

Evaluation und Interpretation Sentiment Analyse

Erarbeitet von
Dr. Katarina Boland

Lernziele	1
Inhalt	2
Einstieg	2
Bedeutung von Sentiment-Labels	2
Sentimente zur Meinungsmessung	5
Abschluss	7
Quellen	8
Weiterführendes Material	9
Disclaimer	9

Lernziele

- Du kannst gebräuchliche Metriken zur Messung der Qualität von (ungerichteter) automatischer Sentiment Analyse aufzählen
- Du kannst bewerten, welche Aussagen auf Grundlage gefundener Sentimente zulässig sind

Inhalt

Einstieg

Sentiment Analyse ist eine Klassifikationsaufgabe.

Quelle [1]

Dementsprechend finden die klassischen Evaluationsmaße hier Anwendung: insbesondere Accuracy, Precision, Recall und die Kombination aus den letzteren beiden, das F-Measure. Auch Wahrheitsmatrizen bieten sich für die Evaluation an.

Wir wollen uns an dieser Stelle nicht auf die Berechnung dieser Maße konzentrieren, sondern gehen davon aus, dass du mit ihrer Hilfe eine gute Qualität deiner automatisch erzeugten Sentiment-Labels sichergestellt hast.

Die Frage ist jetzt: Was genau sagen dir diese Labels denn nun?

Bedeutung von Sentiment-Labels

Zunächst einmal stellt sich die Frage: Was genau bedeutet ein negatives, neutrales oder positives Sentiment bezogen auf den konkreten Datensatz?

Diese Frage mag trivial klingen, aber bei genauerer Betrachtung kann man schon hier unterschiedliche Definitionen finden.

Schauen wir uns beispielsweise diese Tweets an. Wie würdest du diese bewerten?

Ist das Gefühl, das der Tweet vermittelt, insgesamt eher positiv, negativ oder weder noch?



AstraZeneca ist raus, weil eine Handvoll Leute ne Thrombose bekommen haben? Darf ich den Beipackzettel der Pille noch einmal sehen?

Das Sentiment sollte hier als negativ bewertet werden, da die ausgedrückte Gesamtstimmung negativ ist. Die Tatsache, dass gleichzeitig eine positive Haltung gegenüber AstraZeneca ausgedrückt wird, ändert daran nichts. Es ist keine positiv gestimmte Nachricht.

Was aber ist mit diesem Tweet hier?

Ist das Gefühl, das der Tweet vermittelt, insgesamt eher positiv, negativ oder weder noch?



Eilmeldung : #Italien stoppt vorerst sämtliche #Impfungen mit #AstraZeneca (nach #Dänemark)

Hier könnte man sagen, dass einfach eine Schlagzeile wiedergegeben wird, es sich also um neutrale Berichterstattung handelt und der Tweet somit neutral ist. Ebenso kann man aber argumentieren, dass die übermittelte Nachricht eine negative ist, wenn man weiß, dass die Schlagzeile in Kurzform kommuniziert, dass Probleme aufgetreten sind, die nun als zu hohes Risiko eingestuft werden.

Somit wäre dies eine negative behaftete Nachricht (auch wenn man selbst die Maßnahme des Impfstopps als solche ggf. positiv bewertet. Dies betrifft dann die eigene persönliche Einstellung und nicht das, was im Tweet ausgedrückt wird, und das gilt es ja zu klassifizieren).

An dieser Stelle muss man also bei der Annotation sicherstellen, dass die Annotator*innen präzise Anweisungen bekommen:

Ist das Gefühl, das der Tweet vermittelt, insgesamt eher positiv, negativ oder weder noch

*bzw. ist die übermittelte Nachricht oder Schlagzeile positiv, negativ oder gemischt / neutral?
(Hintergrundwissen darf verwendet werden)*

In diesem Fall würde man eine Annotation des Beispiels mit „negativ“ erwirken. Ob das gewünscht ist, oder ob in so einem Fall „neutral“ die bessere Wahl ist, hängt vom Untersuchungsgegenstand ab:

Möchte man z. B. wissen, in welchem Ausmaß positive vs. negative Berichterstattung bzgl. Corona-Impfungen auf Twitter stattfindet, oder möchte man wissen, inwiefern auf Twitter sachlich vs. meinungsbehaftet diskutiert wird?

