Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Medieninformatik

Clickbait Detection with Machine Learning

Bachelorarbeit

Sebastian Köpsel
geb. am: 01.01.1990 in Neubrandenburg

Matrikelnummer 100146

1. Gutachter: Prof. Dr. Matthias Hagen
2. Gutachter: Prof. Dr. Benno Stein

Datum der Abgabe: 31. Februar 2022

Erklärung

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

Weimar, 31. Februar 2022

.....
Sebastian Köpsel

Zusammenfassung

We tried to detect clickbait with machine learning on a self-build corpus. We did ok.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die Medienlandschaft hat sich durch das das Internet stark verändert. Die früher dominierenden Printmedien mussten und müssen mit sinkenden Einnahmen leben und so wird es immer wichtige eine starke und aktuelle Onlinepräsenz zu haben. Damit wird Werbung als Einnahmequelle immer wichtiger. Da Werbung inzwischen einer der wichtigsten Einnahmequellen für viele Internetpublisher ist, sind Besucherzahlen entscheidend. In den vergangenen Jahren hat sich eine Form von Content entwickelt(oder wurde vielmehr wiederentdeckt) die gemeinhin als Clickbait bezeichnet wird. Bei diesen wird durch möglichst reißerische, emotionale und offen gestaltete Überschriften versucht, den Nutzer von Plattformen von Twitter, Facebook und Ähnlichem zum Besuchen zu "überreden". Durch die Verbreitung in sozialen Medien, und der durch viele empfundene Zwang, Artikel mit dieser Form von Überschrift zu lesen, hat Clickbait viele Feinde gemacht, doch die Effektivität dieser Methoden steht außer Frage. Im Rahmen dieser Bachelor Arbeit wurden aus Mangel an Fach-Literatur aber auch um den ... darzustellen einige hundert Clickbait verwandte Blog Einträge ausgewertet und die Meinung der Blogger ist klar: Clickbait ist irreführend, sensationalistisch und enttäuschend.

Um Clickbait zu beschreiben wurden in [1] die Begriffe „Message“ und „Target“ eingeführt. Dabei beschreibt die Message den Teil, der auf den Hauptcontent, das Target, verweist, im Falle Twitter der entsprechende Tweet –Bild—, er hat den Auftrag den Leser zum Click auf den Link zu bewegen.

Daraus ergibt sich das Thema dieser Arbeit: wie gut können wir Clickbait mit Hilfe von Machine Learning erkennen? Denkbare Funktionen dieser Erkennung wäre ein Einsatz in Form eines Spamfilters, der den jeweiligen NewsFeed auf relevante Bereiche und Nachrichten beschränkt und Menschen hilft sich nicht ablenken zu lassen.

Auf Clickbait, seine Formen und Ausprägungen wird genauer in Kapitel 3 eingegangen.

Um Clickbait erkennen zu können wird ein annotierter Korpus benötigt, mit welchen Klassifikatoren trainiert werden können. Da kein relevanter Korpus gefunden wurde, ergab sich die Notwenigkeit diesen selbst zu sammeln und zu annotieren. Dazu haben wir 3000 Tweets der 25 größten Newspublisher auf Twitter gesammelt und anschließend annotiert. Auf die Corpuserstellung und Annotation wird in Kapitel 4 genau eingegangen. Auf dem nun erstellten Korpus konnten nun das Machine-Learning ausgeführt werden. Dazu wurden neben den herkömmlichen Bag-of-Word Features eine vielzahl ein weiterer Features Verwendet, welche sich grob in die Kategorien

- Message-Features: Hier wird die Message selbst in einer Reihe von Features dargestellt, unter anderem als Bag-of-Words und die Flesch-Reading-Ease
- Target-Features: Hier wird das Target der Message dargestellt, mit ähnlichen Features wie die Message selbst
- Meta-Features: Hierbei werden Metainformationen der Message betrachtet, zum Beispiel der User, welcher die Message verfasst hat

unterteilen lassen. Die Details werden in Kapitel 5 beschrieben. Im Folgenden Kapitel 6 wird auf die Ergebnisse dieser Experimente eingegangen. Um den Ergebnisse zu visualisieren und praktisch testen zu können wurde ein Webservice erstellt, bei dem man die Klassifizierung der trainierten Klassifizierer an aktuellen Tweets betrachten kann. Der Aufbau und die Funktion dieses Webservices werden in Kapitel **??** beschrieben. Anschließend werden in Kapitel **??** Ergebnisse noch einmal zusammengefasst, und ein Ausblick auf zukünftige mögliche Aufgaben der Forschung gegeben.

Überschrift muss aus großer auswahl hervorstechen

Kapitel 2

Verwandte Arbeiten

Soft News und Clickbait in News. In „Doing Well and Doing Good: How Soft News and Critical Journalism Are Shrinking the News Audience and Weakening Democracy– And What News Outlets Can Do About It“ zeigen die Autoren schon 2000 eine zunehmende Abwendung der News Publisher von Hard News in Richtung soft news. Da hier der Begriff Clickbait noch nicht klar definiert wurde, soll hier die Definition aus [2] gegeben werden, welche auf dieser Veröffentlichung aufbaut.

> [Soft news is] “typically more personality-centered, less time-bound, more practical, and more incident-based than other news” and associated with “sensationalism”, “human-interest” or “news you can use.”

