Vorstudie

Masterarbeit

Einsatzmöglichkeiten von Machine Learning bei der Analyse von Kundenrückmeldungen

Fabian Thouet (thouefab@students.zhaw.ch)

ZHAW School of Management and Law

Modul «Vorstudie Masterarbeit»

Prof. Dr. Alexandre de Spindler

Herbstsemester 2022

6.12.2022

Inhaltsverzeichnis

[1 Problemstellung 1](#_Toc121231340)

[2 Lösungsvorschlag 2](#_Toc121231341)

[3 Nutzenprognose 3](#_Toc121231342)

[4 Herausforderung 4](#_Toc121231343)

[4.1 Forschungslücke 4](#_Toc121231344)

[4.2 Forschungsfrage 4](#_Toc121231345)

[4.2.1 Hypothese 4](#_Toc121231346)

[5 Methodik 5](#_Toc121231347)

[5.1 Arbeitspakete 5](#_Toc121231348)

[5.1.1 Arbeitspaket 1: Daten 5](#_Toc121231349)

[5.1.2 Arbeitspaket 2: NLP-Prozess 5](#_Toc121231350)

[5.1.3 Arbeitpaket 3: Anwendung 5](#_Toc121231351)

[5.1.4 Arbeitspaket 4: Validierung 5](#_Toc121231352)

[5.2 Ressourcen 6](#_Toc121231353)

[5.2.1 Wissen 6](#_Toc121231354)

[5.2.2 Ressourcen 6](#_Toc121231355)

[6 Validierung 7](#_Toc121231356)

[6.1 Topic Detektion & Sentimente 7](#_Toc121231357)

[6.2 Reproduzierbarkeit 7](#_Toc121231358)

[7 Artefakte 8](#_Toc121231359)

[7.1 Modell 8](#_Toc121231360)

[7.1.1 Topic Detektion 8](#_Toc121231361)

[7.1.2 Sentiment Analysis 9](#_Toc121231362)

[7.2 Statistiken 11](#_Toc121231363)

[7.2.1 Topic Detektion 11](#_Toc121231364)

[7.2.2 Sentiment Analysis 11](#_Toc121231365)

[7.2.3 Durchschnittswerte der Statistiken 13](#_Toc121231366)

[7.2.4 Sentimente pro Topic 14](#_Toc121231367)

[8 Fazit 15](#_Toc121231368)

[Literaturverzeichnis 16](#_Toc121231369)

[Abbildungsverzeichnis 17](#_Toc121231370)

[Tabellenverzeichnis 18](#_Toc121231371)

# Problemstellung

In der heutigen Zeit sind Unternehmen den Kunden näher denn Jeh, dies aufgrund der erweiterten Interaktionsmöglichkeiten, welche eine digitalisierte Welt mit sich bringt, Kunden können Rückmeldungen zum Unternehmen, zu Produkten, oder Mitarbeitern auf verschiedene Arten äussern. Heutzutage geschieht dies zu einem grossen Teil im Internet in der Form von sogenannten Reviews, welche gewöhnlicherweise eine Bewertung und eine Textäusserung zulassen. Ein Ziel von vielen Unternehmen ist es diese Reviews zu sammeln und zu bewerten, dies geschieht auf verschiedene Arten und Weisen, zum Beispiel durch manuelle Bewertungen, Einordnungen durch einen Mitarbeiter oder durch den Einsatz von Algorithmen. Die manuelle Einordnung durch einen Mitarbeiter benötigt jedoch viel Zeit und Aufwand und durch die immer grössere Datenmenge wird dieses Problem nur verschärft, daher liegt der Einsatz von Algorithmen nahe welche Tausende und Abertausende von Rückmeldungen durchsuchen können, um diese zu bewerten.

Die Frage, welche sich hier jedoch stellt, ist wie kann erkannt werden warum ein Produkt, ein Unternehmen oder etwas mit dem Unternehmen im Zusammenhang Stehendes eine schlechte, gute oder neutrale Bewertung erlangt?