Zur Interpretation der Labels muss also die Definition bekannt sein, die bei Entwicklung bzw. Training des Klassifikationsverfahrens gewählt wurde und die somit auch den resultierenden Labels zugrunde liegt.

Dasselbe betrifft die zugrundeliegende Skala: Bedeutet die Kategorie „neutral“, dass weder ein positives noch ein negatives Sentiment gefunden wurde? Oder bedeutet sie, dass positives und negatives Sentiment gleich stark sind?

Gibt es getrennte Angaben zur Stärke von negativem und positivem Sentiment oder werden beide Werte aufaddiert und das überwiegende Sentiment als Label verwendet?

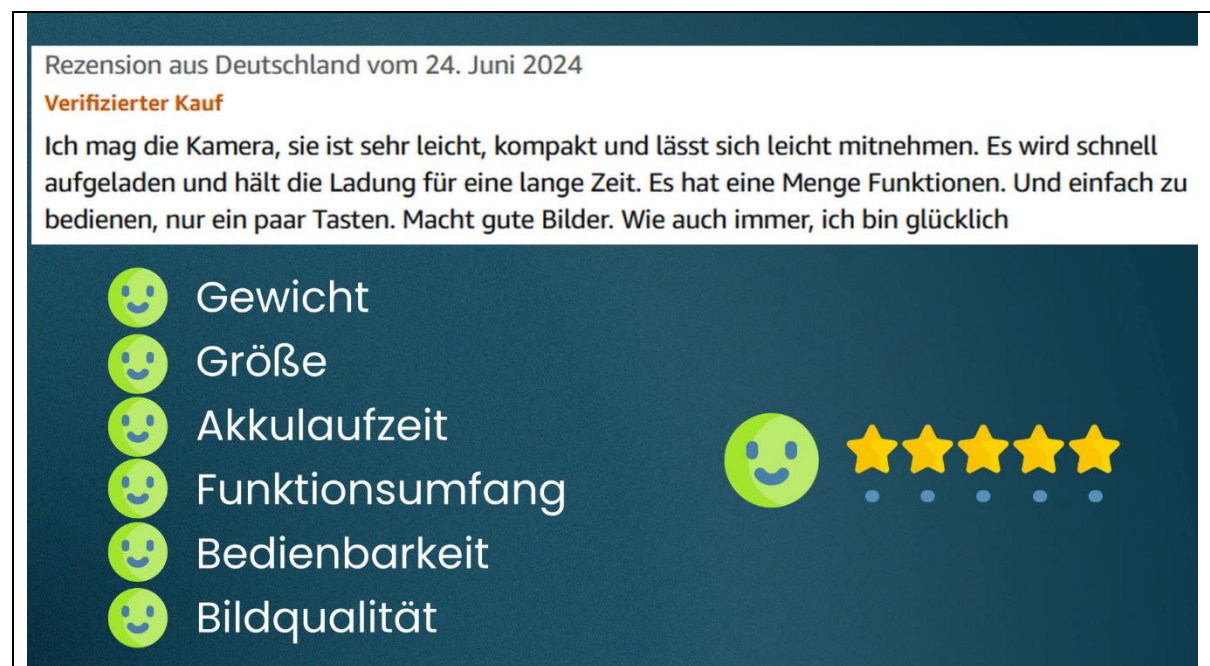
Auch hier musst du prüfen, wie die Klassen definiert sind um die Labels korrekt interpretieren zu können.

Sentimente zur Meinungsmessung

Sentiment Analyse wurde unter anderem eingesetzt, um Bewertungen von Produkten aus Rezensionstexten abzuleiten.

Quelle [1]

Die Idee ist hier, dass wenn die Stimmung im Review eines Produkts positiv ist, also viele positive Aspekte genannt werden, vermutlich auch die Meinung des Autors oder der Autorin zu diesem Produkt insgesamt positiv ist. Nennt er oder sie viel Negatives, so ist die Meinung vermutlich negativ.



Sentiment Analyse kann so also zur Messung von Meinungen eingesetzt werden, im speziellen Fall, um Kaufempfehlungen abzuleiten.

Aber Vorsicht: Derartige Rückschlüsse von Sentimenten auf Meinungen sind nicht immer sinnvoll.

