Informationslinguistik

Melanie Siegel*

Opinion Spam – Meinungsäußerungen als Fake

DOI 10.1515/iwp-2016-0058

Zusammenfassung: Die Meinung der Kunden ist ein wichtiger Geschäftsfaktor geworden. Firmen wollen wissen, was Verbraucher von ihrem Produkt oder ihrer Dienstleistung halten. Sie versuchen daher, sich und ihre Produkte schnell an Kundenbedürfnisse anzupassen und die geäußerten Meinungen als Marketinginstrument einzusetzen. Mit der Zunahme der Relevanz der Kundenmeinungen steigt jedoch auch die Anzahl der Manipulationsversuche. Wir haben es hier mit einem gesellschaftlich und ökonomisch erheblichen Problem zu tun. Der Artikel stellt Forschungsansätze für die englische und chinesische Sprache vor und untersucht die Übertragbarkeit auf die deutsche Sprache. Als Ausgangspunkt für diese Untersuchung wird zunächst der Aufbau eines Korpus für gefälschte Bewertungen im deutschsprachigen Amazon-Portal umrissen. Erste Analysen dieses Korpus zeigen, dass die vorgestellten Forschungsansätze zum großen Teil auch auf die deutsche Sprache übertragbar sind.

Deskriptoren: Textanalyse, Datenanalyse, Produktinformation, Informationswert, Zuverlässigkeit, Bewertung, Chinesisch, Englisch

Opinion Spam – Fake expressions of opinion

Abstract: Customer opinions have become an important business factor. Companies want to know what consumers think of their product and their service trying to quickly adapt themselves and their products to customer needs. They deploy the customer opinions as a marketing tool. Though, with the increase of the relevance of customer opinions, also the number of manipulation attempts increases. This has turned out to be a socially and economically significant problem. This article presents research approaches concerning the English and Chinese languages and examines the transferability of the approaches to the German language. The construction of a corpus for fake reviews in the German Amazon site is outlined. Initial analyses on this corpus show that the methods for English and Chinese are to a large extent transferable to the German language.

Descriptors: Text analysis, Data analysis, Product information, Information value, Reliability, Evaluation, Chinese, English

Opinion Spam – des fausses expressions d'opinion

Résumé : Les opinions des clients sont devenues des facteurs commerciaux importants. Les entreprises veulent savoir ce que les consommateurs pensent de leurs produits ou services. Ils essaient donc eux-mêmes de s'adapter rapidement et de faire en sorte que leurs produits répondent aux besoins des clients. D'autre part, ils utilisent les opinions exprimées comme outil de marketing. En même temps que s'améliore la pertinence des opinions des clients, le nombre de tentatives de manipulation croît également. Nous avons affaire à un problème social et économique important. L'article présente des approches de recherche pour les langues anglaise et chinoise et examine la possibilité de les transposer à la langue allemande. Comme point de départ de cette étude, l'auteur décrit la construction d'un corpus de faux avis sur le site allemand d'Amazon. Des analyses initiales sur ce corpus montrent que les approches de recherche présentés sont transposables dans une large mesure à la langue allemande.

Descripteurs: Analyse de texte, Analyse des données, Information sur des produits, Valeur de l'information, Fiabilité, Evaluation, Chinois, Anglais

Einleitung

Die Meinung der Kunden ist ein wichtiger Geschäftsfaktor geworden. Firmen wollen wissen, was Verbraucher von ihrem Produkt oder ihrer Dienstleistung halten, und versuchen, sich und ihre Produkte schnell an Kundenbedürfnisse anzupassen und die geäußerten Meinungen bestenfalls als Marketinginstrument einzusetzen. In den USA wird man mittlerweile sogar nach Benutzung einer Toilette

***Kontaktperson: Prof. Dr. Melanie Siegel,** h_da, Hochschule Darmstadt, Fachbereich Media, Medien-Campus der h_da, Max-Planck-Straße 2, 64807 Dieburg, E-Mail: melanie.siegel@h-da.de

gebeten, auf einen lachenden oder weinenden Smiley zu drücken, je nachdem, wie die Erfahrung war. Für Firmen kann es existenziell sein, vor einem aufkommenden „Shitstorm" rechtzeitig gewarnt zu werden. Gleichzeitig kann nur derjenige schnell auf Trends reagieren, der diese auch rechtzeitig erkennt.