# Lösungsvorschlag

Im Rahmen der Masterarbeit stellte sich genau diese Frage, wie können Kundenrückmeldungen mit Hilfe von Algorithmen analysiert werden. Da es sich bei den Rückmeldungen handelt, welche einen Text beinhalten wurde, der Einsatz von Natural Language Processing Algorithmen (NLP) genauer betrachtet und für einsetzbar befunden. Da es sich bei dieser Art von Texten um einen Ausdruck von Gefühlen handelt wurde, mit Unterstützung durch den Betreuer dieser Arbeit, Prof. Dr. Alexandre de Spindler, der Einsatz der Algorithmen angestrebt, welche ebendiese Gefühle erkennen können. Dabei handelt es sich um eine sogenannte Sentiment-Analysis und es wurden in einer ersten Phase die Algorithmen «Textblob» und «Vader Sentiment Analysis» verwendet.

Es ist zu erwähnen das es sich bei dieser Arbeit um eine Kooperation mit der Migros Gruppe, im spezifischen dem Migros-Genossenschafts-Bund (MGB), dem Arbeitgeber des Autors, handelt und Kundenrückmeldung vom MGB verwendet werden. Diese Rückmeldungen standen jedoch nicht beim Start der Analyse zur Verfügung, weshalb die Vorstudie mit Beispieldaten der Plattform «Kaggle.com» vollzogen wurde.

# Nutzenprognose

Die Vorstudie und dahingehend auch die Masterarbeit hat verschiedene Nutzen, einerseits wird durch die Erarbeitung des NLP-Prozesses zur Analyse von Kundenrückmeldung ebenfalls ein Beitrag zu den Datenwissenschaften erbracht und auf der anderen Seite wird mit einer Schweizer Firma, dem MGB, zusammengearbeitet welcher sich vom vermehrten Einsatz solcher NLP und auch anderer Datenwissenschaftlicher Auswertungen einen Marktwirtschaftlichen Vorteil erhofft. Dies Aufgrund des geförderten Wissens der Mitarbeiter im MGB, welche in der Zukunft ähnliche Prozesse innerhalb des Unternehmens erarbeiten können oder die bessere Analyse der Kundenrückmeldungen und somit ein besseres Verständnis für Kundenwünsche, welche einen marktwirtschaftlichen Vorteil bieten könnte. Eine genaue Bezifferung dieser Nutzen ist momentan nicht möglich.

# Herausforderung

Der folgende Abschnitt behandelt die Forschungslücke und die Forschungsfrage und stellt zusätzlich die Hypothese der Masterarbeit auf.

## Forschungslücke

Der Einsatz an NLP-Lösungen zur Analyse von Kundenrückmeldung ist sicherlich keine neue Disziplin, jedoch ist das Feld der Datenanalyse mittels Algorithmen im Allgemeinen, im Kontext der Wissenschaften, eine relativ junge Disziplin und kann durch wiederkehrende Aufzeigen oder verschiedenen Kombinationen beim Einsatz von Algorithmen evtl. zu neuen Erkenntnissen oder zur Erweiterung bereits bestehenden Wissens führen. Des Weiteren wurde bereits in den oberen Abschnitten erklärt, dass es sich bei den Daten der Masterarbeit um Daten des MGB handelt und somit die daraus gewonnen Erkenntnisse, neue Erkenntnisse darstellen.

## Forschungsfrage

Die Forschungsfrage ergibt sich aus der Problemstellung, der Nutzenprognose und der beschriebenen Forschungslücke:

* Wie gestaltet sich ein NLP-Prozess zur Analyse von Kundenrückmeldungen mittels NLP?

### Hypothese

Die Forschungsfrage wird mittels einer aufgestellten Hypothese zu beantworten versucht. Diese lautet:

* Dies kann durch eine Kombination von Topic-Detektion Analyse und Sentiment Analyse herausgefunden werden.

# Methodik

Im folgenden Abschnitt wird eine grobe Übersicht geboten, welche dem Leser einen Überblick über den Ablauf der Arbeit geben soll, es werden die zu dem bearbeitenden Punkte in Arbeitspakete unterteilt, mit einen Kurzbeschrieb versehen und aufgelistet.

## Arbeitspakete

In den folgenden Unterkapiteln werden die Arbeitspakete kurz beschrieben.

### Arbeitspaket 1: Daten

Wie bereits beschrieben wurden für die durchgeführte Analyse Daten von Kaggle.com bezogen. Diese Daten dienten dem Autor als Grundlage um die folgenden Arbeitspakete bearbeiten zu können. Es ist jedoch angedacht Daten von der Migros-Gruppe, im speziellen des Migros-Genossenschaftsbundes, zu beziehen. Aufgrund von verschiedenen Abklärungen betreffend des Datenschutzes verzögert sich die zur Verfügungstellung dieser Daten. Der Autor hat im Monat Oktober und November das DGF (Data-Governance-Framework) des MGB durchlaufen, und eine positive Rückmeldung erhalten, jedoch steht die zur Verfügungstellung der Daten noch aus.

