Universität Leipzig
Institut für Informatik
Studiengang Informatik, B.Sc.

Multi-Label Active Learning with Many Irrelevant Examples

Bachelorarbeit

Moritz Leander Brunsch

1. Gutachter: Jun.-Prof. Dr. Martin Potthast

Datum der Abgabe: 25. September 2023

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Leipzig, 25. September 2023

.....
Moritz Leander Brunsch

Zusammenfassung

Textklassifikation ist eines der wichtigsten Teilgebiete des maschinellen Lernens. Dabei ist der Annotationsaufwand gerade im Multi-Label Bereich sehr hoch. Active Learning versucht diesen zu reduzieren, indem es nur die informativsten Beispiele abfragt und annotieren lässt. Multi-Label Active Learning ist bisher jedoch deutlich weniger erforscht als der Single-Label Bereich. In diesem werden neue Querystrategien zudem vorwiegend auf relativ einfachen Datensätzen getestet. Währenddessen ist das Problem irrelevanter Daten, also solcher denen bei der Annotation kein Label zugeordnet werden kann, bisher im Multi-Label Bereich nicht erforscht worden. In dieser Arbeit habe ich dafür drei Textdatensätze mit vielen irrelevanten Beispielen gesucht und angepasst, um ein Testszenario für drei aktuell relevante Multi-Label Querystrategien zusammenzusetzen. Diese Arbeit zeigt, dass die Multi-Label Textklassifikation wie erwartet ein schweres Klassifikationsproblem darstellt. Weiterhin funktionieren auch die getesteten Querystrategien auf den Textdatensätzen nicht gut und haben Probleme, die irrelevanten Beispiele zu identifizieren. Daher bietet Active Learning in diesen Szenarien keinen Vorteil. Es scheint jedoch Faktoren wie Labelanzahl und -verteilung zu geben, die die Klassifikationsqualität stärker beeinflussen als der Anteil irrelevanter Daten.

Inhaltsverzeichnis

1 Einleitung **1**

2 Hintergrund **3**
2.1 Multi-Label Klassifikation 3
2.1.1 Klassenimbalance 3
2.1.2 Metriken 4
2.1.3 Transformer 5
2.2 Active Learning 5
2.2.1 Querystrategien 6

3 Active Learning mit irrelevanten Beispielen **9**
3.1 Irrelevante Beispiele 9
3.2 Active Learning mit irrelevanten Beispielen 11

4 Experiment **12**
4.1 Datensätze 12
4.2 Auswertung der Experimente 16
4.3 Evaluation der Querystrategien 17
4.3.1 Aufbau 17
4.3.2 Ergebnisse 18
4.4 Variation des irrelevanten Anteils 22
4.4.1 Aufbau 22
4.4.2 Ergebnisse 23

5 Diskussion der Experimente **27**
5.1 Evaluierung der Querystrategien 27
5.2 Variation des irrelevanten Anteils 29

6 Fazit **32**

Literaturverzeichnis **33**

Kapitel 1

Einleitung

Textklassifikation ist eines der wichtigsten Felder des modernen maschinellen Lernens. Textdaten sind dabei in großen Mengen im Internet erhältlich. Diese Daten sind jedoch im Fall von aus Onlineforen gescrapeten Daten in der Regel noch ungelabelt. Gerade das Labeln von Textdaten stellt aber einen großen Ressourcenaufwand dar. Zusätzlich gehören Texte oft nicht nur zu einer Kategorie, sondern zu mehreren. So könnte zum Beispiel ein Filmzitat welches sowohl Freude als auch Überraschung ausdrückt den Emotionskategorien "Freude" und "Überraschung" zugeordnet werden. Daher können solche Daten oft nicht nur mit einem, sondern mehreren Labeln beschrieben werden. Solche sogenannten Multi-Label Datensätze stellen einen noch größeren Annotationsaufwand dar, da nicht nur eine passende Kategorie für einen Datenpunkt gefunden werden muss, sondern für jede Kategorie beziehungsweise jedes Label geprüft wird ob dieses zum Beispiel passt.

Active Learning versucht diesen Aufwand zu verringern. Dazu wird vor Beginn der Klassifikation nur ein kleiner Teil der Daten von Expert:innen annotiert. Ausgehend von diesem Pool gelabelter Daten wird ein Klassifikator in mehreren Iterationen trainiert, während sogenannte Querystrategien immer neue, für das Active Learning System möglichst informative Beispiele auswählen und den Expert:innen zur Annotation übergeben. Dadurch sollen nur die informativsten Beispiele gelabelt werden und der Annotationsaufwand verringert sich deutlich.

Trotz der hohen Prävalenz von Multi-Label Datensätzen fokussiert sich ein großer Teil der Active Learning Forschung vor allem auf den Bereich der Single-Label Klassifikation [Ein-Dor et al., 2020, Margatina et al., 2021, Raj and Bach, 2022, Zhang et al., 2022]. Nur wenige neue Querystrategien für Multi-Label Datensätze [Reyes et al., 2018] werden erforscht. Dabei werden diese aber oft auf den gleichen Daten und kaum in schwierigeren Szenarien getestet.

Gerade das Klassenimbalanceproblem [Chawla et al., 2004], welches auftritt, wenn eine oder mehrere Klassen beziehungsweise Label deutlich öfter vertreten sind als andere, beeinträchtigt bekanntermaßen die Klassifikationsqualität maschinellen Lernens [Henning et al., 2023]. Dennoch wird es selten spezifisch thematisiert [Wertz et al., 2022]. Zusätzlich ist das an das Klassenimbalanceproblem angrenzende Problem irrelevanter Beispiele [Mazzoni et al., 2006] im Multi-Label Bereich noch überhaupt nicht erforscht. Irrelevante Beispiele sind hier Beispiele welchen keine Kategorie zugeordnet werden kann. Diese tauchen, weil sie zur eigentlichen Klassifikationsaufgabe keine Informationen beitragen, in den meisten gelabelten Datensätzen nicht auf. Bei Active Learning können solche Beispiele jedoch nicht ausgeschlossen werden und stellen daher ein potentielles Problem für die Klassifikation und insbesondere die Beispielauswahl der Querystrategien dar.

Da das Problem jedoch auch in den bei der Evaluation betrachteten Datensätzen nicht vorkommt, sind meines besten Wissens nach bisher keine Multi-Label Querystrategien auf solchen Daten getestet worden.

Daher versuche ich mit dieser Arbeit folgende Forschungsfragen zu beantworten:

1. Wie gut funktioniert Active Learning auf Multi-Label Datensätzen mit vielen irrelevanten Beispielen, gemessen am F1-Score?

2. Kann man durch das Beobachten verschiedener Metriken wie dem Anteil der irrelevanten Beispiele und der Labelimbalance konkrete Probleme in der Auswahl der Beispiele identifizieren?

3. Wie wirkt sich das Verhältnis von relevanten zu irrelevanten Beispielen (95:5, 80:20, 65:35, 50:50, 35:65, 20:80, 5:95) auf die Klassifikationsqualität aus?

Zur Beantwortung dieser Fragen entwerfe ich ein Experimentsetup für drei der relevantesten Multi-Label Querystrategien, welches diese auf neuen, von mir zusammengetragenen Datensätzen testet, die viele irrelevante Beispiele beinhalten. Um den Einfluss der irrelevanten Beispiele besser einzuschätzen und die dritte Frage zu beantworten, führe ich ein zweites Experiment durch, in dem der Anteil irrelevanter Beispiele in einem der Datensätze variiert wird. Auch hier evaluiere ich die Leistungen aller Querystrategien in diesen Szenarien.

Kapitel 2

Hintergrund

Um ein realistischeres Szenario herzustellen, ist es sinnvoll Textklassifikation in einem Multi-Label Setting zu betrachten. Ein Filmzitat, das mit den Emotionen "Freude" und "Überraschung" annotiert werden soll, lässt sich so abbilden. Das ist in einem Single-Label Szenario nicht einfach möglich.

2.1 Multi-Label Klassifikation

Anders als bei der Binären und Multi-Class Klassifikation wird einem Beispiel bei der Multi-Label Klassifikation nicht eine spezifische Klasse zugeordnet. Stattdessen gibt es eine Menge an sogenannten Labeln. Aus dieser wird für jedes Beispiel ausgewählt, welche dieser Label zu ihm passen. Ein Datenpunkt kann daher nicht nur durch *ein* Label sondern auch durch mehrere gleichzeitig oder keine beschrieben werden.

Formell bedeutet das, dass einem Beispiel x_i aus einer gesamten Labelmenge Y eine Labelmenge $Y_i \subseteq Y$ zugeordnet wird.

2.1.1 Klassenimbalance

Der negative Einfluss von Imbalance in der Klassenverteilung auf die Klassifikation im maschinellen Lernen ist ein viel und lang erforschtes Problem [Chawla et al., 2004]. Das Problem verschärft sich intuitiv bei Multi-Label Daten, da sich durch die hohe Anzahl an Labels und Labelkombination auch deutlich mehr Möglichkeiten für das Auftreten von Imbalance zwischen Labels ergeben [Charte et al., 2013]. Die meisten Ansätze zur Lösung des Problems beschäftigen sich hierbei mit Under- und Oversampling und erreichen damit gute Ergebnisse [Charte et al., 2013, Tahir et al., 2012].

2.1.2 Metriken

Der F_1-Wert wird zur Einschätzung der Klassifikationsqualität benutzt. Dieser bezeichnet das harmonische Mittel zwischen Precision und Recall. Wobei Precision den Anteil richtig positiv klassifizierter Beispiele an allen positiv klassifizierten Beispielen bezeichnet, und Recall den Anteil richtig positiv klassifizierter Beispiele an allen positiven Beispielen bezeichnet.

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{2tp}{2tp + fp + fn} \tag{2.1}$$

Der F_1-Wert liegt zwischen 0 und 1. Dabei bezeichnet 1 perfekte Precision und Recall, während 0 das schlechtmöglichste Ergebnis darstellt.

Die normalisierte Shannon-Entropie [Shannon, 1948] wird zur Einschätzung der Labelimbalance genutzt.

$$H_{norm} = -\frac{\sum_{y \in Y} p(y) ln(p(y))}{ln(|Y|)} \tag{2.2}$$

Wobei $p(y)$ den Anteil der Beispiele der betrachteten Daten bezeichnet, die dieses Label tragen. Die Shannon-Entropie wird durch den Logarithmus der Anzahl der Klassen geteilt um den Wert zu normalisieren und Werte zwischen 0 und 1 zu erzeugen. Eine hohe Entropie zeigt dabei eine gleichmäßige Labelverteilung an.

Labelkardinalität ($LCARD$) und Labeldichte ($LDENS$) [Tsoumakas and Katakis, 2007] sind weitere Metriken, die auch zur generellen Einschätzung der Datensätze dienen. Die Labelkardinalität berechnet sich aus dem Durchschnitt der Anzahl der Label Y pro Beispiel i im Datensatz.

$$LCARD = \frac{1}{n} \sum_{i=1}^{n} |Y_i| \tag{2.3}$$

Sie gibt einen generellen Überblick über die Anzahl an Labels, die ein Beispiel trägt. Die zweite Metrik berechnet sich aus der durchschnittlichen Anzahl an Labels geteilt durch die maximale Anzahl an Labeln.

$$LDENS = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i|}{|Y|} \tag{2.4}$$

Die Labeldichte gibt besonders im Vergleich weitere Einsichten in die Beschaffenheit des Labelraums.