Dies hat auch für die Kunden Vorteile. Verbraucher informieren sich gegenseitig, und fast niemand bucht heutzutage eine Urlaubsreise, ohne sich vorher die Bewertungen anderer Gäste angesehen zu haben. Verbraucher wollen die Meinung anderer Verbraucher wissen, bevor sie ein Produkt kaufen, ein Hotel buchen, einen Politiker wählen, einen Kinofilm ansehen. Auch möchten Verbraucher und Bürger gehört werden, wenn sie ihre Erfahrungen und Ansichten berichten.

Wie wichtig das Thema für Firmen ist, sieht man daran, dass diese aufwändige manuelle und mittlerweile auch automatische Verfahren einsetzen, um die Meinung der Kunden zu eruieren, so wie Presserecherchen, Umfragen, das Verfolgen von Meinungsäußerungen in Newsgroups und Foren und die Auswertung der Kunden-E-Mails.

Mit der Zunahme der Relevanz von Kundenmeinungen steigt jedoch auch die Anzahl der Manipulationsversuche. Systematische Untersuchungen zum Anteil der Fake-Bewertungen an den Bewertungen insgesamt gibt es bisher nicht[1], und ich gehe auch davon aus, dass die Anteile je nach Branche sehr unterschiedlich sind. Schätzungen sprechen von 20 bis 30 Prozent[2]. In einer Befragung von Hoteliers, durchgeführt an der FH Worms, haben aber fast die Hälfte der Hoteliers angegeben, Erfahrungen mit gefälschten Bewertungen zu haben (Conrady 2014). Für die Kunden bedeutet das, dass sie sich nicht mehr auf Online-Bewertungen verlassen können und erheblich getäuscht werden.

Da ihre Glaubwürdigkeit stark unter den Manipulationen leidet, gehen die Betreiber der Online-Portale mittlerweile gegen Opinion Spam vor, wie z. B. Amazon[3].

Fake-Bewertungen haben sich zu einem eigenen Geschäftsmodell entwickelt. Es gibt Plattformen und Anbieter für gekaufte Reviews, wie noch bis Frühjahr 2016 buyamazonreviews.com[4], fiverr.com[5] oder reselleratings.com[6].

Selbst Erpressungen durch Gäste wurden von Hotelbesitzern berichtet, wobei die Gäste einen Preisnachlass forderten und andernfalls mit negativen Bewertungen drohten (Conrady 2014).

Die Zeitschrift Ökotest berichtete in ihrer Online-Ausgabe vom 25. Januar 2013 von kriminellen Betrügern, die in einem Online-Portal ein Produkt angeboten, positiv bewertet, dann verkauft und nie ausgeliefert haben.[7] Opinion Spam war hier ein essenzieller Aspekt der kriminellen Machenschaften.

Wir haben es also mit einem gesellschaftlich und ökonomisch relevanten Problem zu tun. Für die Informationswissenschaft stellt sich daher die Frage, wie Opinion Spam erkannt werden kann, und ob es möglich ist, den Erkennungsprozess durch automatische Verfahren zu unterstützen. Wir untersuchen Methoden, die für die englische und die chinesische Sprache entwickelt worden sind, beginnen mit der Entwicklung eines Korpus mit gefälschten Bewertungen und analysieren, mit welchen Mitteln diese klassifiziert werden können.

Gefälschte Bewertungen

(Liu 2012) unterscheidet zunächst Hype-Spam, bei dem ein Produkt oder eine Firma hochgelobt wird, von Defaming-Spam, bei dem verunglimpft wird. Außerdem unterscheidet er „Fake Reviews", bei denen es um einzelne Produkte oder Dienstleistungen geht, von unspezifischen Bewertungen ganzer Firmen oder Dienstleister („Ich liebe XYZ!"). Letztere sind automatisch noch am besten erkennbar.

1 Die Betreiber von Portalen werden diese Zahlen wohl auch nicht veröffentlichen, sofern sie diese haben.

2 Z. B. (Mukherjee 2015): Diese Schätzungen betreffen aber den amerikanischen Markt.

3 http://www.heise.de/newsticker/meldung/Amazon-verklagt-Haendler-wegen-eingekaufter-Fake-Bewertungen-3227189.html.