### Arbeitspaket 2: NLP-Prozess

Das zweite Arbeitspaket befasst sich mit der Erstellung eines NLP-Prozesses mit welcher sich die die anstehenden Analysen durchführen lassen. Im Prozess kommen verschiedene Algorithmen zur Anwendung, dies sind einerseits Algorithmen wie spaCy welche einen NLP ausführen können, dies wurde eingesetzt, um die Topics zu erkennen, des Weiteren wurden die Algorithmen Textblob und Vader-Sentiment-Analysis angewendet, um die Sentimente der jeweiligen Topics beziehungsweise der Beschriebe zu erkennen.

### Arbeitpaket 3: Anwendung

Die Anwendung dieser Algorithmen sieht vor zuerst die Daten zu beziehen und danach einen Prozess aufzubauen, um diese zu analysieren. Hierbei wurde wie bereits oben beschrieben die genannten Daten gebraucht und danach mittels der beschriebenen Algorithmen analysiert. Diese Daten wurden danach in einer SQL-Datenbank gespeichert und stehen für Visualisierungen und Analysen zur Verfügung.

Dieser Prozess wird wiederholt, sobald die Daten des MGB zur Verfügung stehen, auch wird die Möglichkeit des Einsatzes dieser Algorithmen auf die deutsche Sprache in Betracht gezogen da Deutsche Bewertungen die Mehrheit der abgegebenen Bewertungen in den MGB-Daten ausmacht.

### Arbeitspaket 4: Validierung

Im letzten Arbeitspaket müssen die Auswertungen, welche generiert wurden, validiert werden hiermit ist gemeint, dass die vorgestellte Lösung, das vorgestellte Problem lösen kann. Des Weiteren wird die Reproduzierbarkeit des Nachweises erbracht.

## Ressourcen

Im folgenden Teil werden die Ressourcen, welche für die Durchführung der Arbeit benötigt werden, näher beschrieben, mit Ressourcen sind nicht nur physische Ressourcen gemeint, sondern auch die Fähigkeiten oder das Wissen, welches benötigt wird, um die Arbeit erarbeiten zu können.

### Wissen

Bei der Entwicklung eines NLP-Prozesses braucht man gewisse Vorkenntnisse in Datenwissenschaften, diese hat sich der Autor während des Studiums erarbeitet. Des Weiteren werden in der Analyse die Programmiersprache Python und SQL verwendet, daher sind Vorkenntnisse in beiden Sprachen von Vorteil. Da NLP-Prozesse ein sehr spannendes und häufig untersuchtes Thema sind findet man sehr viele Informationen hierzu im Internet, zum Beispiel auf «w3schools.com» oder «towardsdatascience.com» nebst vielen weiteren Webseiten.

### Ressourcen

Betreffend Software, welche für die Analyse gebraucht wird, wurde Visual Studio Code von Microsoft gebraucht und für die Code-Verwaltung Github.com gebraucht. Beides sind Programme welche kostenlos zur Verfügung stehen, jedoch sei erwähnt, dass der Autor Zugang zu Visual Studio Code durch den Arbeitgeber zur Verfügung gestellt werden kriegt. Eine weitere Erwähnung sei das als Datenbank momentan das kostenlose Programm SQLite gebraucht wird, sollte sich dies als Bedingung durch den Arbeitgeber ändern und eine nicht kostenfreie Datenbank gebraucht werden, wird dies nachgeführt.

# Validierung

Das folgende Kapitel soll einen Überblick geben, ob die erarbeitete Lösung die dargestellte Problematik beantworten kann.

## Topic Detektion & Sentimente

Die Vorstudie hat gezeigt, dass mittels NLP-Prozess Topics und deren Sentimente in Bewertungen von Kunden erkannt werden können, dies lässt den Schluss zu, dass die Erarbeitung eines solchen Prozesses sinnvoll ist. Es sei jedoch gesagt, dass die momentan erarbeitete Lösung nicht der finalen entspricht, da zum Beispiel vielen Annahmen getroffen wurden was die die Bewertung von positiv, negativ, oder neutralen Bewertungen betrifft und sich somit die positiv erkannten Bewertungen der Algorithmen, von denen der Nutzer unterscheiden. Dies müsst sicherlich genauer untersucht werden und bei der weiteren Bearbeitung der Arbeit genau dokumentiert werden.

