). Je größer der Wert der cross-entropy loss function ist, desto größer ist der Unterschied zwischen \hat{Y} und Y (Jurafsky und Martin, 2022). Bleibt der Wert gering, muss die Gewichtung wenig verändert werden. Resultiert jedoch ein hoher Wert, der impliziert, dass die Gewichtung des Merkmals nicht optimal ist, um eine richtige Outputsequenz zu erzeugen, ist es das Ziel, diesen Wert zu minimieren, wofür der SGD benötigt wird, um die maximale logarithmierte Wahrscheinlichkeit \hat{Y} zu erhalten. SGD ermittelt nämlich mittels der Ableitung ∇ der cross-entropy loss function die Steigung dieser und bewegt sich in die entgegengesetzte Richtung der Steigung, um dem Minimum der Funktion näher zu kommen, um so den Wert der loss function zu minimieren (Jurafsky und Martin, 2022). Die Gewichte θ werden somit bei jedem Schritt der Lernrate η , die bestimmt, wie weit sich w bewegen darf, mittels der Gleichung 5.11 aktualisiert.

$$\theta_{t+1} = \theta_t - \eta \nabla L(f(x; \theta), y)$$

Gleichung 5.11 (5.27)

Damit auch beim CRF Gewichte nicht zu sehr an die Trainingsdaten angepasst werden, werden zusätzlich die Regulierungen *Lasso Regression* (L1) und *Ridge Regression* (L2) in die Trainingsphase integriert. L1 reguliert die

Gewichte insoweit, dass es Merkmale entfernt, die den wenigsten Einfluss auf die Identifizierung von Intensivierern haben (Géron, 2019). L2 dient dazu, das Modell an die Trainingsdaten anzupassen und dabei die Gewichte so klein wie möglich zu halten (Géron, 2019). Nähere Details dazu können aus Géron (2019) entnommen werden.

5.3 Testphase

Nachdem mit den Trainingsdaten die Gewichte bestimmt wurden, können diese anschließend an den Testdaten, die dem Modell bisweilen unbekannt waren, angewandt werden. Um die beste Outputsequenz \hat{Y} aus \mathbf{Y} zu erhalten, wird die Gleichung 5.8 verwendet, wobei an Stelle von w_k jene Gewichte eingesetzt werden, die in der Trainingsphase ermittelt wurden.

6. Evaluation und Diskussion

Im Folgenden wird das Modell evaluiert, wobei ebenfalls die Methode der Evaluation erläutert wird, und anschließend die Ergebnisse diskutiert.

6.1 Evaluation

Um sehen zu können, ob `TafIn` im Gegensatz zu anderen Methoden einen Vorteil darstellt, ist es unabdingbar, das Modell zu evaluieren. Als Vergleichsmethoden zu `TafIn` werden zwei weitere Baselines zur Identifikation von Intensivierern verwendet: der PoS-Tagger `SoMeWeTa` von Proisl (2018) und eine heuristische Methode, bei der jedes Token als Intensivierer markiert wurde, das vor einem Adjektiv steht sowie innerhalb einer selbst angefertigten Liste aus 78 Intensivierern (siehe Tabelle 6.1 im Anhang) enthalten ist. Um alle drei Methoden miteinander vergleichen zu können, eignet es sich zunächst, eine *Konfusionsmatrix* zu erstellen, bei der abgebildet wird, wie viele Elemente einer Klasse vom Modell richtig klassifiziert wurden und wie viele nicht (Géron, 2019). Da die Klassifikation aus drei Labels besteht (O, B-ITSF und I-ITSF) handelt es sich hierbei um eine Evaluierung mit mehreren Klassen, wofür eine 3×3 Konfusionsmatrix erstellt wird, die mit den Goldlabels und den Ergebnissen des jeweiligen Modells ausgefüllt wird (Jurafsky und Martin, 2022). Ist die Konfusionsmatrix erstellt, können anschließend die Werte *precision* (P) und *recall* (R) berechnet