2.1.3 Transformer

Zur Klassifikation der Daten stimme ich BERT [Devlin et al., 2018] fein ab. Das transformerbasierte [Vaswani et al., 2017] Klassifikationsmodell wurde auf großen, englischsprachigen Textkorpora vortrainiert. Es besitzt in der Basiskonfiguration 12 Schichten und damit insgesamt 110 Millionen Parameter. Ich benutze hier die "uncased" Version, die die Groß- und Kleinschreibung nicht beachtet.

2.2 Active Learning

Textklassifikation ist ein Feld des maschinellen Lernens, welches große Datenmengen benötigt. Diese sind zwar gerade in Form von gescrapeten Texten aus Onlineforen in riesiger Anzahl verfügbar, jedoch ist der Großteil dieser Daten nicht annotiert. Gerade Multi-Label Daten stellen einen großen Aufwand in der Annotation dar, da die erhöhte Anzahl an Labeln im Gegensatz zum Single-Label Setting auch die Annotationsaufgabe deutlich erschweren. Dadurch bringt die Annotation von Multi-Label Textdaten einen sehr großen Aufwand mit sich.

Ein Ansatz, um diesen Aufwand an Zeit und Ressourcen zu verringern, ist Active Learning. Active Learning macht durch eine Querystrategie, welche selbstständig Beispiele auswählt, die annotiert und klassifiziert werden sollen, Supervised Learning möglich, wenn nur wenige gelabelte Beispiele vorhanden sind. Somit soll der Annotationsaufwand verringert werden, da nicht alle Beispiele im Vorhinein ausgewertet werden müssen.

Im Active Learning gibt es dabei mehrere Ansätze, hier werde ich mich auf den Poolbasierten Ansatz [Lewis and Gale, 1994] fokussieren, bei dem ein großer Pool nicht annotierter Daten S_U vorhanden ist und dieser immer wieder durch die Querystrategie abgefragt wird, um neue Beispiele zu selektieren, die einem kleineren Pool annotierter Beispiele S_L hinzugefügt werden.

Bei diesem Ansatz wird anfangs zumeist zufällig eine Anzahl nicht annotierter Beispiele ausgewählt, um den Klassifikator initialisieren zu können. Diese werden durch Expert:innen gelabelt und dem Pool annotierter Daten hinzugefügt. Aus diesem Pool wird dann zum ersten Mal das Klassifikationmodell trainiert. Anschließend nutzt eine Querystrategie eine Auswahl aus Variablen wie zum Beispiel den Vorhersagen des Klassifikators für den ungelabelten Pool [Li and Guo, 2013] und der Struktur des Labelraums [Reyes et al., 2018], um nach festgelegten Kriterien eine bestimmte Anzahl an Datenpunkten auszuwählen. Die selektierten Beispiele werden wiederum von Expert:innen annotiert und ebenfalls dem Pool annotierter Beispiele hinzugefügt. Hierauf wird das Klassifikationsmodell nun erneut trainiert, um neue Vorhersagen und Werte für die

nächste Active Learning Iteration zu erzeugen.

Diese Schritte werden so lange wiederholt, bis bestimmte Abbruchkriterien erreicht sind und das Active Learning System stoppt. Das einfachste dieser Kriterien, welches auch in dieser Arbeit genutzt wird, ist das Erreichen einer bestimmten Anzahl an Iterationen. Vlachos [2008] und Laws and Schütze [2008] schlagen jedoch auch komplexere Kriterien vor, welche sich zum Beispiel an der Klassifikationsqualität und der Beispielauswahl orientieren.

2.2.1 Querystrategien

Querystrategien sind für die Auswahl der Beispiele während der Active Learning Iterationen zuständig. Durch sie wird eine Bewertung c berechnet, die genutzt wird um ein Ranking der nicht annotierten Beispiele aufzustellen. Die meisten Querystrategien werten für dieses Ranking die informativsten Beispiele am höchsten aus [Reyes and Ventura, 2018]. Aus diesem Ranking wird dann eine bestimmte Anzahl noch nicht gelabelter Beispiele ausgewählt.

Hierbei gibt es zwei Strategien. Die sogenannten myopischen Querystrategien [Li and Guo, 2013, Reyes et al., 2018] sind dafür optimiert, jeweils einzelne Beispiele auszuwählen. Dies ist gerade bei großen Datenmengen und Transformermodellen ineffizient. Letztere benötigen sehr viel Zeit, um neu trainiert zu werden und profitieren somit davon, dass weniger Active Learning Iterationen mit höheren Querygrößen durchgeführt werden. Deshalb werden myopische Strategien in vielen Experimenten so angepasst, dass sie greedy die n besten Beispiele ihres Rankings auswählen. Durch diese Anpassung läuft man jedoch Gefahr, dass die ausgewählten Beispiele untereinander starke Redundanz zeigen.

Sogenannte Batch-Mode Querystrategien sind wiederum für die Auswahl von mehreren Beispielen auf einmal optimiert. Sie wenden weitere Kriterien an, um Diversität unter den ausgewählten Beispielen zu garantieren und somit eine größere Breite an Informationen zu bieten [Chakraborty et al., 2011, Reyes and Ventura, 2018, Zhang et al., 2014]. Diese Strategien benötigen jedoch für ihre Berechnungen deutlich mehr Zeit und Ressourcen [Reyes and Ventura, 2018], da im Gegensatz zu myopischen Strategien ganze Mengen von Beispielen ausgewertet werden müssen und nicht nur einzelne Beispiele.

Aufgrund dieser erhöhten Zeit- und Ressourcenintensität werden für das Experiment in dieser Arbeit nur myopische Strategien genutzt, die jedoch mehrere Beispiele auf einmal auswählen.

Welche Kriterien zur Erstellung des Rankings genutzt werden, und wie diese gewichtet werden, ist stark von der Auswahl der Querystrategie abhängig. Für die in dieser Arbeit genutzten werden diese hier nun kurz vorgestellt.

Label Cardinality Inconsistency (LCI) ist eine von Li and Guo [2013] vorgeschlagene Querystrategie, die die Labelkardinalität der Beispiele nutzt um deren Informationsgehalt einzuschätzen. Dafür wird die prognostizierte Labelanzahl der individuellen, ungelabelten Beispiele mit der Labelkardinalität des Pools gelabelter Beispiele verglichen.

$$c(x_i) = \left\| |Y_i| - \frac{1}{|S_L|} \sum_{j=1}^{|S_L|} |Y_j| \right\|_2 \tag{2.5}$$

Um den Vergleichswert c bei einem Beispiel x_i zu bestimmen, wird dafür die euklidische Distanz der beiden Werte berechnet, wobei j hier alle Beispiele aus S_L bezeichnet.

Je höher die Distanz der beiden Werte ist, desto unterschiedlicher und somit informativer wird das jeweilige Beispiel gewertet.

Adaptive Active Learning (AAL) wurde ebenfalls von Li and Guo [2013] vorgestellt und verbindet LCI mit Uncertainty Sampling. Uncertainty Sampling ist eine der frühesten und einfachsten Gruppen von Querystrategien im Active Learning und wurde bereits viel erforscht.

$$c_u(x_i) = \frac{1}{max_{l \in Y}(|p_{il} - (1 - p_{il})|)} \tag{2.6}$$

Die hier genutzte Version des Uncertainty Sampling definiert sich, indem für jedes Beispiel x_i das Label, welches die größte Spanne zwischen der positiven und negativen Vorhersage des Klassifikators besitzt, herausgesucht wird. Das Inverse dieser Spanne wird als Maß c_U für das Ranking der Beispiele genutzt. Ein höherer Wert steht hierbei für eine geringere Spanne bei der Labelvorhersage, da der innere Teil der Gleichung 2.6 immer zwischen null und eins liegt. Kleinere Spannen beziehungsweise unsicherere Vorhersagen führen also zu höheren Bewertungen durch das Uncertainty Sampling.

Die beiden Teile der Querystrategie werden nach der Berechnung aller Werte zu gleichen Teilen verknüpft und gleichwertig betrachtet. Somit wird bei jedem Beispiel sowohl die Distanz der Labelkardinalität als auch die generelle Unsicherheit in der Labelvorhersage des Beispiels gewertet.

Category Vector Inconsistency and Ranking of Score (CVIRS) verbindet ebenfalls zwei Metriken und wurde von Reyes et al. [2018] aufgestellt.

Die erste Metrik wird als Ranglistenproblem beschrieben. Hier wird für jedes ungelabelte Beispiel x_i für jedes zugehörige Label l die Unsicherheit der Vorhersage m_{il} bestimmt.

$$m_{il} = (|p_{il} - (1 - p_{il})|) \tag{2.7}$$

Hier steht p_{il} für die vom Klassifikator bestimmte Wahrscheinlichkeit, dass ein Beispiel x_i dem Label l zugeordnet wird. Diese Unsicherheitsspannen m werden nun für jedes Label unabhängig in Ranglisten τ_l eingeordnet, welche um das erste Maß zu berechnen, mit Bordas Methode [Dwork et al., 2001] als $s(i)$ zusammengefasst werden. Beispiele mit kleinen Unsicherheitsmargen m werden dabei höher eingeordnet als Datenpunkte deren Marge größer ist.

Der zweite Teil der Bewertung beschäftigt sich mit der Beschaffenheit des Labelraums. Die Methode vergleicht dafür die vorhergesagten Label der ungelabelten Beispiele mit den bekannten echten Labeln der gelabelten Beispiele. Dabei berechnet sie die Distanz der Labelvektoren und rankt die Beispiele dementsprechend. Das basiert auf der Annahme, dass der Labelraum der ungelabelten Beispiele dem der gelabelten Beispiele ähnlich ist. Daher werden die Beispiele, die die größte Distanz zum bisher bekannten Labelraum besitzen durch das zweite Maß $v(i)$ am höchsten eingeordnet, da sie vermeintlich die meisten Informationen beinhalten, da ihre Labelstruktur sich stark von der bekannten unterscheidet.

$$c(x_i) = s(i) \cdot v(i) \tag{2.8}$$

Zuletzt werden die beiden Metriken multiplikativ verbunden. Dies ist möglich, da beide Maße die als am informativsten eingeschätzten Beispiele mit den höchsten Werten belegen. Somit sind auch die von der Querystrategie ausgewählten Beispiele jene, welche im Wert c die höchsten Werte zugeordnet bekommen.

Kapitel 3

Active Learning mit irrelevanten Beispielen

Aktuelle Forschung zu Textklassifikation im Multi-Label Active Learning zeigt, dass es oft kaum besser als Random Sampling funktioniert [Wang and Liu, 2023, Wertz et al., 2023]. In extrem imbalancierten Szenarien werden viele Strategien laut Wertz et al. [2022] sogar von Random Sampling outperformt. Gleichzeitig konzentriert sich die Erforschung neuer Querystrategien vor allem auf den selben Pool von Multi-Label Datensätzen [Li and Guo, 2013, Reyes et al., 2018, Wang and Liu, 2023]. Diese bestimmten Daten sind somit schon sehr gut einschätzbar und bringen eine gute Basis um neue Strategien zu vergleichen. Sie bieten jedoch kaum neue Einsichten in die Funktionalität von Active Learning Systemen in der echten Welt.