4 Inzwischen wird man bei dieser Seite auf eine Seite von amazon umgeleitet, die droht: „If we determine that you have attempted to manipulate reviews or violated our guidelines in any other manner, we may immediately suspend or terminate your Amazon privileges, remove reviews, and delist related products." (https://www.amazon.com/gp/help/customer/display.html?nodeId=201749630&ref=cm_udrp_bar).

5 https://www.fiverr.com/search/gigs?utf8=%E2 %9C%93&search_in=everywhere&source=top-bar&locale=de&en_query=&query=bewertungen&page=1&filter=rating.

6 http://resellerratings.com/.

7 http://www.oekotest.de/cgi/index.cgi?artnr=11617;gartnr=91;bernr=23.

Fake Reviewers

Wer sind nun die Personen, die gefälschte Bewertungen verfassen? Erwähnt wurden schon die Auftragnehmer in speziellen „Marketing"-Firmen und die Personen mit kriminellem Hintergrund. Diese Fälle sind eindeutig. Unklarer ist jedoch die Einschätzung, wenn Freunde und Familie eines Buchautors dessen neues Buch bei Amazon bewerten oder die Bewertung eines neuen Autos durch die Mitarbeiter des Herstellers. Was ist darüber hinaus davon zu halten, wenn eine Firma einen Preis dafür ausschreibt, dass Kunden Bewertungen schreiben?

Es wird deutlich, dass es hier keine klaren Grenzen gibt und sich die Erkennung von Opinion Spam daher zunächst auf die klaren Fälle beschränken muss.

Daten als Basis für die Entdeckung

(Liu 2012) stellt Quintupel auf, um Bewertungen zu beschreiben. Diese enthalten die relevanten Elemente einer Empfehlung: Die Entität, die bewertet wird, der bewertete Aspekt dieser Entität, die Meinung dazu, der Reviewer und der Zeitpunkt des Reviews. Auf diese Informationen wird auch zugegriffen, wenn man versucht, Fälschungen zu entdecken. (Liu 2012) beschreibt die Daten, die für die Entdeckung von Fälschungen zur Verfügung stehen auf drei Ebenen: Textebene, Meta-Daten und Produktinformationen.

Auf der Textebene erfährt man Essentielles über die Inhalte (wenn die Formulierungen für unterschiedlicher Produkte sehr ähnlich sind, oder etwa die Kritik eines Buchs einfach aus dem Klappentext kopiert wurde), über die Professionalität (Satzlänge, Fehleranteile), über die benutzten Wörter und über syntaktische und semantische Hinweise auf Lügen.

Die Ebene der Meta-Daten informiert über die vergebenen Sterne (z. B. die 5-Sterne-Bewertung von Amazon), die User-ID, über den Zeitpunkt des Postings, IP- und MAC-Adressen, Standort des Computers, und die Reihenfolge von Klicks.

Die Produktebene berichtet über den Verkaufsrang und die Produkteigenschaften.

Auf der Basis dieser Informationen wird versucht, Bewertungen zu klassifizieren. Dabei existiert ein wesentliches Problem: Es steht kein solide validiertes Korpus von gefälschten Bewertungen für die deutsche Sprache zur Verfügung mit dem man z. B. Data Mining-Systeme trainieren oder testen könnte. Das erste Korpus für das Englische wurde 2011 entwickelt und in (Ott, et al. 2011) beschrieben. Es enthält 400 echte und 400 gefälschte Bewertungen. Dafür nahmen (Ott, et al. 2011) aus TripAdvisor fünf Sterne-Empfehlungen für die 20 populärsten Hotels in der Gegend von Chicago als echte Bewertungen und beauftragten mit Amazon Mechanical Turk[8] Versuchspersonen, gefälschte Kritiken für diese Hotels zu verfassen.

(Wang, Liu und Yu 2012) erstellten ein Korpus für das Englische durch Testpersonen, die gefälschte Bewertungen für Produkte, die sie nicht kannten, schreiben sollten. Diese Vorgehensweise mit Versuchspersonen ist jedoch problematisch, da unklar ist, ob die Ergebnisse authentisch sind.