Angenommen, du möchtest zum Beispiel Diskurse in sozialen Medien analysieren. Wenn du jetzt Tweets zum Thema Lockdown mit negativem Sentiment findest, muss das nicht zwingend heißen, dass die entsprechenden Autor*innen der Maßnahme der Lockdowns negativ gegenüberstehen. Denn die Themen Lockdowns und Corona sind an sich mit vielen negativen Assoziationen behaftet und auch Befürworter*innen von Lockdowns müssen nicht euphorisch über dieses Thema schreiben.



gut zusammengefasst "Ein konsequenter Lockdown würde natürlich kurz wehtun, ähnlich wie ein Nadelstich beim Impfen. Aber was wir derzeit machen ist wie der Versuch, sich mit einem Löffel ein Loch in den Arm zu bohren."

Ein beliebiger Text über etwas ist nicht einer Rezension gleichzusetzen.

Zudem kann auch mit negativen Sentimenten Zustimmung ausgedrückt werden und umgekehrt, wenn man Sentiment nicht zielgerichtet betrachtet. Beispiele siehst du hier.

Quelle [2]

	Sentiment	Sentiment (Ziel: Trump)	Stance (Ziel: Trump)
<i>Trump ist der einzige Kandidat, den ich voll und ganz unterstütze.</i>			
<i>Die verlogenen Medien, die Trump angreifen, machen mich wütend.</i>			
<i>Jeb Bush ist der einzige verünftige Kandidat der Republikaner.</i>		—	

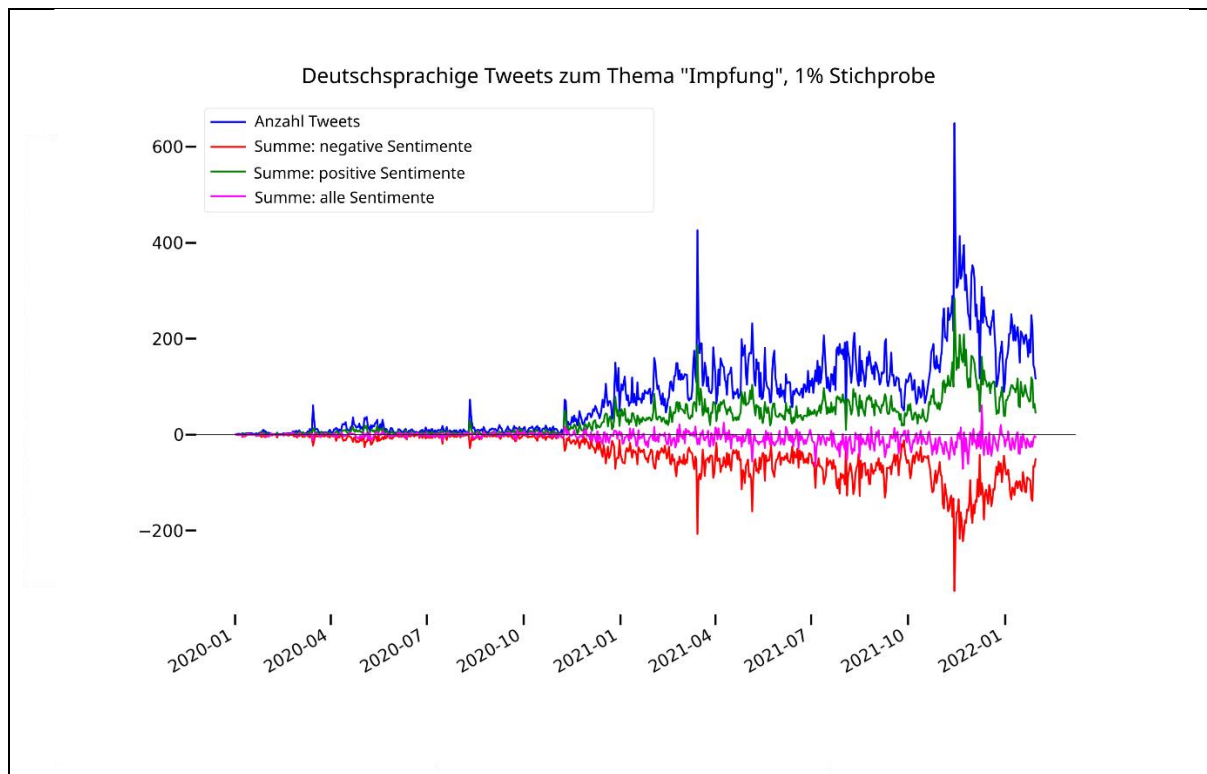
Sentiment, gerichtet und ungerichtet, vs. Stance. Adaptiert aus Quelle [2]

Sentimente können also unter Umständen zur Messung von Meinungen eingesetzt werden, aber nicht immer auf direkte Weise. Sentimente sind nicht automatisch ein Signal für Zustimmung oder Ablehnung und spiegeln auch nicht direkt positive oder negative Bewertungen wider.