3.1 Irrelevante Beispiele

Ein Problem, welches im Multi-Label Active Learning Bereich gar nicht erforscht wurde, sind irrelevante Beispiele. Ein Beispiel wird dabei als irrelevant bezeichnet, wenn ihm bei der Annotation keine Label zugeordnet wurden. Es ist sehr wahrscheinlich, dass diese Beispiele in Daten aus der realen Welt häufig auftreten. Beispielsweise gibt es bei der Annotation von Filmzitaten mit Emotionen auch viele Beispiele, die keine Emotionen ausdrücken. Diesen kann dann kein Emotionslabel zugeordnet werden. Das Problem tritt des Weiteren noch stärker auf, wenn die Arten und Ausdrucksweisen von Hate Speech in Internetforen wie gab.com untersucht werden sollen. Hier zeigt ein großer Teil aller Beiträge voraussichtlich keine Vorkommnisse von expliziter Hate Speech. Diese Beiträge sind dann im Kontext der Klassifikationsaufgabe irrelevant.

Irrelevante Beispiele sind eine Erweiterung des Imbalanceproblems, da sie die Labelverteilung der Beispiele verschieben und so im Verhältnis zur Gesamt-

zahl der Beispiele für jedes Label weniger Daten vorhanden sind, auf denen ein Klassifikator trainiert werden kann.

Irrelevante Beispiele bringen somit keine wirkliche Informationen über die Unterscheidung der Label und somit das eigentliche Klassifikationsproblem mit sich und scheinen der Qualität der Klassifikation dadurch intuitiv sogar schädlich zu sein. Daher wird dieses Problem in Szenarien ohne Active Learning umgangen, indem solche Beispiele direkt bei der Erstellung der Trainingsdatensätze gefiltert werden. Das ist in einem Active Learning Szenario jedoch nicht möglich, da die echten Label der Daten im Voraus nicht bekannt sind. Der vorherige große Annotationsschritt wird im Active Learning explizit umgangen um Zeit und Ressourcen zu sparen. Stattdessen werden die Daten innerhalb der Active Learning Iterationen annotiert. Zu diesem Zeitpunkt kann realistisch nicht mehr nach irrelevanten Beispielen gefiltert werden, weil die Querystrategien ohne das Wissen um die echten Label der Beispiele Daten auswählen und somit nicht einfach nach relevanten Beispielen gefiltert werden kann.

Auch eine Filterung nach dem Annotationsprozess erscheint nicht wirklich sinnvoll, da die Beispielwahl durch die Querystrategien zu diesem Zeitpunkt schon abgeschlossen ist. Ein mögliches Verfahren in diesem Szenario wäre, dass weniger Beispiele als geplant an den Klassifikator weitergegeben werden. Das könnte im Extremfall bedeuten, dass, wenn keins der ausgewählten Beispiele Label trägt, dem Klassifikator auch keine Beispiele übergeben würden und er erneut auf den gleichen Beispielen trainiert werden müsste. Diese Möglichkeit ist nicht erstrebenswert und das Vorgehen daher nicht sinnvoll einzusetzen.

Ein zweites Vorgehen ist, vor dem nächsten Trainingsschritt so lange weitere Beispiele abzufragen, bis eine bestimmte Mindestanzahl relevanter Beispiele ausgewählt wurde. Hierfür müssen jedoch potentiell alle Rankings mehrmals unter Einbezug der jeweils neu gelabelten Beispiele neu berechnet werden. Dies bringt gerade bei ressourcenintensiven Strategien wie CVIRS unter Umständen einen großen Rechenaufwand mit sich. Außerdem müssten die neuen Daten alle annotiert werden, was dem Ziel von Active Learning, so wenige Daten wie möglich zu annotieren, entgegenwirkt.

Aus diesen Gründen ist eine Filterung irrelevanter Beispiele in einem Active Learning Szenario nicht sinnvoll umzusetzen. Laut Mazzoni et al. [2006] schaden diese Daten in Active Learning Szenarien jedoch, da Random Sampling in mehreren Multi-Class Szenarien durch ihr Vorhandensein nicht mehr übertroffen wurde.

Irrelevante Beispiele sind deshalb ein relevantes Problem für die Klassifikation in Active Learning Szenarien. Sie erhöhen voraussichtlich die Schwierigkeit der Klassifikation deutlich, da sie wie beim Klassenimbalanceproblem die Labelhäufigkeit stark verringern können. Anders als bei sonstiger Imbalan-

ce geschieht das jedoch für alle Klassen gleichmäßig. Außerdem unterscheidet sich das Problem insofern vom herkömmlichen Klassenimbalanceproblem im Multi-Label Setting, als dass es nicht einige Label gibt, für die deutlich mehr Positivbeispiele vorhanden sind als für andere, sondern dass es für alle Label mehr Negativbeispiele gibt. So produziert dieses Problem voraussichtlich kein Bias, welches einem Label zugute kommt, sondern vor allem einen Bias gegen alle Label. Dies könnte jedoch Label, die sowieso wenig vorhanden sind, stärker treffen als häufige Label. Deren Beispiele könnten dann durch Querystrategien noch leichter übersehen werden als sonst, da sie von einer Fülle an irrelevanten Beispielen überschattet werden. Die Hinzufügung dieses Problems stellt daher voraussichtlich eine deutlich Erschwerung der herkömmlichen Multi-Label Active Learning Szenarien dar.

3.2 Active Learning mit irrelevanten Beispielen

Trotz dieser Probleme wurde der Einfluss irrelevanter Beispiele auf Active Learning bisher kaum erforscht und es ist mir keine Multi-Label Querystrategie bekannt, die auf solchen Datensätzen getestet wurde. Mazzoni et al. [2006] ist meines besten Wissens nach die einzige Arbeit, die sich mit irrelevanten Beispielen im Active Learning beschäftigt. Darin werden allerdings nur Multi-Class und keine Multi-Label Szenarien betrachtet.

Es existiert zudem keine Benchmark, mit der Multi-Label Querystrategien in Szenarien mit vielen irrelevanten Beispielen verglichen werden können.

An dieser Stelle habe ich angesetzt, indem ich Datensätze mit vielen irrelevanten Beispielen gesammelt und daraus ein Experimentsetup für die Querystrategien LCI, AAL und CVIRS aufgebaut habe. Da solche Datensätze in annehmbarer Qualität und Größe aus den oben genannten Gründen sehr rar sind, musste ich hier auch andere Datensätze anpassen, um diese den Anforderungen gerecht werden zu lassen.

Des Weiteren ist der konkrete Einfluss verschiedener Anteile irrelevanter Beispiele auf die Multi-Label Klassifikation kaum bekannt. Dazu habe ich in einem zweiten Experiment das Verhältnis irrelevanter und relevanter Daten auf einem einzelnen Datensatz variiert.

Kapitel 4

Experiment

Für das Experiment wurden mehrere große Datensätze mit vielen irrelevanten Beispielen gesucht und ausgewählt. Aus diesen wurde mit Hilfe der small-text Bibliothek [Schröder et al., 2021] in Python ein Active Learning Experimentsetup für die Querystrategien LCI, AAL und CVIRS erstellt, um deren Effektivität auf Datensätzen mit vielen irrelevanten Beispielen zu evaluieren. In einem zweiten Active Learning Experiment wurde der Einfluss des Anteils irrelevanter Beispiele auf die Klassifikation getestet, indem dieser auf einem Datensatz variiert wurde.

4.1 Datensätze

Da bei maschinellem Lernen, das kein Active Learning nutzt, die Klassen- beziehungsweise Labelverteilung a priori bekannt ist, kann diese in den meisten Fällen so angepasst werden, dass sie die Klassifikation nicht beeinträchtigt. Dabei können auch irrelevante Beispiele entfernt werden. Daher ist der Großteil der Multi-Label Datensätze und somit auch der herkömmlich zur Evaluation von Querystrategien genutzten Multi-Label Datensätze auf diese Weise optimiert. Solche herkömmlich gebrauchten Datensätze sind somit für das Experiment nicht anwendbar und es müssen andere Daten genutzt werden.

Für die Auswahl passender Datensätze gelten zwei Kriterien, um sie von den weiter verbreiteten Daten abzugrenzen.

Erstens ist es wichtig, dass sie eine ausreichende Anzahl an Beispielen beinhalten, damit umfangreiche Auswahlmöglichkeiten für die Querystrategien gegeben sind. Hier wurde eine ungefähre Abgrenzung von mindestens 20.000 Beispielen pro Datensatz gewählt.

Zweitens soll der Anteil irrelevanter Beispiele erheblich sein. Dafür wurde die Grenze bei etwa einem Drittel irrelevanter Beispiele gesetzt. Datensätze,

die einen geringeres Maß an Irrelevanz zeigen, wurden für das Experiment nicht betrachtet.

Da die Auswahl geeigneter Datensätze schon durch diese Einschränkungen sehr gering war, wurden keine weiteren Anforderungen an die Labelanzahl und -verteilung gestellt. Generell ist zu erwarten, dass die Klassifikationsqualität mit steigender Labelanzahl sinkt, da die Komplexität des Klassifikationsproblems steigt.

4.1.1 Gab Hate Korpus

Das Gab Hate Korpus [Kennedy et al., 2022] ist ein Datensatz aus Posts des amerikanischen Kurznachrichtendiensts gab.com. Der Datensatz beinhaltet 27.665 Dokumente, welche mit diversen hierarchischen Labeln zum Thema Hate Speech und Zielgruppen dieser annotiert sind. Aus dieser Labelmenge wurden jedoch, wegen fehlender Übereinstimmung zwischen den annotierenden Personen bei den anderen Labeln, nur die drei top-level Label "assaults on human dignity" (HD), "call for violence" (CV) und "violent/offensive" (VO) ausgewählt. [Kennedy et al., 2022]

Nur etwa zwölf Prozent aller Dokumente sind mit mindestens einem der drei Label versehen. Daher können knapp 88% der Daten als irrelevant bezeichnet werden.

Der Datensatz hat also eine ausreichende Anzahl an Daten und viele irrelevante Dokumente. Er ist daher für das Experiment geeignet, wobei er durch die geringe Labelanzahl eine relativ einfache Klassifikationsaufgabe darstellt. Dies wird vermutlich durch den hohen Anteil irrelevanter Beispiele erschwert.

4.1.2 Subtitle Sentiments

Subtitle Sentiments (ursprünglich xed_en_fi) ist ein von Öhman et al. [2020] veröffentlichter Datensatz bestehend aus mit Emotionen annotierten Filmzitaten, welche aus dem OPUS OpenSubtitle Korpus entnommen wurden.

Diese sind zwischen etwa vier Zeichen und drei Sätzen lang. Sie sind nach Plutchiks Theorie der Emotionen [Plutchik, 1980] mit den acht Labeln "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise" und "trust" versehen. Zusätzlich wurde das Label "neutral" hinzugefügt, um Elemente abzudecken, die keine der Emotionen zeigen. Alle Elemente wurden ohne Kontextwissen über die Herkunft der Zitate ausgewertet. Daher sind sie für eine maschinelle Auswertung, die den Kontext der korrespondierenden Filme nicht einbezieht, geeignet.