(Li, et al. 2015) berichten über ein Korpus für das Chinesische mit sechs Millionen Bewertungen für Restaurants in Shanghai, das auch Metadaten, z. B. Zeit des Postings, IP-Adressen und Standorte der Computer enthält. Die chinesische Firma Dianping benutzt dabei einen eigenen Filter, dessen Algorithmus Firmengeheimnis ist. Doch der Wissenschaft steht immerhin die Information zur Verfügung, welche Bewertungen gefiltert worden sind. Sie klassifizieren Reviewer (und ihre IP-Adressen), von deren Bewertungen mehr als die Hälfte von Dianping als gefälscht eingestuft werden, als nicht vertrauenswürdig.

(Sandulescu und Ester 2015) haben Zugriff auf Empfehlungen, die von Portalen automatisch gefiltert wurden und damit auf ein Korpus von Bewertungen mit Spam-Verdacht. Unklar sind jedoch Precision und Recall der Filtermethoden, die zu diesem Korpus geführt haben. Das Portal Yelp stellt die gefilterten Beiträge auf den Bewertungsseiten zur Verfügung.[9] Es ist jedoch auf den ersten Blick nicht ersichtlich, auf welcher Grundlage diese gefiltert worden sind. Sehr wahrscheinlich basieren diese Filter eher auf einer Kategorisierung des Verhaltens der Reviewer als auf deren Formulierungen. Als Gold-Standard für eine Evaluation oder für Machine Learning-Verfahren eignen sich diese deutschen Beispiele nicht.

(Shojaee, et al. 2015) erstellen ein Korpus, indem sie Annotatoren nach gefälschten Bewertungen in einem Online-Portal suchen lassen und dann das Inter-Annotator Agreement (also die Übereinstimmung zwischen den Annotatoren) messen. Dies ist sicher eine recht zuverlässige Methode, auch wenn Fake-Bewertungen nicht leicht zu erkennen sind. Die Autoren unterstützen den Annotationsprozess dadurch, dass Sie den Annotatoren gezielte Fragen stellen[10] und alle Bewertungen eines Reviewers gleichzeitig präsentieren.

8 https://www.mturk.com.

9 Z. B.: https://www.yelp.de/biz/die-kaffee-d%C3 %BCsseldorf [26.7.2016].

10 Z. B.: „Is this review unrelated to the product?".

Fake-Erkennung als Klassifikationsaufgabe

Die Erkennung von Opinion Spam ist eine klassische Klassifikationsaufgabe, die Dokumente (Bewertungen) als gefälscht oder als nicht gefälscht einordnen soll. Im Folgenden werden Methoden für diese Klassifikationsaufgabe vorgestellt. Diese Methoden unterscheiden sich vor allem dadurch, auf welche Daten sie sich beziehen.

Klassifikation mit Meta-Daten

(Hooi, et al. 2016) klassifizieren Reviewer durch ihr Verhalten. Wenn diese beispielsweise ausschließlich positive Bewertungen in großer Menge abgeben, ist dies ebenso verdächtig, wie wenn sie viele Empfehlungen in sehr kurzer Zeit abgeben. Mit diesen Meta-Daten trainieren sie einen Bayesianisches Modell des Data Mining, um verdächtige Reviewer zu finden.

(Wang, Liu und Yu 2012) betrachteten die Relation zwischen dem Reviewer, den Bewertungen und dem Shop, der das Produkt anbietet. Sie teilten also die Klassifikationsaufgabe in drei Unteraufgaben ein und stellten fest, dass nicht vertrauenswürdige Reviewer eine Beziehung zum Shop hatten. Weiterhin fanden sie heraus, dass es gute und schlechte Shops gibt und dass diese schlechte Spammer engagierten.

Fake-Bewertungen sind nicht ehrlich. Für die Klassifikation der Bewertungen betrachtet man also die Korrelationen der Vertrauenswürdigkeit der Reviewer, der Echtheit der Bewertungen und der Seriosität des Shops:

- Die Vertrauenswürdigkeit des Reviewers ist abhängig von der Anzahl seiner echten Bewertungen.
- Ein Shop ist seriös, wenn viele vertrauenswürdige Reviewer ihn positiv beurteilen, und weniger serös, wenn viele vertrauenswürdige Autoren ihn negativ bewerten.
- Die Echtheit einer Bewertung hängt von der Seriosität des Shops und der Übereinstimmung mit anderen Empfehlungen in einem gegebenen Zeitfenster ab.