werden. P bestimmt den prozentualen Anteil der Daten, die vom Modell als O, B-ITSF und I-ITSF markiert wurden und auch im Goldlabel solche sind. R berechnet den prozentualen Wert, wie viele der mit Goldlabel als O, B-ITSF und I-ITSF markierten Tokens auch vom Modell richtig identifiziert wurden. P und R sind wie folgt definiert:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Die jeweiligen Elemente im Zähler und Nenner der Gleichungen werden im Folgenden anhand der Klasse B-ITSF beispielhaft erläutert. *True positive* repräsentiert die Anzahl an Klassifikationen, die das Modell als B-ITSF markiert hat und die auch im Goldlabel als solche markiert sind. *False positive* beinhaltet die Anzahl an Klassifikationen, die von dem Modell als B-ITSF markiert wurden, im Goldlabel jedoch I-ITSF oder O sind. *False negative* ist die Anzahl an Klassifikationen, die das Modell nicht als B-ITSF markiert hat, obwohl sie im Goldlabel als B-ITSF markiert sind. Da es sich hier nicht um eine binäre Klassifikation handelt, müssen die Berechnungen von P und R etwas abgeändert werden. Dafür gibt es zwei Möglichkeiten. *Macroaveraging*, Makro-Mittelwertbildung, berechnet für jede einzelne Klasse P und R Werte und bestimmt anschließend, wie gut diese im Durchschnitt sind. *Microaveraging*, Mikro-Mittelwertbildung, sammelt alle Klassifikationen innerhalb einer 2×2 Konfusionsmatrix und bestimmt anschließend deren P und R Werte (Jurafsky und Martin, 2022). Die Klassen, die häufiger vergeben wurden, haben bei letzterer Option einen größeren Einfluss auf das Ergebnis der P und R Werte, sodass weniger frequente Klassen, wie B-ITSF und I-ITSF untergehen können. Daher werden nur die macroaveraging P und R Werte berechnet, um statistisch besser widerspiegeln zu können, wie die Klassifikationen von B-ITSF und I-ITSF abgeschnitten haben (Jurafsky und Martin, 2022). Tabelle 6.2 repräsentiert die zu den jeweiligen Modellen einzelnen macroaveraging Werte zu P, R und F_1 , die mit `classification_report`⁸ von `sklearn` ermittelt wurden. F_1 ist eine Metrik, die den harmonischen Durchschnitt der P und R Werte repräsentiert und demnach

⁸ https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

verwendet werden kann, um P und R Werte in einem zu kombinieren (Géron, 2019). Je höher P und R eines Modells sind, desto höher ist der F_1 Wert.

	Macroaverage P	Macroaverage R	Macroaverage F_1
Baseline 1	0.45	0.49	0.47
Baseline 2	0.41	0.51	0.44
TafIn	0.84	0.77	0.80

Tabelle 6.2. Abbildung der einzelnen macroaverage P, R und F_1 Werte der einzelnen evaluierten Modelle.

Wie die Modelle im Einzelnen bei der Klassifikation der jeweiligen Klassen abgeschnitten haben, ist tabellarisch im Anhang vorzufinden, weswegen hier nur auf besondere Auffälligkeiten eingegangen wird. Bereits anhand der durchschnittlichen F_1 Werte der Modelle ist zu erkennen, dass TafIn bessere P und R Werte erzielt. Während die Baselines 1 und 2 nur circa 41 und 45 Prozent der Daten als solche markieren, wie sie es im Goldlabel auch wirklich sind (P), sind es bei TafIn über 80 Prozent. Der prozentuale Anteil der Daten, die vom Modell richtig identifiziert wurden (R), ist bei allen drei Modellen gering, jedoch liefert TafIn auch hier bessere Ergebnisse mit 77 Prozent, 26 Prozentpunkten mehr als Baseline 2, die 51 Prozent erzielt, und 28 Prozentpunkten mehr als Baseline 1, die 49 Prozent der jeweiligen Klassen erkennt. Weiter auffällig sind die Werte bei der Klasse I-ITSF, die Fähigkeit, Sequenzen von Intensivierern zu identifizieren. Von allen drei Modellen stellt sich TafIn als einziges heraus, das Sequenzen von Intensivierern als solche erkennt (Tabelle 6.3, 6.4 und 6.5 im Anhang). Baseline 1 und 2 haben für die Klasse I-ITSF jeweils P, R und F_1 Werte von 0 Prozent. TafIn hingegen hat einen P Wert von 60 Prozent und einen R Wert von 46 Prozent. Das bedeutet, dass mehr als die Hälfte der von TafIn markierten I-ITSF auch im Goldlabel als solche markiert sind und etwas weniger als die Hälfte der im Goldlabel als I-ITSF markierten Tokens tatsächlich erkannt wurden.