Neutrale Elemente werden hier als irrelevant betrachtet, da sie keine der Emotionen beinhalten. Das Label wurde somit für das Experiment entfernt.

Insgesamt wurden 9.675 Elemente als neutral bewertet, während 17.528 mit mindestens einer Emotion annotiert wurden. Somit sind etwa 35,5% des Datensatzes als irrelevant einzuschätzen.

Mit einer ähnlichen Anzahl an Datenpunkten wie das Gab Hate Korpus, jedoch unterschiedlicher Labelverteilung und einem deutlich niedrigeren Prozentsatz irrelevanter Daten, eignet sich dieser Datensatz gut für das Experiment.

4.1.3 ArXiv Paper Abstracts

Der arXiv Paper Abstracts Datensatz [Paul and Rakshit, 2021] wurde auf besteht aus einer Sammlung der Titel und Abstracts von Papern die auf der Plattform arXiv veröffentlicht wurden. Diese sind mit den jeweiligen arXiv Kategorien annotiert.

Der Datensatz enthält 41.125 Dokumente, die mit mindestens einem von 147 arXiv Kategorielabeln annotiert sind. Ursprünglich enthielt der Datensatz zusätzliche Label welche nicht in der offiziellen Labeltaxonomie[1] vorhanden waren. Diese wurden jedoch vor der Nutzung entfernt, um eine möglichst einheitliche Annotationsqualität zu erreichen. Das häufigste dieser verbliebenen Label ist *"cs.CV"*. Dieses Label ist für etwa ein Drittel der Daten das einzige Label, das diesen zugeordnet ist. Während der ursprüngliche Datensatz keine der Definition von irrelevanten Daten entsprechenden Beispiele enthält, lässt sich durch das generelle Entfernen dieses Labels von allen Daten ein solcher, den Kriterien entsprechender Datensatz herstellen. Dessen Anteil irrelevanter Daten beträgt dann etwa 34%.

Die Labelverteilung des resultierenden Datensatzes ist weiterhin sehr unbalanciert. Einige Label sind mehr als 1.000 Mal vorhanden, während viele nur ein Mal vorkommen. Das häufigste Label kommt 23.303 Mal vor. Dies erschwert die Klassifikationsaufgabe stark.

Des Weiteren sind die Beispiele des Datensatzes deutlich länger als die der zwei Vorherigen. Es ergeben sich Satzlängen von bis zu 565 Token. Diese Länge überschreitet die maximale Menge von Token die BERT für ein Beispiel aufnehmen kann. Da dies jedoch weniger als 0,1% Prozent der Daten betrifft, kann diese Inkompatibilität vernachlässigt werden. Die betroffenen Daten werden durch BERT automatisch bei der maximalen Länge von 512 abgeschnitten.

Durch diese Anpassungen und Betrachtungen lässt sich sagen, dass der Datensatz mit einem Anteil irrelevanter Daten von 34% und einer Gesamtzahl über 40.000 Beispielen für das Experiment geeignet ist. Er bietet auch durch die hohe Anzahl an Labeln und deren hohe Imbalance eine deutlich schwerere Klassifikationsaufgabe als die vorherigen zwei Datensätze.

[1] https://arxiv.org/category_taxonomy

Name	Beispiele	Label	Irrel.	LCARD	LDENS
Gab Hate Korpus	27.665	3	0,88	0,1543	0,0514
Subtitle Sentiments	27.203	8	0,36	0,8242	0,103
arXiv Paper Abstracts	41.125	146	0,34	1,4345	0,0098

Tabelle 4.1: Vergleich der in den Experimenten genutzten Datensätze. Anzahl der Datenpunkte (Beispiele), Anzahl der Label (Label), Anteil irrelevanter Daten (Irrel.), Labelkardinalität (LCARD), Labeldichte (LDENS).

4.1.4 Analyse und Vergleich der Datensätze

Die drei Datensätze sind, unterscheiden sich trotz der Auswahlkriterien, die sie alle erfüllen, in anderen Metriken stark voneinander. So besitzt das Gab Hate Korpus mit fast 90% den größten Anteil an irrelevanten Beispielen. Die anderen beiden Datensätze kommen nur auf etwa ein Drittel.

Die Labelkardinalität ist bei den Datensätzen mit höherer Anzahl an Labeln ebenso höher als bei den Datensätzen mit weniger Labeln. Das lässt sich unter anderem dadurch erklären, dass der Subtitle Sentiments Datensatz und der arXiv Paper Abstracts Datensatz deutlich weniger irrelevante Beispiele besitzen. Dadurch sind in diesen Datensammlungen mehr Datenpunkte mit mindestens einem Label annotiert. Dies steigert die Labelkardinalität deutlich.

Außerdem können Beispiele in den beiden letzten Datensätzen mit mehr als drei Labeln annotiert werden. Gleichzeitig ist das jedoch die höchste Anzahl von Labeln, die ein Beispiel aus dem Gab Hate Korpus tragen kann. Auch das könnte einen Grund für den Unterschied der Labelkardinalitäten darstellen.

Ein anderes Bild zeigt sich bei der Betrachtung der Labeldichte. Während der arXiv Paper Abstracts Datensatz etwa die zehnfache Labelkardinalität des Gab Hate Korpus aufzeigt, fällt die Labeldichte im größeren Datensatz bei nur einem Fünftel der Labeldichte des Gab Hate Korpus deutlich geringer aus. Im arXiv Datensatz sind Datenpunkte gemessen an der maximalen Anzahl der Label mit deutlich weniger Labeln annotiert. Daraus lässt sich schließen, dass hier mehr Beispiele als im Gab Datensatz ausgewertet werden müssen, um überhaupt für jedes Label ein Beispiel zu finden. Es ist also insgesamt schwieriger, Informationen über Label zu sammeln. Dadurch erschwert sich die Klassifikationsaufgabe des arXiv Datensatzes gegenüber den anderen Datensätzen.

Der Subtitle Sentiments Datensatz hat währenddessen die größte Labeldichte. Dies erwächst hier im Vergleich mit dem Gab Hate Korpus vermutlich ausschließlich aus dem größeren Anteil an relevanten Beispielen. Bei einer Hochrechnung der Labeldichte des Gab Hate Korpus, wenn dieser nur 33%

irrelevanter Daten besäße, ergibt sich eine deutlich höhere Labeldichte als die des Subtitle Sentiments Datensatzes.

Abschließend lässt sich hier also sagen, dass das Gab Hate Korpus abgesehen von seinem hohen Anteil an irrelevanten Daten eine relativ einfache Klassifikationsaufgabe repräsentiert. Die irrelevanten Daten könnten diese jedoch deutlich erschweren.

Die voraussichtlich mit Abstand komplexeste und schwerste Klassifikationsaufgabe stellt hier der arXiv Paper Abstracts Datensatz dar. Dies liegt vor allem an der großen Imbalance der Label, der geringen Labeldichte und der generell hohen Anzahl an Labeln. Hier sind die schlechtesten Ergebnisse zu erwarten.

Der Subtitle Sentiments Datensatz lässt sich mit seiner Komplexität zwischen den anderen Datensätzen einordnen. Durch die höhere Labelanzahl als die des Gab Hate Korpus und die im Vergleich mit diesem niedrige Labeldichte scheint er deutlich schwerer als dieser zu klassifizieren zu sein. Er stellt aber gleichzeitig durch die höhere Labelkardinalität und -dichte, sowie relativ balancierte Klassenverteilung und niedrigere Labelanzahl eine deutlich leichtere Klassifikationsaufgabe dar als der arXiv Paper Abstracts Datensatz. Aus diesen Gründen wird er als ausreichend schwere aber nicht zu komplexe Klassifikationsaufgabe für die Variation des irrelevanten Anteils ausgewählt.

4.2 Auswertung der Experimente

Die Berechnung der F_1-Scores unterscheidet sich bei der Multi-Label Klassifikation von der herkömmlichen Methode, da nicht nur ein sondern mehrere Label betrachtet werden. Hier gibt es zwei einfache Methoden um den Wert anzupassen.

Die erste Möglichkeit besteht darin, global alle richtig positiven, richtig negativen und falsch negativen Klassifikationen zu sammeln, um aus diesen direkt den F_1-Score zu berechnen. Diese Methode nennt sich Micro-F_1.

Die zweite Möglichkeit ist der Macro-F_1 und besteht darin den F_1-Score separat für jedes Label zu bestimmen und dann deren Durchschnitt zu berechnen. Dies legt ein größeres Gewicht auf die richtige Klassifikation der einzelnen Label und favorisiert im Vergleich mit dem Micro-F_1 die Label, die im Datensatz weniger oft vorhanden sind.

Da in diesen Datensätzen durch die irrelevanten Beispiele eine generelle Imbalance herrscht und es dadurch generell mehr negative Beispiele für jedes Label gibt, ist ein Bias gegen alle Label zu erwarten. Dadurch sind global viele richtig negative Einschätzungen zu erwarten, wodurch der Micro-F_1-Score favorisiert und voraussichtlich ein besseres Ergebnis suggerieren würde.

Ein realistischeres Bild zeigt in diesem Fall der Macro-F_1-Score, da durch ihn keine der Klassen unterrepräsentiert werden und der Einfluss der irrelevanten Beispiele verringert wird.

Daher wird hier für die Auswertung der Active Learning Simulationen der Macro-F_1-Score genutzt.

Die Shannon-Entropie nutzt für die Berechnung ausschließlich die positiv ausgewerteten Label und lässt somit irrelevante Beispiele außen vor. Sie ist nach der Normalisierung daher gut für die Einschätzung der Klassenimbalance unabhängig der irrelevanten Beispiele geeignet.

Gemeinsam mit dem Anteil der irrelevanten Beispiele und der Labelkardinalität wird sie zur Auswertung der abgefragten Beispiele pro Query verwendet.

4.3 Evaluation der Querystrategien

Das erste Experiment dient zur Einschätzung der generellen Leistung von Active Learning Algorithmen in schwereren Szenarien als den bisher geteteten [Li and Guo, 2013, Reyes et al., 2018]. Außerdem soll ein erster Einblick in den Einfluss irrelevanter Beispiele auf die Klassifikation und die Beispielauswahl der Querystrategien gewährt werden. Dafür werden die ausgewählten Strategien auf den vorgestellten Daten ausgeführt und getestet.

4.3.1 Aufbau

Im Laufe des ersten Active Learning Experiments werden die Datensätze Gab Hate Korpus, Subtitle Sentiments und arXiv Paper Abstracts mit der Random Baseline und den Querystrategien LCI, AAL und CVIRS kombiniert. Daraus ergeben sich insgesamt neun Kombinationen der Variablen im Experiment.

Für jede Kombination werden insgesamt 20 Active Learning Iterationen ausgeführt. In jedem der 20 Durchläufe wird zuerst eine Query ausgeführt, in der die jeweilige Querystrategie anhand der vorherigen Klassifikation jeweils 50 Beispiele auswählt, die von ihr als wichtigste oder informativste Beispiele bewertet werden. An dieser Stelle werden dem Active Learning Modell die wahren Label der abgefragten Datenpunkte aus dem Datensatz offenbart, um den Annotationsschritt durch Expert:innen zu simulieren. Diese nun gelabelten Beispiele werden dann dem existierenden Pool solcher hinzugefügt. Auf diesem Pool wird dann das Klassifikationsmodell BERT Base Uncased trainiert. Dessen Ausgaben und Vorhersagen werden für die Query der nächsten Active Learning Iteration genutzt.