## Reproduzierbarkeit

Wie bereits in oberen Kapiteln beschrieben ist eines der Arbeitspakete die Reproduzierbarkeit des Nachweises, hierfür wird in der Vorstudie, die öffentlich zugängliche Codeverwaltung Github verwendet.

LINK:

# Artefakte

In den folgenden Abschnitten werden die Ergebnisse aus der Studie präsentiert. Die heisst es wird einerseits das Modell mit verschiedenen Code-Abschnitten abgebildet und präsentiert. Des Weiteren werden Statistiken über die generierten Daten hier abgebildet.

## Modell

Das Modell wird in Microsofts Visual Studio Code erarbeitet, es werden wie bereits beschrieben zuerst Daten von «Kaggle.com» benutzt, um das Modell zu erarbeiten, welche im späteren Verlauf durch Daten des MGB ersetzt werden. Bei den Daten von «Kaggle.com» handelt es sich um Online-Kundenrückmeldungen zu weiblichen Bekleidungsgegenständen (*Kaggle*, o. J.).

Als Programmiersprache wird Python verwendet da es sich um eine Sprache handelt, welche während des Masterstudiengangs verwendet wurde und in der Praxis eine hohe Verwendung findet.

### Topic Detektion

Für die Topic Detektion wurde der Algorithmus spaCy verwendet, dieser wurde laut spaCy speziell für den Einsatz von Text-Analysen entwickelt, da er grosse Mengen an Text, verarbeiten und verstehen kann (*SpaCy 101*, o. J.). Mit verstehen ist ein maschineller Prozess gemeint, spaCy kategorisiert die Text, Topic, Worte und interpretiert diese während der Analyse.

In der nachfolgenden Darstellung kann das Importieren des spaCy Algorithmus betrachtet werden. Es wird zuerst die sogenannte Library, die Bibliothek, importiert und danach das bereits eingebaute NLP-Verfahren auf den Bewertungstext (Review Text) angewandt.

import spacy

nlp = spacy.load("en\_core\_web\_sm")

reviews\_df["Review Text"] = reviews\_df["Review Text"].apply(nlp)

In der unten abgebildeten zweiten Darstellung wird der Text genauer durchleuchtet und sogenannte Topics erkannt, bei einer Kundenrückmeldung handelt es sich bei einem Topic zum Beispiel um ein T-Shirt welches im Text erwähnt wird.

topics = []

i = 1

for index, row in reviews\_df.iterrows():

    for sentence in row["Review Text"].sents:

        sentenceDocument = nlp(sentence.text)

        descriptiveTerm = ""

        target = ""

        for token in sentenceDocument:

            if token.dep\_ == "nsubj" and token.pos\_ == "NOUN":

                target = token.text

            if token.pos\_ == "ADJ":

                prepend = ""

                for child in token.children:

                    if (child.pos\_ != "ADV"):

                        continue

                    prepend += child.text + " "

                descriptiveTerm = prepend + token.text

        if (target):    # (target and descriptiveTerm)

            topics.append({"topic": target, "description": sentence.text,  "label": row["Rating"], "document": i})  # descriptiveTerm vs. sentence.text

    i = i + 1

### Sentiment Analysis

In diesem Teil werden die verwendeten Sentiment-Analyse Algorithmen aufgezeigt und beschrieben.

#### Textblob

Bei «Textblob» handelt es sich um Algorithmus mit welchem sich Textdaten zum Beispiel klassifizieren, übersetzen lassen oder auch eine Sentiment-Analyse durchgeführt werden kann (*TextBlob*, o. J.).

Im unten dargestellten Code sieht man, dass Textblob auf die erkannten Topics zugreift und sie im Kontext des Textes beurteilt. Dies geschieht bei Textblob auf zwei Arten einerseits die «polarity» und «subjectivity». Bei der Polarität handelt es sich um die Bewertung ob der Text positiv oder negativ ist und bei der Subjektivität bewertet Textblob den Text, ob er objektiv oder eher subjektiv geschrieben wurde.

from textblob import TextBlob

for topic in topics:

    sentiment = TextBlob(topic["description"]).sentiment

    topic["textblob"] = {"polarity": sentiment.polarity, "subjectivity": sentiment.subjectivity}

In der unten angefügten Grafik kann man die Auswertung von Textblob erkennen, es wurde jeweils eine «polarity», eine positive, negative oder neutrale Bewertung der Objektivität der Nutzerbewertung und eine Bewertung der Subjektivität der Nutzerbewertung errechnet, nämlich «subjectivity».