6.2 Diskussion und Optimierungsmöglichkeiten

Betrachtet man die Werte von TafIn, sind diese nicht ausreichend für eine komplett automatische Annotation von Intensivierern von Adjektiven. Werden die Evaluationen der Baselines 1 und 2 jedoch hinzugezogen, ist zu erkennen,

dass `TafIn` wesentlich besser bei der Identifizierung ist als diese. Demnach ist es jedenfalls hilfreicher bei der Erkennung von Intensivierern als bisherige Methoden, die nicht manuell stattfinden. Denn der Evaluation nach ist `TafIn` nicht nur doppelt so gut darin, einzelne Tokens als Intensivierer zu erkennen, sondern auch im Vergleich als einziges in der Lage, Sequenzen zu annotieren. Weiter handelt es sich hierbei um einen ersten konkreten Versuch der Identifizierung, sodass es eine Reihe an Optimierungsmöglichkeiten gibt, wodurch die P und R Ergebnisse von `TafIn` verbessert werden können. Das bedeutet, selbst wenn `TafIn` allein bisher keine zufriedenstellenden Ergebnisse erreicht, ist es im Vergleich zu anderen maschinellen Methoden dem Ziel der automatischen Annotation ein großes Stück näher gekommen, womit es eine gute erste Grundlage bildet.

Diese Grundlage kann wie folgt optimiert werden: zum Beispiel kann eine überarbeitete Annotation bereits Auswirkungen auf die Ergebnisse haben, für den Fall, dass Daten falsch oder unterschiedlich annotiert wurden, wodurch das Trainieren der Gewichte negativ beeinflusst werden kann. Weiter fehlt die Information zu Tokens, ob sie Inhalte semantisch steigern können oder nicht, welche – so die Hypothese – ebenfalls einen großen Einfluss auf die Ergebnisse haben sollte. Eine dritte Optimierungsmöglichkeit wäre die Integration der bisher verwendeten Wörterlisten, mit denen ein ähnlicher Effekt wie die Verwendung eines Ortlexikons bei der NER entstehen sollte (Jurafsky und Martin, 2022). Zuletzt wäre es noch eine Möglichkeit zu testen, ob die Entfernung von *stopwords*, ‚Stoppwörter‘, also Wörtern, die häufig in einem Korpus auftreten, jedoch keine bedeutungstragende Relevanz bei der Erfassung gewünschter Inhalte besitzen, einen Einfluss hat.

7. Fazit

Das Ziel dieser Arbeit war es zum einen, einen Überblick über Intensivierer und ihren Eigenschaften zu gewinnen und zum anderen ein Modell zu implementieren, das Intensivierer von Adjektiven automatisch identifiziert und als solche markiert. Bei der Erläuterung von Intensivierern wurde schnell ersichtlich, dass es sich dabei um ein viel untersuchtes, schnell wandelndes und somit komplexes Thema handelt. All diese komplexen Eigenschaften einer solch rapide wandelnden Klasse, die in verschiedenen Formen auftreten kann, in dessen