Dabei lässt sich Soft News als Obergruppe von Clickbait verstehen, da es die Merkmale Sensationsmacherei, Menschliches Interesse und News die man gebrauchen kann teilt, jedoch spezifischer ist in der Hinsicht das Clickbait sich primär durch den Überschriftenstil auszeichnet. Soft News zeigt sich in zu zunehmender Sensationsmacherei, Aufmerksamkeit für „Human Interest“, Kriminalität und Katastrophen widerspiegelt. Auch zeigen sie einen Anstieg von Selbstreferenz in News, ein für Clickbait typisches Mittel. Linguistische Forschung über Clickbait. "Click bait: Forward-reference as lure in online news headlines"[2] beschäftigt sich detailiertmit der "clickbaitisierung"dänischer News. Bei der durchgeführten Studie wurden 100.000 Überschriften 10 verchiedener dänischer News Websiten auf untersucht, untersucht wurde hierbei insbesondere der vorwärts referenzierende Teil von Clickbait. Es konnten in dem Datensatz 17,2% der Nachrichten als Clickbait identifiziert werden, wobei der größte Anteil von Clickbait den sogenannten ßoft content"Kategorien Sport, Wetter, „Lifestyle“ und „Gadget“ gefunden wurde. Es ließ sich zudem ein Trend zum Boulevardzeitungsstil und Kommerzialisierung feststellen. Die Autoren

verwendeten zur erkennung von Clickbait hauptsächlich die dänischen Wörter für „here“, „this“ und „we“ [1]. Damit haben die Autoren die Grundlage für eine automatische Erkennung von Clickbait gebildet, sie erkennen jedoch auch das ihre Ergebnisse nicht völlig umfassend sind, sondern eher als Teil einer Tendenzbewertung verstanden werten sollten. Wir stimmen mit der Bewertung, das vorwärtsreferenzierende Überschriften vorrangig das Ziel haben, Leser zum clicken des Artikels zu ködern um Werbeeinnahmen zu erzielen überein.

Psychologische Funktionsweise Clickbaits. [3] untersuchte Faktoren die zur viralität von Online Content beitragen und kam zum Schluss das Besorgnis- und Wut erregende Geschichten häufiger in Mailinglisten geteilt wurden. Das emotionale Geschichten und Überschriften mehr aufmerktsamkeit Erfahren und dementsprechen höhere viralität und einen höheren köderfaktor bei clickbait haben konnte auch in [4] gezeigt werden. In journalistischen Artikeln [5]) wird häufiger die „Curiosity Gap“ genannt. Dieser Ausdruck bezieht sich auf die Arbeit Loewensteins, welcher in [6] geprägt wurde. Demnach empfinden Menschen oftmals das verlangen eine entstehende Wissenslücke zu schliessen falls sie auf eine solche Treffen, im falle Clickbaits auf das Thema nicht vollständig erklärende, vorwärts referenzierende Überschriften. Das öffentliche Angelegenheiten bei Lesern deutlich unbeliebter sind als Kriminalität und enternainmeint konnte in [7] dargestellt werden. Listicles, also Artikel die als Listen aufgebaut sind, sind eine Ausprägung von Clickbait. In dem New Yorker Artikel „A List of Reasons Why Our Brains Love Lists “ argumentiert Konnikova mithilfe von [8] das sich diese Form von Wissensaufbereitung besonders eignet

Paper lesen!

MORE

[1] her, sådan, derfor, så and dette, denne, dette

Kapitel 3

Clickbait

3.1 Definition Clickbait

3.2 Motivation

3.3 Problematik

3.4 Meinungen

Kapitel 4

Korpuserstellung

4.1 Ansätze

4.2 Sampling Process

4.3 Annotationsprozess

4.4 Analyse

Clickbait pro Publisher

Listicles

Besonderheiten

Probleme

Kapitel 5

Maschine Learning Experimente

5.1 Preprocessing

5.2 Attribute Selection

5.3 Classifier

5.4 Experimente

Kapitel 6

Auswertung der Ergebnisse

6.1 Vergleich und Auswertung

6.2 Erklärung

Kapitel 7

Erstellung eines Webservices

7.1 Funktion

7.2 Aufbau

7.3 Bewertung

Kapitel 8

Zusammenfassung und Ausblick

Literaturverzeichnis

[1] Dr. M. Potthast Prof. Dr. M. Hagen and S. Köpsel. Clickbait detection.

[2] Jonas Nygaard Blom and Kenneth Reinecke Hansen. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100, 2015.

[3] Jonah Berger and Katherine L Milkman. What makes online content viral? *Journal of marketing research*, 49(2):192–205, 2012.

[4] Marco Guerini and Jacopo Staiano. Deep feelings: A massive cross-lingual study on the relation between emotions and virality. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 299–305. International World Wide Web Conferences Steering Committee, 2015.

[5] Guarino Ben. The 5 psychological reasons you can't resist clickbait, 06, year = 2015, url = https://www.inverse.com/article/4069-the-5-psychological-reasons-you-can-t-resist-clickbait urldate = 2016-01-06.

[6] George Loewenstein. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75, 1994.

[7] Elena Hensinger, Ilias Flaounas, and Nello Cristianini. Modelling and explaining online news preferences. In *Pattern Recognition-Applications and Methods*, pages 65–77. Springer, 2013.

[8] Claude Messner and Michaela Wänke. Unconscious information processing reduces information overload and increases product satisfaction. *Journal of Consumer Psychology*, 21(1):9–13, 2011.