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Abbildung 1 Textblob Sentiment Analysis (Eigene Darstellung)

#### Vader Sentiment Analysis

Bei «Vader Sentiment Analysis» (Vader) findet die Bewertung ähnlich statt, es klassifiziert diese nach positiven, neutralen oder negativen Sentimenten und erstellt darüber hinaus noch ein Gesamtergebnis für den analysierten Text (*Vader Sentiment Analysis*, o. J.).

from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

sia = SentimentIntensityAnalyzer()

for topic in topics:

    topic["sia"] = sia.polarity\_scores(topic["description"])

Im unten angefügten Bild kann man das Ergebnis des Vader Algorithmus erkennen, welcher ähnlich wie Textblob eine positive, negative, neutrale Bewertung der Nutzerbewertung errechnet hat, zusätzlich errechnet Vader noch einen Gesamtbewertung, den SIA Compound.

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Abbildung 2 Vader Sentiment Analysis (Eigene Darstellung)

## Statistiken

Im folgenden Bereich werden Statistiken zu den erkannten Topics aufgezeigt.

### Topic Detektion

Beim momentanen Durchlauf wurden insgesamt 34’341 Topics erkannt, hierbei sei angemerkt, dass es sich nicht um 34’341 verschiedene Topics handelt, sondern um die Gesamtzahl der bisherig erkannten. Es wurden zu diesem Zeitpunkt 1'967 Unikate erkannt.

Von den 34’341 wurden die folgenden positiven, negative und neutralen Bewertungen erkannt, da diese in Form einer Zahl berechnet worden sind bzw. abgegeben worden sind wurden diese in einen Text verwandelt.

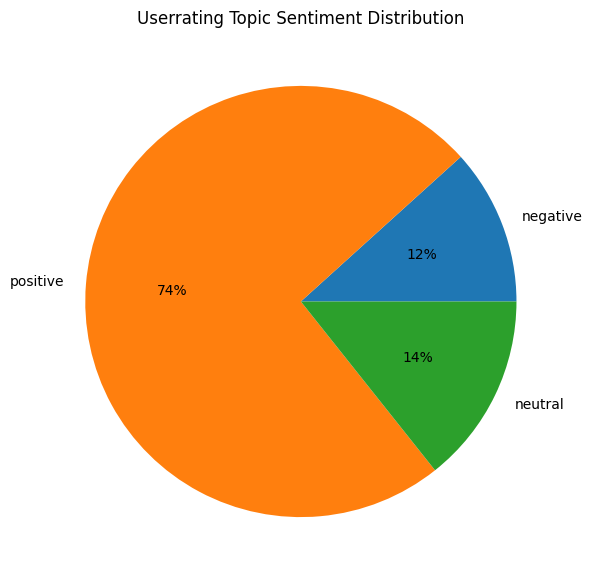
|  |  |  |  |
| --- | --- | --- | --- |
| Bewertungsart | Positiv | Negativ | Neutral |
| Nutzer | 25’408 | 4’914 | 4’019 |
| Vader | 22’083 | 8’246 | 4’012 |
| Textblob | 20’764 | 10’215 | 3’362 |

Tabelle 1 Analyse der Bewertungen (Nutzer / Vader / Textblob) (Eigene Darstellung)

### Sentiment Analysis

Im folgenden Bereich werden Visualisierungen der einzelnen Analysen aufgezeigt.

#### Kreisdiagramme

Auf der linken Seite kann man die oben in Zahlen beschriebene Statistik anhand einen Kreisdiagrammes betrachten. Gemäss der Statistik haben 74% der Nutzer eine Positive, 34% eine neutrale und 12% eine negative Bewertung abgegeben.