Skopus verschiedene Elemente stehen können, innerhalb eines ersten Ansatzes der automatischen Identifizierung zu erkennen, ist ein großes Ziel. Daher wurde hier der Fokus daraufgelegt, ein erstes Modell und somit ein Grundgerüst für die Identifikation von Intensivierern zu etablieren. Dafür wurde sich zunächst auf die Identifikation von Intensivierern von Adjektiven konzentriert. Um diesen ersten Schritt zu erreichen, mussten annotierte Korpusdaten mit einer neuen Annotation versehen und vorbereitet werden. Anschließend konnte `TafIn`, basierend auf dem CRF Algorithmus trainiert werden, wobei zu jedem ausgewählten Merkmal, das bei der Erkennung von Intensivierern helfen soll, Gewichte berechnet wurden, die bestimmen, wie hilfreich es für die Identifizierung ist. Für die Evaluierung der Ergebnisse wurden zwei weitere Baselines hinzugezogen: die Sammelklasse PTKIFG von `SoMeWeTa` sowie die gängige Identifizierung mittels Wörterlisten. Die Resultate zeigen jedoch eindeutig, dass `TafIn` im Vergleich zu den beiden Baselines eine bessere Alternative darstellt, vor allem im Hinblick auf die Erkennung von Sequenzen von Intensivierern, und somit die gewünschte, erste Grundlage in Richtung automatische Identifikation von Intensivierern von Adjektiven darstellt. Weiter bestehen einige Optimierungsmöglichkeiten, die bisher nicht umgesetzt werden konnten, weil sie den Rahmen dieser Arbeit gesprengt hätten, mit denen es aber möglich und realistisch ist, noch bessere Werte mit `TafIn` zu erreichen. Demnach ist es wünschenswert, wenn `TafIn` als Motivation oder Ansatz verwendet wird, um das Ziel der automatischen Annotation weiter zu verfolgen. Unter anderem dafür ist `TafIn`⁹ öffentlich zugänglich und lädt zu einer Weiterentwicklung ein. Hierbei können die bereits genannten oder auch neue Optimierungsmöglichkeiten umgesetzt werden sowie die Identifikation auf Intensivierern von Negationen, Adverbien, Verben oder auch Klassifikation verschiedener Subkategorien (Kapitel 3.2.2) und so weiter ausgeweitet werden.

⁹ <https://github.com/imgeyuez/TafIn--Tagger-for-Intensifiers>

Abkürzungsverzeichnis

B-ITSF	Beginning of intensifier ,Anfang eines Intensivierers‘
CMC	Computer-mediated communication ,computervermittelte Kommunikation‘
CRF	Conditional Random Field
HMM	Hidden Markov Modell
I-ITSF	Inside of intensifier ,Innerhalb eines Intensivierers‘
L1	Lasso Regression
L2	Ridge Regression
LR	Logistische Regression
NB	Naive Bayes
O	Outside of intensifier ,Außerhalb eines Intensivierers‘
P	Precision ,Präzision‘
PoS	Part of Speech
R	Recall ,Wiedererkennung‘
SGD	Stochastic Gradient Descent ,stochastischer Gradientenabstieg‘
TafIn	Tagger für Intensivierer

Literaturverzeichnis

- Hans Altmann und Suzan Hahneemann. 2007. *Syntax fürs Examen*. Band 1 aus *Linguistik fürs Examen*. Vandenhoeck & Ruprecht, Göttingen, 3. aktualisierte Auflage.
- Susanne Augenstein. 1998. *Funktionen von Jugendsprache*. Band 192 aus *Germanistische Linguistik*. Max Niemeyer Verlag, Tübingen.
- Nils Bahlo, Tabea Becker, Zeynep Kalkvan-Aydın, Netaya Lotze, Konstanze Marx, Christian Schwarz, und Yazgül Şimşek. 2019. *Jugendsprache*. J.B. Metzler, Stuttgart.
- Rocío Baños. 2013. “That is so cool”: investigating the translation of adverbial intensifiers in English-Spanish dubbing through a parallel corpus of sitcoms. *Perspectives*, 21(4):526–542.
- Michael Beißwenger, Sabine Bartsch, Stefan Evert, und Kay-Michael Würzner. 2016. *EmpiriST 2015: A shared task on the automatic linguistic annotation of computer mediated communication and web corpora*. In *Proceedings of the 10th web as corpus workshop*, Seiten 44–56.
- Manfred Bierwisch. 1987. *Grammatische und konzeptuelle Aspekte von Dimensionsadjektiven*. Band 26 aus *Studia grammatica*. Akademie-Verlag, Berlin.
- Wolfgang Boettcher. 2009. *Grammatik verstehen*. Band 1 von 3. Max Niemeyer Verlag, Tübingen.
- Eva Breindl. 2009. *Intensitätspartikel*. In L. Hoffmann, Ed., *Handbuch der deutschen Wortarten*. Walter de Gruyter, Berlin/New York, Seiten 397–422.
- Eva Breindl. 2018a. *Fokuspartikel*. In *Leibniz-Institut für deutsche Sprache: "Systematische Grammatik"*. Grammatisches Informationssystem grammis.
- Eva Breindl. 2018b. *Partikel*. In *Leibniz-Institut für deutsche Sprache: "Systematische Grammatik"*. Grammatisches Informationssystem grammis.
- Jorge Carrillo-de-Albornoz und Laura Plaza. 2013. *An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification*. *Journal of the American Society for Information Science and Technology*, 64(8):1618–1633.