Die maximale Tokenanzahl wird für die beiden kleineren Datensätze auf 128 Tokens beschränkt. Diese Länge schließt mehr als 99% der Daten des Gab Hate

Korpus ein, während sie sogar den gesamten Subtitle Sentiments Datensatz abdeckt. Bei der Verarbeitung des arXiv Paper Abstracts Datensatzes musste die maximale Tokenanzahl aufgrund der längeren Textdaten auf 512 erhöht werden. Auch hier wird von dieser Längeneinschränkung nach der Anpassung nur noch ein Bruchteil der Daten beeinträchtigt.

Vor der ersten Query wird das Modell auf einem zufällig gewählten Pool aus gelabelten Beispielen trainiert, um die nötigen Parameter für die anschließende Auswahl der Beispiele durch die Querystrategien bereitstellen zu können. Der anfängliche Pool gelabelter Beispielen beträgt beim Gab Hate Korpus und dem Subtitle Sentiments Datensatz 50 Beispiele. Aufgrund der deutlich größeren Anzahl an Labeln beim arXiv Paper Abstracts Datensatz startet dieser mit einem Pool von 300 gelabelten Beispielen, welches ungefähr der doppelten Anzahl an Labeln in diesem Datensatz entspricht.

In den späteren Abfragen wird die Querygröße zwischen den Datensätzen nicht angepasst, um Vergleichbarkeit herzustellen. Des Weiteren sind die genutzten Querystrategien nicht für große Beispielzahlen optimiert und würden mit größeren Querys kaum Verbesserungen zeigen.

Zur Ausführung der Active Learning Iterationen wird die Small-Text Bibliothek genutzt. Diese stellt die benötigten Querystrategien und das Klassifikationsmodell bereit.

Um den Einfluss zufälliger Parameter einzuschränken und ein reproduzierbares Ergebnis zu ermöglichen, wird das Experiment für die Variablenkombinationen des ersten Teils und alle Verhältnisse des zweiten Teils insgesamt zehn Mal durchgeführt. Eine Ausnahme bildet hier der arXiv Paper Abstracts Datensatz, welcher durch die hohe Anzahl an Labeln einen großen Zeitaufwand in der Klassifikation und Bewertung durch die Querystrategien darstellt. Um diesen Zeitaufwand einzuschränken wurden hier nur fünf Wiederholungen ausgeführt. Die Auswertung erfolgt dann auf den Durchschnittswerten aller dadurch entstehenden Resultate.

4.3.2 Ergebnisse

Datensatz	**Querystrategie**			
	Random	LCI	CVIRS	AAL
Gab Hate Korpus	0,287	0,32	**0,323**	0,276
Subtitle Sentiments	**0,108**	0,105	0,104	0,102
arXiv Paper Abstracts	0,0056	0,0056	0,0057	**0,0057**

Tabelle 4.2: Macro-F_1-Wert für alle Kombinationen aus Datensätzen und Querystrategien nach der letzten Active Learning Iteration.

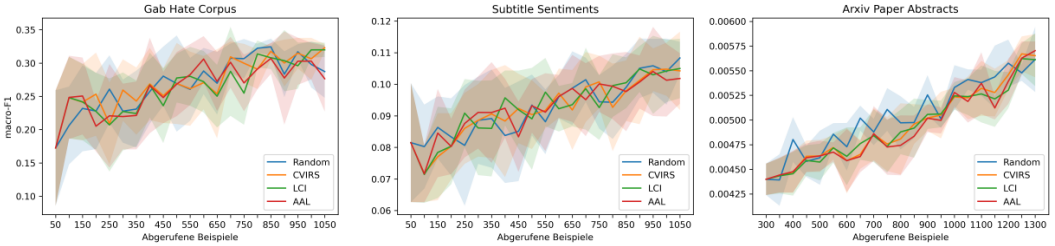

Abbildung 4.1: Macro-F_1 Lernkurven der Querystrategien auf den drei Testdatensätzen. Die ausgefüllten Flächen um die Kurven repräsentieren dabei die Standardabweichung.

Der Macro-F_1-Wert fällt bei allen Datensätzen und Querystrategien, wie in Tabelle 4.2 zu sehen, sehr niedrig aus. Dabei ist der Wert des Gab Datensatzes mit Abstand am höchsten, während der Wert des arXiv Datensatzes wie erwartet ein Bruchteil der anderen beträgt.

Aus Abbildung 4.1 wird ersichtlich, dass der Wert bei allen Kombinationen im Laufe der Versuche gleichmäßig ansteigt, wobei jedoch die Stärke des Anstiegs vom Datensatz abhängt. Die Leistung auf dem Gab Datensatz steigt dabei am stärksten um bis zu 0,15. Das stellt fast eine Verdopplung des F_1-Scores dar. Der geringste Anstieg liegt mit etwa 0,0013 beim arXiv Datensatz. Prozentual stellt dies ein Wachstum von etwa 28% dar. Der Subtitle Sentiments Datensatz zeigt einen Zuwachs des Macro-F_1-Scores von 0,027 beziehungweise etwa 33%.

Somit wächst der F_1-Score des Gab Datensatzes sowohl absolut, als auch prozentual am schnellsten und der des arXiv Paper Abstracts Datensatzes am langsamsten.

Die Leistung der individuellen Querystrategien hebt sich währenddessen nur marginal voneinander ab. Auf dem Gab Datensatz zeigt Tabelle 4.1, dass LCI und CVIRS deutlich besser als die Random Baseline und AAL funktionieren. Bei den beiden ersteren Konfigurationen fällt der F_1-Wert nach der letzten Iteration ungefähr so hoch aus wie die Summe des Durchschnitts der Random-Konfiguration und deren Standardabweichung. Der Abstand zwischen den beiden Querystrategien ist hier jedoch vernachlässigbar.

Aus dem ersten Diagramm der Abbildung 4.1 geht jedoch hervor, dass alle Querystrategien bis zur letzten Iteration sehr ähnliche Ergebnisse wie die Random Baseline zeigen. Die Zufallsstrategie erreicht auch schon vor LCI und CVIRS eine ähnlich hohe Spitze im F_1-Wert wie diese.

Die Standardabweichung der Random-Konfiguration bleibt wie zu erwarten ungefähr gleich. Lediglich in den Iterationen zwischen 750 und 850 abgefragten Beispielen sinkt sie kurz deutlich, steigt danach jedoch wieder an. Währenddessen sinkt sie bei den Strategien LCI und CVIRS vor allem während der

letzten sechs Iterationen deutlich.

Der Macro-F_1-Wert der Kombination mit der AAL Strategie ist im Gegensatz zu den anderen getesteten Strategien geringer als der der Baseline. Er steigt ab ungefähr 600 Beispielen nicht mehr bedeutend an, sondern zirkuliert um einen Wert von ungefähr 0,29. Die zugehörige Standardabweichung schwankt während des gesamten Experiments stark und sinkt auch in den letzten Iterationen kaum.

Auch im Subtitle Sentiments Datensatz besitzt die Konfiguration der AAL Strategie das niedrigste Endergebnis, während die Random Baseline das beste erzielt. Diese Unterschiede sind jedoch sehr gering. Die Querystrategien heben sich im Laufe des Experiments nicht voneinander ab. Auch die Standardabweichung bleibt über das gesamte Experiment stabil.

Auf dem arXiv Paper Abstracts Datensatz bleiben alle Strategien in fast allen Stadien des Experiments hinter der Baseline zurück. Nur in den letzten zwei Iterationen zeigen CVIRS, LCI und AAL in Abbildung 4.1 sichtbare höhere Werte als die Baseline. Diese sind jedoch im Kontext des Maßstabs unbedeutend. Die Standardabweichung ist hier im Vergleich zu den anderen beiden Datensätzen relativ gering und bleibt ebenso über das gesamte Experiment etwa gleich.

Analyse der Querys

Die Analyse der abgefragten Beispiele pro Query, vergleiche Abbildung 4.2, zeigt zusätzlich, dass sich das Auswahlverhalten der Querystrategien kaum voneinander abhebt. Der Anteil der abgefragten irrelevanten Beispiele ist bei allen Kombinationen relativ stabil und ändert sich kaum. Der Wert schwankt hierbei jeweils etwa um den Wert des gesamten Datensatzes. Die Querystrategien zeigen hier jedoch mehr Varianz und generell eine höhere Abweichung vom Mittelwert als die Random Baseline.

Beim Gab Hate Korpus und dem Subtitle Sentiments Datensatz werden durch die Querystrategien durchschnittlich mehr irrelevante Beispiele ausgewählt als durch die Zufallsstrategie. Dieser Effekt ist auf dem zweitem Datensatz stärker zu sehen als auf den Gab Daten. Die AAL Strategie hat hierbei die niedrigsten Spitzen.

Ein anderes Bild zeigt sich beim arXiv Paper Abstracts Datensatz, wo durch alle Querystrategien eher weniger irrelevante Beispiele abgefragt werden. Die CVIRS Strategie schwankt hier anfangs noch sehr in ihrer Beispielauswahl, pegelt sich jedoch in den letzten Querys mehr auf Werte ein, die niedriger als die Baseline liegen. LCI fragt von Anfang an relativ stabil weniger irrelevante Beispiele ab. Der Anteil irrelevanter Beispiele die die AAL Strategie abfragt, ist das ganze Experiment relativ sprunghaft, wählt aber durchschnittlich auch

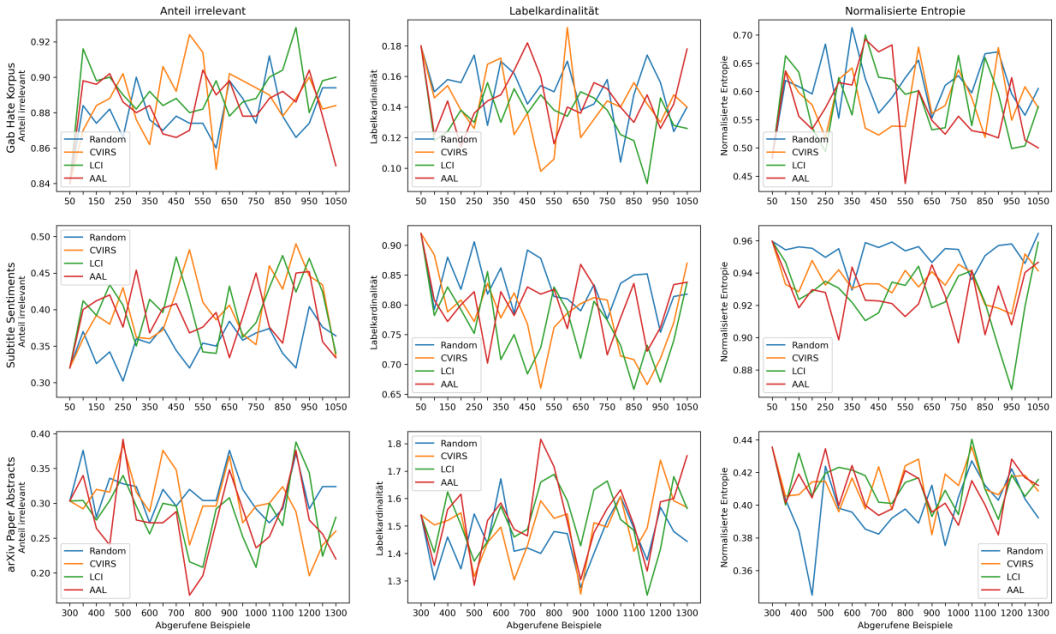

Abbildung 4.2: Zusammensetzung der abgefragten Beispiele pro Query für alle Kombinationen aus betrachteten Querystrategien und Datensätzen auf den drei Datensätzen, gemessen am Anteil der irrelevanten Beispiele (Anteil irrelevant) der Labelkardinalität und der normalisierten Entropie.

weniger irrelevante Bespiele als die Baseline aus.