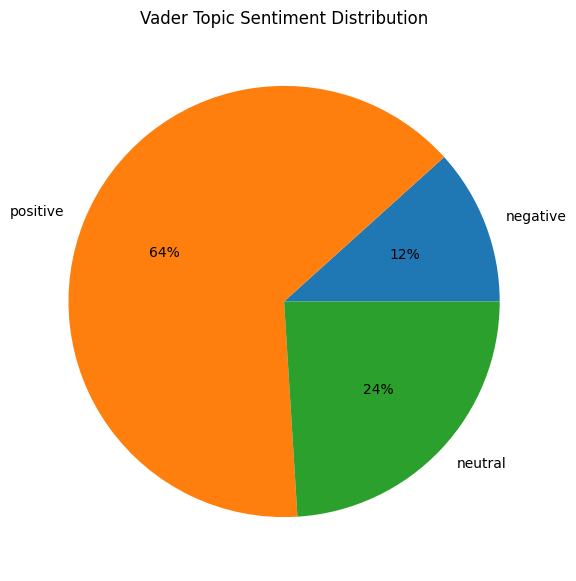


Abbildung 3: Nutzerbewertung (Eigene Darstellung)

Auf der rechten Seite kann man die Auswertung des Vader-Algorithmus erkennen. Die Auswertung ergab 64% positive, 24 neutrale und 12% negative Bewertungen.

Abbildung 4: Vaderbewertung (Eigene Darstellung)

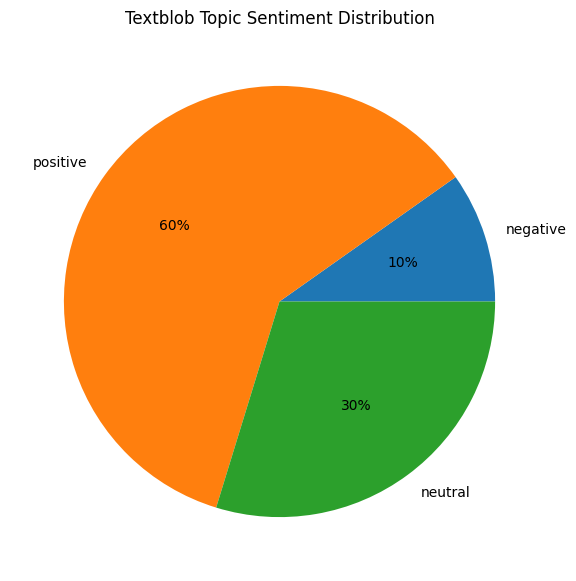
Das letzte Kreisdiagramm zeigt die Auswertung des Textblob-Algorithmus welcher 60% positive, 30% neutrale und 10% negative errechnet hat.

Abbildung 5: Textblobbewertung (Eigene Darstellung)

#### Balkendiagramme

Die unten dargestellten drei Balkendiagramme zeigen die Verteilung positiver, negativer oder neutraler Bewertungen und zeigen das gleiche Bild wie bei den Kreisdiagrammen in anderer Darstellungsform.

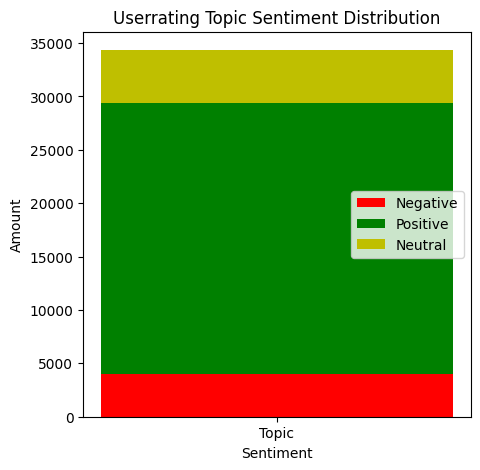
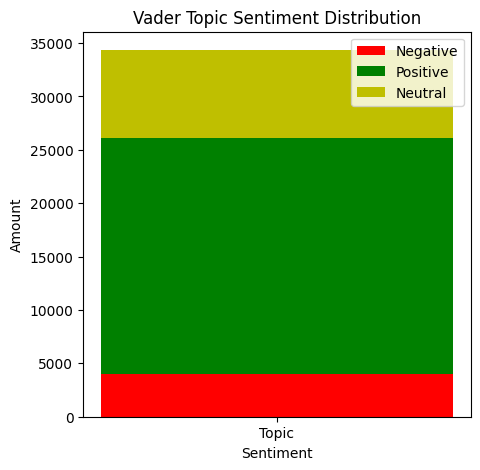


Abbildung 6: Nutzerbewertung (Eigene Darstellung) Abbildung 7: Vaderbewertung (Eigene Darstellung)