- Akshay Chavan. 2019. [Introduction to Conditional Random Fields \(CRFs\)](#). *AI Time Journal*.
- EmpiriST. 2015. [GSCL Shared Task: Automatic linguistic annotation of computer-mediated communication / social media](#). *EmpiriST 2015*.
- Bernhard Engelen. 1990. Sehr und Konsorten. *Zielsprache Deutsch Zeitschrift für Unterrichtsmethodik und angewandte Sprachwissenschaft*, 21.
- Robert Fuchs. 2017. [Do women \(still\) use more intensifiers than men? Recent change in the sociolinguistics of intensifiers in British English](#). *International Journal of Corpus Linguistics*, 22(3):345–374.
- Aurélien Géron. 2019. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Sebastopol, 2. Auflage.
- Matthias Granzow-Emden. 2019. [Deutsche Grammatik verstehen und unterrichten](#). Narr Francke Attempto Verlag, Tübingen, 3. überarbeitete und erweiterte Auflage.
- Gerhard Helbig und Joachim Buscha. 2007. *Deutsche Grammatik*. Langenscheidt, Berlin.
- Elke Hentschel und Harald Weydt. 2013. *Handbuch der deutschen Grammatik*. Walter de Gruyter, Berlin/Boston, 4., vollständig überarbeitete Auflage.
- Rika Ito und Sali Tagliamonte. 2003. [Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers](#). *Language in Society*, 32(2):257–279.
- Daniel Jurafsky und James H. Martin. 2022. [Speech and language processing](#). 3. Auflage.
- Ilja Kirschbaum. 2002. [Schrecklich nett und voll verrückt](#). Dissertation, Heinrich-Heine-Universität Düsseldorf, Düsseldorf.
- John D. Lafferty, Andrew McCallum, und Fernando C. N. Pereira. 2001. [Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data](#). *Proceedings of the Eighteenth International Conference on Machine Learning 2001*:282–289.
- Olga Laitenberger. 2016. [Die semantische Entwicklung von Intensivierern](#). Dissertation, Universität Göttingen, Göttingen.

- Charles Van Os. 1989. *Aspekte der Intensivierung im Deutschen*. Band 37 aus *Studien zur deutschen Grammatik*. Narr, Tübingen.
- Thomas Proisl. 2018. SoMeWeTa: [A part-of-speech tagger for German social media and web texts](#). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*:665–670.
- Luise F. Konstanz Pusch. 1981. Ganz. In *Partikeln und Deutschunterricht*, Seiten 31–44. Groos, Heidelberg.
- Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, und Jan Svartvik. 2000. *A comprehensive grammar of the English language*. Longman, Harlow, 16. Auflage.
- Fabian Renz-Gabriel. 2021. [Mega gut und sau schlecht](#). In Martin Evertz-Rittich und Frank Kirchhoff (Hgg.), *Geschriebene und gesprochene Sprache als Modalitäten eines Sprachsystems*, Band 575 aus *Linguistische Arbeiten*, Seiten 79–114. De Gruyter.
- Tatjana Scheffler, Michael Richter, und Roeland van Hout. Noch unveröffentlicht. *Tracing and classifying German intensifiers via information theory*.
- Tatjana Scheffler, Lesley-Ann Kern, und Hannah Seemann. Noch unveröffentlicht. *Individuelle linguistische Variabilität in sozialen Medien*. Bochum.
- Anne Schiller, Simone Teufel, und Christine Stöckert. 1999. [Guidelines für das Tagging deutscher Textcorpora mit STTS \(Kleines und großes Tagset\)](#). Tübingen.
- Fei Sha und Fernando Pereira. 2003. [Shallow parsing with conditional random fields](#). *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 1:134–141.
- Utpal Kumar Sikdar und Björn Gambäck. 2016. [Language identification in code-switched text using Conditional Random Fields and Babelnet](#). *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, Seiten 127–131.
- James M. Stratton. 2020. [Adjective Intensifiers in German](#). *Journal of Germanic Linguistics*, 32(2):183–215.