Sie sind vielmehr als Stimmungen zu sehen, die mit bestimmten Themen zu einem gegebenen Zeitpunkt assoziiert sind. Somit können Trends und Änderungen in der öffentlichen Stimmung analysiert werden.

Diese Assoziationen können wiederum sehr interessante Informationen liefern. In diesem Bild siehst du die Gesamtanzahl pro Tweets pro Tag zum Thema Impfungen im Kontext von Corona (blau eingefärbt).

Quelle [3]



Die Sentimente sind zusätzlich dargestellt.

Sie wurden mit dem SentiStrength-Tool gemessen, was getrennte Werte für positives und für negatives Sentiment pro Tweet ausgibt.

Quelle [4]

Die aufsummierten positiven Sentimente aller Tweets an einem Tag sind in Grün dargestellt, aufsummierte negative Sentimente in Rot, die Summe aus positiven und negativen Sentimenten in Pink.

In demselben Zeitraum, als große Debatten über die Sicherheit des AstraZeneca-Wirkstoffs stattfanden und ein Impfstopp mit diesem in Deutschland, konnten eine höhere Anzahl an Tweets und stärkere negative Sentimente verzeichnet werden.

Ob es sich hier um kausale Zusammenhänge handelt, muss natürlich getrennt betrachtet werden.

Abschluss

Je nach konkretem Anwendungsfall können Sentimente unterschiedliche Dinge aussagen. Die Definition der Labels in verschiedenen Datensätzen und Klassifikationsverfahren können

voneinander abweichen und auch die Eigenschaften der untersuchten Texte spielen eine Rolle.

Es muss also immer hinterfragt werden, was genau wie gemessen wurde: Es wäre irreführend zu sagen, dass Sentimente generell Zustimmung oder Ablehnung signalisieren oder sie als allgemeines Stimmungsbild fungieren. In bestimmten Fällen kann dies allerdings zutreffen.

Ob Sentiment Analyse für deinen Anwendungsfall die erhofften Informationen liefert, musst du also im Vorhinein genau überlegen, noch bevor du ein Tool auswählst und anwendest.

Quellen

- Quelle [1] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Springer International Publishing. <https://doi.org/10.1007/978-3-031-02145-9>
- Quelle [2] Indira Sen, Fabian Flöck, and Claudia Wagner. 2020. On the Reliability and Validity of Detecting Approval of Political Actors in Tweets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1413–1426, Online. Association for Computational Linguistics.
- Quelle [3] Boland, K., Starke, C., Bensmann, F., Marcinkowski, F., & Dietze, S. (2024). A computational analysis of German online discourses about COVID-19 vaccinations to inform policy-making in times of crisis (Preprint). <https://doi.org/10.2196/preprints.63909>
- Quelle [4] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. J. Am. Soc. Inf. Sci. Technol., 61(12), 2544–2558.

Weiterführendes Material

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems, 226, 107134. <https://doi.org/10.1016/j.knosys.2021.107134>

Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. In S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, & D. Jurgens (Hrsg.), Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (S. 502–518). Association for Computational Linguistics. <https://doi.org/10.18653/v1/S17-2088>

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica, 22(3), 276–282. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052>

Disclaimer

Transkript zu dem Video „06 Clustering: vom Sortieren bis zum Explorieren: Evaluation und Interpretation Sentiment Analyse“, Katarina Boland.

Dieses Transkript wurde im Rahmen des Projekts ai4all des Heine Center for Artificial Intelligence and Data Science (HeiCAD) an der Heinrich-Heine-Universität Düsseldorf unter der Creative Commons Lizenz [CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/) veröffentlicht. Ausgenommen von der Lizenz sind die verwendeten Logos, alle in den Quellen ausgewiesenen Fremdmaterialien sowie alle als Quellen gekennzeichneten Elemente.