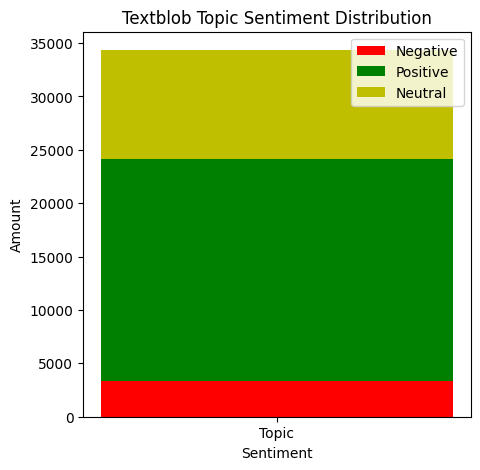


Abbildung 8: Textblobbewertung (Eigene Darstellung)

### Durchschnittswerte der Statistiken

Eine weitere Erkenntnis dieser Vorstudie sind die erkannten Topics im Durschnitt pro Review, dies wurde mittels folgendermassen erreicht:

Die Daten wurden nach einmaliger Analyse in einer SQL-Datenbank gespeichert, mit dieser wurde weitergearbeitet und die Statistiken im unteren Bereich erstellt.

topicquery = '''SELECT

        Description,

        Document,

        (SIA\_COMPOUND + Textblob\_Polarity)/2 as sentiment,

        COUNT(Topic)

        FROM review

        GROUP BY

        Description'''

Es wurde mittels «Select-Statements» die relevanten Bereiche, Beschrieb (Description), DokumentID (Document), Durchschnittsberechnung der Sentimente (SIA\_Compound + Textblob\_Polarity) und die Anzahl Topics ausgewählt. Diese wurden mittels «Group-By Statement» den Beschrieben zugewiesen.

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Abbildung 9 SQL-Select-Statement (Eigene Darstellung)

#### Durchschnitt Topics pro Beschrieb & Dokument

Im Anschluss wurde eine Berechnung des Durchschnitts via SQL-Befehl durchgeführt, welche die durchschnittliche Anzahl Topics pro Beschrieb wiedergibt. Dies ergab einen Durchschnittswert von **1.012173** erkannten Topics pro Beschrieb. Wenn nach DokumentID gruppiert wird ergibt sich ein Durchschnittswert gefundener Topics pro Dokument von **1.952747**.

#### Durchschnitt Sentimente

Dies ergab einen Durchschnittswert von **1.012173** erkannten Sentimente pro Beschrieb. Wenn nach DokumentID gruppiert wird ergibt sich ein Durchschnittswert gefundener Sentimente pro Dokument von **1.952747**. Die Sentimente pro Topic werden in Kapitel 7.2.4 nochmals genauer betrachtet.

### Sentimente pro Topic

Der folgende SQL-Befehl gibt zeigt die Topics gruppiert nach gefundenen Sentimenten. Da bei der im Kapitel 7.1.1 auch Topics enthalten sein können, welche keine Sentimente aufweisen.

topicsentimentquery = '''SELECT

        (SIA\_COMPOUND + Textblob\_Polarity)/2 as sentiment,

        Topic

        From review

        GROUP BY

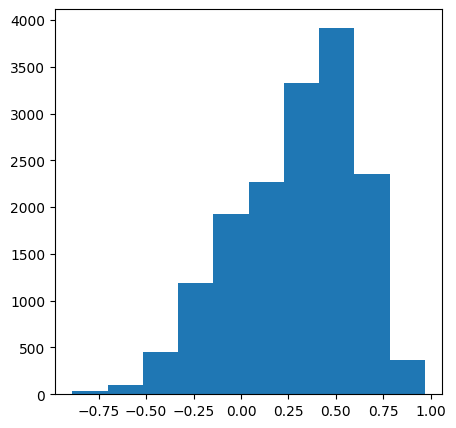
        sentiment'''

Nachdem via den oberen SQL-Befehlen die Topics den Sentimenten zugewiesen wurde, ergab sich eine Gesamtanzahl gefundener Topics von 15'936.