Fast alle Kombinationen fragen jedoch in der letzten Query weniger irrelvante Beispiele ab als die Baseline. Die einzige Ausnahme hiervon ist die LCI Strategie auf dem Gab Datensatz.

Auch die normalisierte Entropie ist bei allen Datensätzen relativ gleichbleibend. Sie liegt bei allen Konfigurationen durchschnittlich leicht unter dem Wert der vollständigen Daten.

Die Entropiewerte der getesteten Querystrategien, zu sehen in der dritten Spalte in Abbildung 4.2, auf dem Gab Hate Korpus und dem arXiv Datensatz lassen sich kaum von der Random Baseline unterscheiden. Hierbei sind die Werte des Gab Datensatzes jedoch mit einer Spanne von 0,5 bis 0,7 deutlich breiter gefächert als die des arXiv Datensatzes mit einer Verteilung zwischen etwa 0,35 und 0,44.

Auf dem Subtitle Sentiments Datensatz zeigen beide Querystrategien jedoch niedrigere Werte als die Baseline. Diese sind für den Großteil des Experiments relativ stabil um ungefähr 0,03 niedriger als die Werte der Baseline. Während der letzten Iterationen löst sich diese Verbindung jedoch und die Verteilung erscheint deutlich zufälliger.

4.4 Variation des irrelevanten Anteils

Das zweite Experiment dient dazu, den Einfluss irrelevanter Daten auf die Klassifikation beim Active Learning genauer zu testen. Hierfür wird der Anteil irrelevanter Beispiele in einem Datensatz variiert und die Effektivität der Active Learning Simulationen mit verschiedenen irrelevanten Anteilen beobachtet. Für diese Evaluation habe ich den Subtitle Sentiments Datensatz ausgewählt, da seine Komplexität zwischen der der anderen beiden Datensätze liegt. Dies wird auch durch die Ergebnisse des ersten Experiments bestätigt.

4.4.1 Aufbau

Um den Anteil irrelevanter Beispiele im Subtitle Sentiments Datensatz zu variieren, wird der Datensatz vor Beginn der Active Learning Simulation jeweils so neu zusammengestellt, dass sich im resultierenden Datensatz 5%, 20%, 35%, 50%, 65%, 80% oder 95% irrelevanter Daten befinden. Um dies zu bewerkstelligen wird aus dem Teil der Daten, der im ursprünglichen Datensatz für das erwünschte Verhältnis zu viel vorhanden ist, zufällig eine passende Anzahl Beispiele ausgewählt. Für die Auswahl wird stratifiziertes Sampling anhand der Labelmengen der Beispiele wie von Sechidis et al. [2011] vorgeschlagen genutzt.

Bei den Anteilen irrelevanter Beispiele zwischen 95% und 35% liegt der gewünschte Anteil über dem gegebenen Anteil im Datensatz. Daher werden die relevanten Beispiele undersamplet, während die vollständige Anzahl irrelevanter Beispiele genutzt wird. Bei den beiden anderen Konfigurationen mit 20% und 5% irrelevanter Beispiele liegt der gewünschte Anteil irrelevanter Beispiele unter dem realen und die irrelevanten Beispiele werden undersamplet.

Durch diese Art des Samplings besitzen die Datensätze unterschiedliche Gesamtgrößen. Die höchste Anzahl an Beispielen besitzt der Datensatz mit einem Anteil von 35% irrelevanter Beispiele mit 26.966, die geringste besitzt der Datensatz mit 95% irrelevanten Daten mit 10.184 Datenpunkten. Dies stellt eine sehr große Spanne dar, welche das Experiment negativ beeinflussen könnte. Sollte die Anzahl der Beispiele eine Auswirkung auf die Klassifikation haben, ist zu erwarten, dass die kleineren Datensätze durch die geringere Anzahl an Beispielen unproportional schlechtere Macro-F_1-Werte zeigen.

In Kombination mit den vier Querystrategien Random, LCI, AAL und CVIRS ergeben sich durch das Sampling 28 Konfigurationen, welche nach dem jeweiligen Anteil irrelevanter Beispiele gesammelt in sieben Gruppen betrachtet werden. Für jede dieser Konfigurationen wird wie in Abschnitt 4.3.1 beschrieben ein Active Learning Experiment mit jeweils 20 Iterationen simuliert. Auch hier werden die durchschnittlichen Ergebnisse aus jeweils zehn Läufen betrachtet.

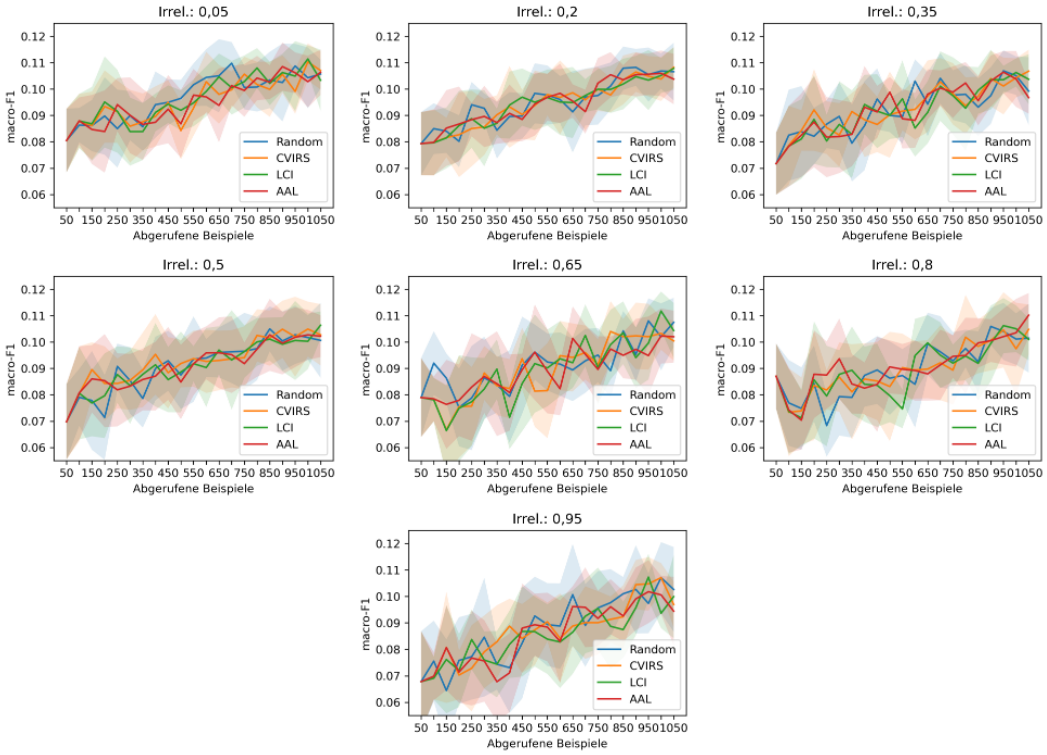

Abbildung 4.3: Macro-F_1 Lernkurven der getesteten Querystrategien auf den neu gesampleten Datensätzen mit 5% bis 95% irrelevanter Daten, hier gekennzeichnet mit Irrel.: 0,05 bis Irrel.: 0,95.

4.4.2 Ergebnisse

Im zweiten Experiment unterscheidet sich der Macro-F_1-Score zwischen den verschiedenen Anteilen irrelevanter Beispiele kaum. Der Wert wächst wie in Abbildung 4.3 zu sehen bei allen Experimenten von etwa 0,08 auf etwa 0,105 an. Dieser Wert entspricht etwa den Ergebnissen auf dem Subtitle Sentiments Datensatz aus dem ersten Experiment (siehe Abbildung 4.1). Entgegen diesem Trend stehen die Werte aus den Experimenten mit 35%, 50% und 95% irrelevanter Daten. Hier wächst der Macro-$F1$-Score von etwa 0,07 auf etwa 0,1 an.

Die F_1-Werte aller Konfigurationen auf den letzten drei Datensätzen zeigen sich insgesamt etwas sprunghafter als die der restlichen Datensätze.

Analyse der Querys

Die abgerufenen Anteile irrelevanter Beispiele gleichen in allen Experimenten etwa dem jeweiligen Anteil irrelevanter Daten im Trainingdatensatz. Die Que-

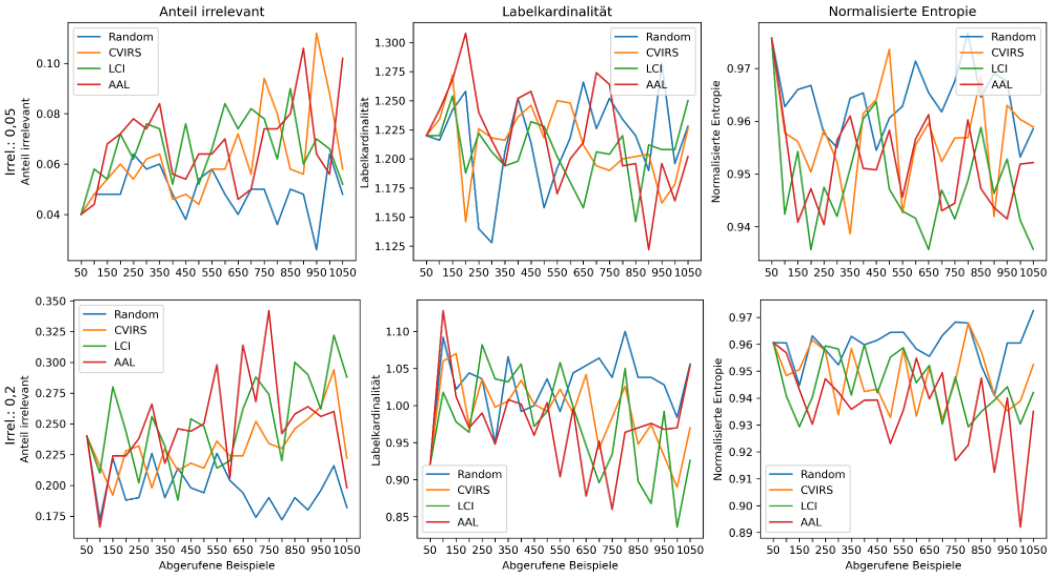

Abbildung 4.4: Zusammensetzung der abgefragten Beispiele pro Query für die getesteten Querystrategien auf den neu gesampleten Datensätzen mit weniger als 35% irrelevanten Daten, gemessen am Anteil irrelevanter Beispiele, der Labelkardinalität und der normalisierten Entropie.

rystrategien zeigen hier keine bedeutenden Unterschiede. Generell lässt sich beobachten, dass die Querystrategien alle tendenziell mehr irrelevante Beispiele abrufen als die Random Baseline.