Als gravierendes Problem für die Evaluation dieser Methode wird auch von (Wang, Liu und Yu 2012) hervorgehoben, dass es kein ausreichend großes Korpus mit Bewertungen gibt, bei denen klar annotiert ist, ob sie echt oder falsch sind.

Auf Basis der Daten aus ihrem chinesischen Korpus konnten (Li, et al. 2015) Regelmäßigkeiten bezüglich Zeit und Ort feststellen. So sind Spammer häufiger an Wochentagen aktiv und meist weiter von Shanghai entfernt als vertrauenswürdige Autoren, die zum einen eher lokal gebunden und zum anderen vorzugsweise am Wochenende Restaurants bewerten.

(Li, et al. 2015) zeigen so die Möglichkeit auf, die Klassifikation von Reviewern auf der Basis von vorklassifizierten Daten zu lernen. Verknüpft mit den Methoden von (Wang, Liu und Yu 2012) könnte man mit diesen Ergebnissen Shops und letztlich Bewertungen klassifizieren.

Die Zeit des Postings von Bewertungen betrachteten (Ye, Kumar und Akoglu 2016) unter verschiedenen Aspekten. Dabei interpretierten sie Folgendes als Alarmsignale:

- Wenn sich die durchschnittliche Bewertung eines Produkts plötzlich ändert
- Wenn plötzlich extrem viele Bewertungen zu einem Produkt erscheinen
- Wenn Reviewer regelmäßig jeden Tag Bewertungen posten

Klassifikation mit linguistischer Information

(Sandulescu und Ester 2015) stellten fest, dass Spammer zwar häufig ihre Benutzernamen wechselten, aber oft sehr ähnliche Texte für unterschiedliche Produkte verfassten. Die Autoren berechneten daher die Ähnlichkeit zwischen Bewertungen unter Verwendung von Synonymie-Beziehungen aus WordNet sowie Informationen über Lemmatisierung von Wörtern und ihren syntaktischen Kategorien. Sie bestimmten dann in Experimenten einen Grenzwert, ab dem die Bewertungen anderen so ähnlich waren, dass sie als gefälscht klassifiziert werden konnten. Als Datenbasis für die Evaluation nahmen sie Bewertungen aus Yelp und Trustpilot, wobei sie die von den Firmen gefilterten Bewertungen als gefälscht klassifizierten. Damit knüpften sie an Experimente von (Jindal und Liu 2008) an, die ebenfalls nach Empfehlungen suchten, die ein hohes Maß an Ähnlichkeit aufwiesen und damit ein Korpus von Fake Reviews aufbauten.

(Banerjee, Chua und Kim 2015) wiesen darauf hin, dass Fake-Bewertungen oft ohne den vorherigen Kauf eines Produkts oder den Aufenthalt in einem bewerteten Hotel stattfanden. Auch Amazon markiert Bewertungen mit „verifizierter Kauf“, wenn ein Produkt über Amazon bestellt wurde. Hier sind also Metadaten vorhanden, die bei einer Authentifizierung halfen, so dass ein Korpus aus echten Bewertungen aufgebaut werden konnte. Das Korpus gefälschter Bewertungen wurde durch Versuchspersonen erstellt. Dieses nutzten die Autoren, um linguis-

tische Hinweise in Bewertungen zu sammeln, die auf Fakes hindeuten. Sie unterschieden dabei vier linguistische Merkmale: Verständlichkeit, Detailgenauigkeit, Schreibstil und Kognitionsindikatoren. Für die Bestimmung der Verständlichkeit wurden Standard-Verständlichkeitsmaße wie der „Flesch-Kincaid Grade Level" (Kincaid, et al. 1975) berechnet. Der Faktor Detailgenauigkeit erschloss sich durch die verwendeten syntaktischen Kategorien (POS), denn in informativen Texten gibt es mehr Nomen, Adjektive, Artikel und Präpositionen als Verben, Konjunktionen, Adverbien und Pronomen. Dazu kam die Berechnung der lexikalischen Diversität. Die Einordnung des Schreibstils beruhte auf der Benutzung von Emotionswörtern, dem Tempus der Verben, dem Gebrauch von verstärkenden Wörtern wie „always" oder „never" und Satzzeichen wie Fragezeichen oder Ausrufezeichen. Als Kognitionsindikatoren wurden sprachliche Merkmale, die auf Lügen hindeuteten, wie z. B. der Gebrauch von Wörtern wie „should" und „may" oder auch der Gebrauch von Füllwörtern charakterisiert. Mithilfe dieser Merkmale trainierten sie Machine Learning-Systeme.