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

Abbildung 10 Anzahl Sentimente welche ein Topic aufweisen und deren Verteilung (Eigene Darstellung)

In der oben abgebildeten Abbildung kann man den Durchschnittswert der Sentimente erkennen, dieser liegt bei **0.290079**, dies bedeutet das mehrheitlich positive Beschriebe und somit Topics erkannt wurden, was sich auf mit den in Kapitel 7.2.2.1 abgebildeten Kreisdiagrammen deckt. Des Weiteren kann man festhalten dass das negativste Sentiment einen Wert von **-0.887883** aufweist und das positivste **0.968500**. Die Gesamtverteilung kann man im nachfolgenden Histogramm erkennen.

In der links abgebildeten Grafik kann man die Verteilung der Sentimente erkennen, des Weiteren kann man erkennen, dass die Verteilung stark positiv ist.

Abbildung 11 Histogramm Verteilung Sentimente (Eigene Darstellung)

# Fazit

Die Vorstudie zeigt auf wie ein NLP-Prozess zur Analyse von Kundendaten aussehen kann, und hat darüber hinaus aufgrund der angewendeten Algorithmen und weiterer Analyseschritte einige Resultate hervorgebracht. Da die MGB-Daten, stand 06.12.2022, noch nicht verfügbar waren, sind Rückschlüsse auf die Verlässlichkeit im Vergleich zu den im Unternehmen angewendeten Prozessen nicht möglich. Jedoch besteht hier die Hoffnung das die Daten im Verlaufe der Masterarbeit zur Verfügung stehen, um Vergleiche und Rückschlüsse ziehen zu können.

Die Vorstudie zeigt jedoch auf, dass eine Topic Detektion mittels spaCy und darauffolgender Sentiment-Analyse mittels Textblob und Vader zu interessanten Auswertungen führt, welche den Gebrauch für weiterführende Analyseschritte rechtfertigen.

.

# Literaturverzeichnis

*spaCy 101: Everything you need to know · spaCy Usage Documentation*. (o. J.). SpaCy 101: Everything You Need to Know. Abgerufen 6. Dezember 2022, von https://spacy.io/usage/spacy-101

*TextBlob: Simplified Text Processing — TextBlob 0.16.0 documentation*. (o. J.). Abgerufen 6. Dezember 2022, von https://textblob.readthedocs.io/en/dev/

*Welcome to VaderSentiment’s documentation! —VaderSentiment 3.3.1 documentation*. (o. J.). Abgerufen 6. Dezember 2022, von https://vadersentiment.readthedocs.io/en/latest/

*Women’s E-Commerce Clothing Reviews*. (o. J.). Abgerufen 6. Dezember 2022, von https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews

# Abbildungsverzeichnis

[Abbildung 1 Textblob Sentiment Analysis (Eigene Darstellung) 10](#_Toc121231231)

[Abbildung 2 Vader Sentiment Analysis (Eigene Darstellung) 10](#_Toc121231232)

[Abbildung 3: Nutzerbewertung (Eigene Darstellung) 11](https://d.docs.live.net/600a89876a86c2d1/Desktop/ZHAW/Masterarbeit/Vorstudie_Masterarbeit.docx#_Toc121231233)

[Abbildung 4: Vaderbewertung (Eigene Darstellung) 11](https://d.docs.live.net/600a89876a86c2d1/Desktop/ZHAW/Masterarbeit/Vorstudie_Masterarbeit.docx#_Toc121231234)

[Abbildung 5: Textblobbewertung (Eigene Darstellung) 12](https://d.docs.live.net/600a89876a86c2d1/Desktop/ZHAW/Masterarbeit/Vorstudie_Masterarbeit.docx#_Toc121231235)

[Abbildung 6: Nutzerbewertung (Eigene Darstellung) Abbildung 7: Vaderbewertung (Eigene Darstellung) 12](#_Toc121231236)

[Abbildung 8: Textblobbewertung (Eigene Darstellung) 12](#_Toc121231237)

[Abbildung 9 SQL-Select-Statement (Eigene Darstellung) 13](#_Toc121231238)

[Abbildung 10 Anzahl Sentimente welche ein Topic aufweisen und deren Verteilung (Eigene Darstellung) 14](#_Toc121231239)

[Abbildung 11 Histogramm Verteilung Sentimente (Eigene Darstellung) 14](https://d.docs.live.net/600a89876a86c2d1/Desktop/ZHAW/Masterarbeit/Vorstudie_Masterarbeit.docx#_Toc121231240)

# Tabellenverzeichnis

[Tabelle 1 Analyse der Bewertungen (Nutzer / Vader / Textblob) (Eigene Darstellung) 11](#_Toc121231224)