Am stärksten ist dieser Effekt in der Gruppe der Simulationen mit 35% irrelevanten Beispielen zu sehen. Hier weichen die Querystrategien teilweise um mehr als 15 Prozentpunkte vom erwarteten Wert von 35% ab. Am stärksten ist dieser Effekt bei der LCI Strategie zu sehen. Die anderen Strategien übersteigen hier keine Abweichung von 10%

Auf den beiden Datensätzen mit den wenigsten irrelevanten Daten in Abbildung 4.4 zeigt sich bei allen Querystrategien im Laufe des Experiments ein leichter Aufwärtstrend im Anteil der abgefragten irrelevanten Daten. Die konsistentesten Abweichungen zeigen sich dabei in der Gruppe mit 20% irrelevanten Beispielen. Hier weichen alle drei Querystrategien in der zweiten Hälfte der Experimente mindestens um etwa 5 Prozentpunkte von der Random Baseline ab.

Je höher der Anteil irrelevanter Beispiele in den Datensätzen, desto weniger stark zeigt sich dieser Effekt. Während die echten Querystrategien LCI, AAL und CVIRS bei 80% irrelevanten Daten in den letzten Querys höhere Anteile irrelevanter Daten abfragen, bewegen sie sich im Rest des Experimentsverlauf sehr nah an der Baseline.

Abbildung 4.5: Zusammensetzung der abgefragten Beispiele pro Query für die getesteten Querystrategien auf den neu gesampleten Datensätzen mit 35% oder mehr irrelevanten Daten, gemessen am Anteil irrelevanter Beispiele, der Labelkardinalität und der normalisierten Entropie.

Der letzte Datensatz zeigt wiederum keine bedeutenden und konsistenten Abweichungen von der Baseline. Die Querystrategien fragen hier durschschnittlich etwa den gleichen Anteil irrelevanter Daten ab wie die Baseline und zeigen dabei auch keine großen Spitzen.

Analog sinkt auch die Labelkardinalität mit steigendem Anteil irrelevanter Beispiele. Sie spiegelt auch die Irregularitäten der Querystrategien, die beim Anteil der abgerufenen irrelevanten Beispiele auftreten, wider.

Die Klassenverteilung innerhalb der Querys scheint über die meisten Konfigurationen hinweg relativ stabil. Die Querystrategien liegen bei allen Konfigurationen zwischen 5% und 50% irrelevanter Daten etwa drei bis sechs Hundertstel unter dem Wert der normalisierten Entropie der Baseline. Die Konfiguration mit 65% irrelevanten Daten zeigt am Ende eine größere Spanne, wo AAL und LCI jeweils 7,5 und 10 Hunderstel unter der Baseline liegen (Abbildung 4.5). Bei der Konfiguration mit 80% irrelevanten Daten fällt der Wert der normalisierten Entropie der Querstrategien deutlich weiter unter die Baseline. Hier liegt er teilweise fast 15 Hunderstel unter dem Kontrollwert.

Des Weiteren lässt sich beobachten, dass die durchschnittliche normalisierte Entropie mit einem steigenden Anteil irrelevanter Daten sinkt. Während sich die Entropie in den ersten zwei Datensätzen relativ konsistent im Bereich von 0,95 bewegt, sinkt sie im dritten Datensatz auf durchschnittlich etwa 0,93-0,94 für die Querystrategien, während die Baseline hier bei 0,95 bleibt. In den folgenden zwei Datensätzen mit 50% und 65% irrelevanten Beispielen sinkt die Entropie weiter um jeweils etwa 0.02 auf respektive etwa 0,92 und 0,9. Hier fallen jedoch nicht nur die echten Querystrategien, sondern auch die Werte der Random Baseline ab. Ein deutlich größerer Abfall der normalisierten Entropie zeigt sich bei 80% irrelevanten Daten. Hier sinkt die Klassenbalance auf Werte um etwa 0,8. Außerdem entwickelt sich hier eine deutlich höhere Instabilität des Werts. Während er sich vorher in einem Experiment im Bereich von etwa 5 Hundertsteln verändert hat, gibt es hier eine Spanne von Werten von 0,7 bis etwa 0,93. So springt die normalisierte Entropie von LCI in den letzten drei Querys von etwa 0,7 auf etwa 0,92 und fällt wieder zurück auf etwa 0,73. Diese Entwicklung setzt sich im Datensatz mit 95% irrelevanten Daten deutlich fort. Der Durchschnitt liegt hier bei etwa 0,6 mit einer Spanne von Werten zwischen 0,41 und 0,8. Dies entspricht ungefähr der achtfachen Spanne der ersten fünf Datensätze.

Mit steigendem Anteil der irrelevanten Daten sinkt hier also die Klassenbalance. Dieser Effekt zeigt sich aber erst in den letzten Datensätzen wirklich bedeutend.

Es ist keine klare Korrelation zwischen dem Ergebnis des Macro-$F1$-Scores und den verschiedenen Metriken zur Queryauswertung zu erkennen.

Kapitel 5

Diskussion der Experimente

Auch wenn Active Learning in keinem der Experimente Vorteile gegenüber der Random Baseline gezeigt hat, lassen sich einige Erkenntnisse aus den Versuchen ablesen. Der generell niedrige F_1-Wert zeigt, dass die getesteten Datensätze schwer zu lösende Probleme darstellen.

5.1 Evaluierung der Querystrategien

Wie erwartet, stellt sich der Gab Hate Korpus als leichteste Klassifikationsaufgabe heraus und zeigt somit die besten Ergebnisse. Währenddessen funktioniert die Klassifikation des arXiv Paper Abstracts Datensatzes am schlechtesten. Der Subtitle Sentiments Datensatz siedelt sich erwartungsgemäß dazwischen an.

Active Learning

Entgegen der Erwartungen zeigt sich jedoch in keinem der drei Datensätze des ersten Experiments, dass die Querystrategien die Klassifikation unterstützen. Sie heben sich in Bezug auf den F_1-Wert weder bedeutend von der Baseline noch voneinander ab. Das entspricht den Ergebnissen von Wertz et al. [2022] für extrem imbalancierte Szenarien. Daher ist dieses Ergebnis zumindest für die Klassifikation des arXiv Datensatzes plausibel. Die anderen Datensätze zeigen jedoch genauso kein funktionierendes Active Learning, obwohl sie deutlich einfachere Probleme darstellen.

Die Endergebnisse des Gab Hate Korpus' in Tabelle 4.2 suggerieren zwar, dass LCI und CVIRS die Baseline übertreffen, tun dies aber nicht wirklich. Die Konfiguration der Random Baseline erreicht schon einige Iterationen früher bei ähnlicher Standardabweichung einen höheren F_1-Wert als die Endergebnisse der beiden genannten Strategien. Danach fällt sie wieder auf einen niedrigeren

Wert als den der beiden Strategien ab. Diese Entwicklungen können daher als zufallsbedingt eingeschätzt werden, da sich die Strategien auch im sonstigen Verlauf des Experiments nicht bedeutend von der Random Baseline unterscheiden. Dieser Schluss wird außerdem dadurch unterstützt, dass die AAL Strategie, welche sonst sehr ähnliche Ergebnisse wie die anderen Strategien liefert, hier sogar hinter der Baseline zurückbleibt.

Im Subtitle Sentiments Datensatz zeigt sich das umgekehrte Ergebnis, da die Baseline die Querystrategien zu übertreffen scheint. Auch diese Entwicklung kann aber dem Zufall zugeschrieben werden.

Bezogen auf die erste Forschungsfrage lässt sich also sagen, dass die Klassifikation sehr schlechte Ergebnisse produziert. Active Learning funktioniert in diesem Szenario nicht wirklich und bringt keinen erheblichen Vorteil in der Klassifikation, da die Kombinationen in denen die Querystrategien eingebunden sind in ihrer Klassifikationsqualität keinen bedeutenden Unterschied zur Baseline zeigen.

Funktionsweise der Querystrategien

Die Querystrategien scheinen die Beispiele in Hinsicht auf Irrelevanz genauso zufällig auszuwählen wie die Baseline. Während sie um dem Klassifikator den meisten Informationsgehalt zu liefern, vermutlich weniger irrelevante Beispiele auswählen sollten, zeigt sich kein solcher Trend. Der größte Unterschied zwischen Baseline und Querystrategien ist dabei im Subtitle Sentiments Datensatz zu sehen, wo durch die Strategien jedoch mehr irrelevante Beispiele als von der Baseline abgerufen werden. Auch hier zeigen die Querystrategien also nicht den erwünschten Effekt.

In Hinsicht auf die Klassenverteilung bieten die Querystrategien ebenso keinen Vorteil gegenüber der Baseline. Dies zeigt sich in den normalisierten Entropie pro Query, die hier für alle Querystrategien ähnlich oder niedriger liegt als die der Baseline. Der Gab Datensatz zeigt hier jedoch deutlich größere Schwankungen in der Klassenimbalance pro Query als die anderen Datensätze. Das lässt sich vermutlich durch den hohen Anteil irrelevanter Beispiele erklären. In einer Query von 50 Beispielen sind bei 12 Prozent relevanten Beispielen durchschnittlich nur 6 relevante Datenpunkte vorhanden. Durch diese geringe Stichprobengröße ergibt sich bei zufälliger Auswahl eine große mögliche Varianz in der Klassenverteilung.

Auch bei der normalisierten Entropie zeigt sich der eindeutigste Trend entgegen der Baseline jedoch beim Subtitle Sentiments Datensatz. In diesem besitzen alle Querys der echten Querystrategien eine größere Imbalance als die Zufallsstrategie.

Somit ist der Subtitle Sentiments Datensatz der, der in allen Metriken zur

Analyse der Querys die klarsten Ergebnisse zeigt. Daher stellt sich die Frage, worin sich der Datensatz von den anderen unterscheidet. Innerhalb dieser drei Datensätze trägt er die höchste Labeldichte und normalisierte Entropie. Während die Wichtigkeit einer balancierten Klassenverteilung bekannt ist [Wertz et al., 2022], ist der Einfluss der Labeldichte bisher kaum erforscht worden. Daher lässt sich hier keine klare Aussage treffen, ob eine der beiden Metriken diesen Unterschied in der Klassifikation auslöst.

Generell lässt sich aus den gemessenen Metriken nicht ablesen, dass die Querystrategien informationsreichere Beispiele abrufen als die Baseline. Daher hat Active Learning auch in der Beispielauswahl nicht funktioniert und die zweite Forschungsfrage kann ebenfalls mit nein beantwortet werden.

5.2 Variation des irrelevanten Anteils

Allgemein lässt sich zum zweiten Active Learning Experiment sagen, dass der Anteil irrelevanter Beispiele entgegen den Erwartungen in diesem Szenario keinen Einfluss auf die Höhe des Macro-F_1-Werts und somit die Klassifikationsqualität hat. Zwar zeigt der Datensatz mit dem höchsten Anteil irrelevanter Beispiele das niedrigste Ergebnis, ein ähnliches wird aber auch auf den Datensätzen mit 35% und 50% irrelevanten Daten erreicht.