- James M. Stratton. 2022. [Tapping into German Adjective Variation: A Variationist Sociolinguistic Approach](#). *Journal of Germanic Linguistics*, 34(1):63–102.
- Charles Sutton und Andrew McCallum. 2007. [An introduction to Conditional Random Fields for relational learning](#). In *Introduction to Statistical Relational Learning*, Seiten 1–35. The MIT Press.
- Sali A. Tagliamonte. 2016. [So sick or so cool? The language of youth on the internet](#). *Language in Society*, 45(1):1–32.
- Sali A. Tagliamonte und Derek Denis. 2008. [Linguistic ruin? Lol! Instant messaging and teen language](#). *American Speech*, 83(1):3–34.

abartig	ganz	total
absolut	ganz und gar	überaus
ausgesprochen	genug	übertrieben
außergewöhnlich	geradezu	ultra
außerordentlich	gewaltig	unendlich
äußerst	gotteserbärmlich	unermesslich
bedeutend	herrlich	ungeheuer
bei weitem	höchst	ungemein
beileibe	immens	ungewöhnlich
beinahe	irre	unglaublich
bemerkenswert	irrsinnig	unheimlich
besonders	kaum	verdammt
bestialisch	komplett	verhältnismäßig
beträchtlich	krass	viel
denkbar	lange	voll
durchaus	mäßig	vollauf
echt	mega	völlig
ein bisschen	nahezu	vollkommen
einigermmaßen	recht	weit
enorm	relativ	weitaus
entschieden	sau	wenig
entsetzlich	scheiße	wesentlich
etwas	schweine	ziemlich
extrem	so	zu
fast	super	zucker
furchtbar	tierisch	zutiefst

Tabelle 6.1. Liste ausgewählter Intensivierer für Baseline 2.

Baseline 1	Precision	Recall	F_1
0	1.00	0.99	0.99
B-ITSF	0.37	0.47	0.41
I-ITSF	0.00	0.00	0.00
macro avg	0.45	0.49	0.47

Tabelle 6.3. Darstellung der precision, recall und F_1 Werte der Baseline 1: Identifizierung mittels SoMeWeTa.

Baseline 2	Precision	Recall	F_1
0	1.00	0.98	0.99
B-ITSF	0.24	0.55	0.33
I-ITSF	0.00	0.00	0.00
macro avg	0.41	0.51	0.44

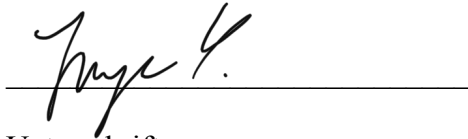
Tabelle 6.4. Darstellung der precision, recall und F_1 Werte der Baseline 2: Identifizierung mittels einer Wörterliste.

TafIn	Precision	Recall	F_1
0	1.00	1.00	1.00
B-ITSF	0.92	0.85	0.89
I-ITSF	0.60	0.46	0.52
macro avg	0.84	0.77	0.80

Tabelle 6.5. Darstellung der precision, recall und F_1 Werte von TafIn.

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die Arbeit selbständig angefertigt, außer den im Quellen- und Literaturverzeichnis sowie in den Anmerkungen genannten Hilfsmitteln keine weiteren benutzt und alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen sind, unter Angabe der Quellen als Entlehnung kenntlich gemacht habe.

A handwritten signature in black ink, appearing to read 'Mayer 4.', is written over a horizontal line.

Unterschrift