Der größte Unterschied der sich bei den größeren Anteilen irrelevanter Daten abbildet ist, dass die F_1-Scores der Active Learning Simulationen auf den Datensätzen zwischen 65% und 95% deutlich sprunghafter sind als auf den Datensätzen mit weniger irrelevanten Beispielen. Das gleicht der Beobachtung des Gab Hate Korpus im ersten Experiment, welcher auch deutlich größere Sprünge in der Klassifikation gezeigt hat als die anderen Datensätze.

Die Datensätze mit weniger irrelevanten Beispielen zeigen außerdem deutlich, dass die Querystrategien mehr irrelevante Beispiele abrufen als die Baseline. Dieser Trend verschwindet jedoch immer weiter, je mehr irrelevante Beispiele vorhanden sind. Dies könnte daran liegen, dass selbst bei einer gezielten Auswahl kaum mehr irrelevante Beispiele abgefragt werden können. Bei zufälliger Auswahl sind in einer Query auf dem letzten Datensatz durchschnittlich 48 von 50 Beispielen irrelevant. Dieser Wert lässt sich kaum noch weiter steigern. Generell bestätigt dies das Negativergebnis für Active Learning aus dem ersten Experiment, da die Querystrategien ihren Zweck nicht erfüllen.

Des Weiteren fällt die normalisierte Entropie der Querys je mehr irrelevante Beispiele vorhanden sind. Während der Wert in den ersten Datensätzen noch kleine Sprünge macht, sinkt die Klassenbalance pro Query in den letzten zwei Datensätzen mit größeren Schritten. Außerdem wird die gemessene Entropie mit steigendem Anteil irrelevanter Daten auch immer sprunghafter.

Es ist wahrscheinich, dass auch hier wie beim Gab Hate Korpus die sinkende Anzahl relevanter Beispiele pro Query und somit die Stichprobengröße für die normalisierte Entropie der Grund für diese Entwicklung ist.

Einfluss irrelevanter Beispiele

Zur dritten Forschungsfrage lässt sich sagen, dass der Einfluss irrelevanter Beispiele auf die Klassifikationsqualität in diesen Szenarien kaum erkennbar ist. Das könnte an weiteren Problemen liegen, die den Einfluss der irrelevanten Daten überschatten. So wird die Leistung der Querystrategien auch deutlich von der generell schlechten Klassifikationsqualität beeinflusst, da sie auf deren Vorhersagen beruhen und diese zur Entscheidung, welche der Beispiele abgerufen werden sollen, gebrauchen. Weitere einflussreiche Faktoren die die Klassifikation und Querystrategien beeinflussen könnten, sind die Labelanzahl und -verteilung.

Der Anteil irrelevanter Daten scheint aber zumindest leichten Einfluss auf die Stabilität des F_1-Werts zu haben.

Außerdem beeinflusst er die Effektivität der Querystrategien. Bei niedrigen Anteilen irrelevanter Beispiele zeigen sie noch Unterschiede zur Baseline, auch wenn diese nicht die erhofften sind. Bei höheren Anteilen irrelevanter Daten verschwinden diese Unterschiede jedoch und die Querystrategien scheinen sowohl in Bezug auf den Anteil irrelevanter Beispiele als auch auf die Klassenimbalance immer sprunghafter in ihrer Beispielauswahl.

Limitierungen

Eine der Limitierungen des zweiten Experiments ist, dass die Datensätze sich nicht nur in ihrem Anteil an irrelevanten Beispielen unterscheiden, sondern auch unterschiedliche Gesamtzahlen an Beispielen haben. Generell sind immer so viele relevante Beispiele wie bei diesem Anteil irrelevanter Beispiele möglich in den Datensätzen vorhanden. Während in den beiden Datensätzen mit 5% und 20% irrelevanter Beispiele alle 17.528 möglichen relevanten Beispiele enthalten sind, sinkt ab dem Datensatz mit 35% irrelevanter Daten mit steigendem Anteil letzterer die Anzahl an relevanten Beispielen, während die Anzahl irrelevanter Beispiele gleich bleibt. Die ursprüngliche Erwartung war hier, dass die Klassifikationqualität, gemessen anhand des Macro-F_1-Scores mit zunehmendem Anteil irrelevanter Beispiele sinkt. Außerdem scheint es intuitiv, dass die Klassifikation ebenfalls mit sinkender Anzahl relevanter Beispiele an Effektivität verliert. Wie in den Lernkurven jedoch zu beobachten ist, unterscheidet sich der Macro-F_1-Wert der neu zusamengestellten Datensätze nur marginal. Gerade im Vergleich der Kurven der irrelevanten Anteile von 35% und 50%,

wo sich der erste große Abfall relevanter Beispiele ereignet, zeigt sich kaum ein Unterschied in den Ergebnissen.

Der konsistenteste Unterschied der Kurven ist die höhere Sprunghaftigkeit der F_1-Werte bei den Datensätzen mit mehr als 50% irrelevanten Daten. Diese Datensätze sind generell die kleinsten, da sie alle weniger als 15.000 Beispiele besitzen, während die anderen Datensätz als mehr als 18.500 Datenpunkte beinhalten. Diese Diskrepanz könnte eine Ursache für die erhöhte Instabilität sein. Sie lässt sich jedoch ohne weitere Experimente nicht sicher vom Einfluss der vielen irrelevanten Beispiele abgrenzen.

Um diese Bedenken endgültig zu eliminieren, müsste ein weiteres Experiment durchgeführt werden, in dem die Größe der Datensätze gleich gehalten wird, und ausschließlich der Anteil irrelevanter Beispiele variiert wird. Hier lässt sich jedoch anhand der gemessenen Daten kein klarer Einfluss der Datensatzgröße ablesen.

Kapitel 6

Fazit

In dieser Arbeit habe ich drei Textdatensätze mit großem Anteil irrelevanten Daten gesucht und zusammengestellt, um ein Experimentsetup für Active Learning Simulationen für das Problem vieler irrelevanter Beispiele zusammenzusetzen. Auf diesem habe ich die drei Multi-Label Querystrategien LCI, AAL und CVIRS angewendet.

Die Evaluation des Experiments zeigt, dass die Klassifikationsqualität generell sehr schlecht ausfällt. Dieses Ergebnis war durch die hohe Komplexität der Datensätze zu erwarten. Außerdem funktioniert jedoch auch Active Learning selbst auf diesen Daten nicht, da es keinen bedeutenden Vorteil gegenüber der Baseline erreicht. Es ist daher für die Klassifikation der Multi-Label Textdaten in diesem Szenario nicht sinnvoll anwendbar.

Dabei scheint das Vorhandensein irrelevanter Daten hier aber keinen großen Einfluss auf das Ergebnis der Klassifikation und die Leistung der Active Learning Simulationen zu haben. Anders als zunächst angenommen, sind irrelevante Beispiele auf den hier evaluierten Daten nicht das einflussreichste Problem.

Die Querystrategien unterscheiden in ihrer Auswahl die irrelevanten Beispiele dabei nur bei einem der Datensätze von den relevanten. Auch dort existiert dieser klare Unterschied jedoch nur bei Anteilen irrelevanter Beispiele unter 50%. Die Strategien scheinen also das Potential zu haben, irrelevante Beispiele zu detektieren, tun dies aber unter Anderem aufgrund anderer Faktoren in der Klassifikation hier nicht.

Literaturverzeichnis

Shayok Chakraborty, Vineeth Balasubramanian, and Sethuraman Panchanathan. Optimal batch selection for active learning in multi-label classification. In *Proceedings of the 19th ACM international conference on multimedia*, pages 1413–1416, 2011.

Francisco Charte, Antonio Rivera, María José del Jesus, and Francisco Herrera. A first approach to deal with imbalance in multi-label datasets. In *Hybrid Artificial Intelligent Systems: 8th International Conference, HAIS 2013, Salamanca, Spain, September 11-13, 2013. Proceedings 8*, pages 150–160. Springer, 2013.

Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active learning for bert: an empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, 2020.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*,

pages 523–540, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.38. URL `https://aclanthology.org/2023.eacl-main.38`.

Brendan Kennedy, Mohammad Atari, and Aida Mostafazadeh et al. Davani. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Lang Resources & Evaluation*, 56:79–108, 2022.

Florian Laws and Hinrich Schütze. Stopping criteria for active learning of named entity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 465–472, 2008.

David D Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and development in Information Retrieval*, pages 3–12, 1994.

Xin Li and Yuhong Guo. Active learning with multi-label svm classification. In *IjCAI*, volume 13, pages 1479–1485. Citeseer, 2013.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.51. URL `https://aclanthology.org/2021.emnlp-main.51`.

Dominic Mazzoni, Kiri L Wagstaff, and Michael C Burl. Active learning with irrelevant examples. In *European Conference on Machine Learning*, pages 695–702. Springer, 2006.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. Xed: A multilingual dataset for sentiment analysis and emotion detection. In *The 28th International Conference on Computational Linguistics (COLING 2020)*, 2020.

Sayak Paul and Soumik Rakshit. arxiv paper abstracts, 2021. URL `https://www.kaggle.com/datasets/spsayakpaul/arxiv-paper-abstracts`. accessed 05.2023.

Robert Plutchik. *A general psychoevolutionary theory of emotion*, chapter 1, pages 3–31. 1980.

Anant Raj and Francis Bach. Convergence of uncertainty sampling for active learning. In *International Conference on Machine Learning*, pages 18310–18331. PMLR, 2022.

Oscar Reyes and Sebastián Ventura. Evolutionary strategy to perform batch-mode active learning on multi-label data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(4):1–26, 2018.

Oscar Reyes, Carlos Morell, and Sebastián Ventura. Effective active learning strategy for multi-label learning. *Neurocomputing*, 273:494–508, 2018.

Christopher Schröder, Lydia Müller, Andreas Niekler, and Martin Potthast. Small-text: Active learning for text classification in python. *arXiv preprint arXiv:2107.10314*, 2021.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pages 145–158. Springer, 2011.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Muhammad Atif Tahir, Josef Kittler, and Fei Yan. Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recognition*, 45(10):3738–3750, 2012.

Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3 (3):1–13, 2007.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Andreas Vlachos. A stopping criterion for active learning. *Computer Speech & Language*, 22(3):295–312, 2008.

Mengqi Wang and Ming Liu. An empirical study on active learning for multi-label text classification. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 94–102, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.insights-1.12. URL `https://aclanthology.org/2023.insights-1.12`.

Lukas Wertz, Katsiaryna Mirylenka, Jonas Kuhn, and Jasmina Bogojeska. Investigating active learning sampling strategies for extreme multi label text classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4597–4605, 2022.

Lukas Wertz, Jasmina Bogojeska, Katsiaryna Mirylenka, and Jonas Kuhn. Reinforced active learning for low-resource, domain-specific, multi-label text classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10959–10977, 2023.

Bang Zhang, Yang Wang, and Fang Chen. Multilabel image classification via high-order label correlation driven active learning. *IEEE Transactions on Image Processing*, 23(3):1430–1441, 2014.

Jifan Zhang, Julian Katz-Samuels, and Robert Nowak. Galaxy: Graph-based active learning at the extreme. In *International Conference on Machine Learning*, pages 26223–26238. PMLR, 2022.