Im Ergebnis zeigte sich, dass linguistische Merkmale bei der Unterscheidung echter von gefälschten Reviews helfen. Allerdings ist kritisch anzumerken, dass das Korpus der gefälschten Bewertungen in einer künstlichen Experimentsituation entstand und seine Authentizität daher in Frage gestellt werden kann.

Beobachtungen über Opinion Spam im deutschsprachigen Amazon-Portal

Mit Studierenden der Hochschule Darmstadt haben wir Beispiele für Opinion Spam im deutschsprachigen Amazon-Portal gesucht und diese dann analysiert. Überraschend war, wie viele eindeutige Beispiele sich in kurzer Zeit finden lassen.

Das Annotationsschema für das Korpus deutscher Opinion Spam beinhaltet die Ebenen, die auch (Liu 2012) beschrieb: Textebene, Meta-Daten und Produktinformationen.

- Textebene
  - Überschrift
  - Text
- Meta-Daten
  - Review-URL
  - Rating
    - Anzahl der Bewertungen
    - Anzahl der Bewertungen mit fünf Sternen
    - Anzahl der Bewertungen mit einem Stern
    - diese Bewertung
  - Reviewer
    - User-ID
    - verifizierter Kauf
  - Bewertung des Reviews als hilfreich
    - Datum
- Produktinformation
  - Produktname
  - Verkaufsrang
  - Shop

Dazu kommt jeweils ein kurzer Text mit einer Begründung für die Einordnung als Spam.

Das so entstandene Korpus mit hundert Beispielen für Opinion Spam kann jetzt für weitergehende Untersuchungen deutschsprachiger gefälschter Bewertungen genutzt werden. Für automatische Lernverfahren ist das jedoch noch zu klein.

Erste Beobachtungen zeigen, dass die Forschungsergebnisse für das Englische und das Chinesische zum Teil auf das Deutsche übertragbar sind:

So konzentrierten wir uns auf Hype-Spam und Fake-Reviews, weil vor allem diese im deutschen Amazon-Portal vorhanden waren. Wenige Beispiele für Defaming Spam gab es aber ebenfalls. Anders als von (Wang, Liu und Yu 2012) beobachtet, scheint im deutschen Amazon-Portal der Shop nicht ausschlaggebend zu sein. Wenn gefälschte Bewertungen gefunden und zusätzliche Empfehlungen zu Produkten im selben Shop analysiert wurden, so fanden sich nur sehr selten weitere gefälschte Einträge. Es müsste somit untersucht werden, ob eher die Herstellerfirma (z. B. im Fall von technischen Geräten) oder der Autor, Komponist oder ähnliches Opinion Spam in Auftrag geben. In einer weiteren Untersuchung sollte diese Information dann in das Korpus mit aufgenommen werden.

Wie auch bei (Hooi, et al. 2016) traten häufig verdächtige Reviewer in Erscheinung, die den gleichen Text unter dem selben Datum für verschiedene Produkte verwendeten. Dies ist auch ein Ansatzpunkt für eine Erweiterung des Korpus, denn weitere Bewertungen von notorischen Spammern können aufgenommen werden.

Dazu scheint das Datum eine Rolle zu spielen, etwa wenn es direkt nach Erscheinen einer CD sehr viele positive Reviews innerhalb weniger Tage gibt und später dann in erster Linie negative. Dazu zeigte sich, ebenso wie bei (Li, et al. 2015), dass die Spammer meist an Wochentagen und nur in Ausnahmefällen an Wochenenden agierten; nur 25 der gefälschten Bewertungen entstanden am Wochenende, 75 dagegen an einem Wochentag.

Anders als bei (Banerjee, Chua und Kim 2015) handelte es sich bei den gefälschten Einträgen im deutschen Amazon-Portal oft um einen verifizierten Kauf, im Korpus in 84 Fällen von 100. Dies verweist auf eine gewisse Professionalität der Spammer, die entweder direkt von den Shops beauftragt werden oder die Produkte bestellen und danach zurücksenden. Jedenfalls scheint für das deutsche Amazon-Portal die Methode des Korpusaufbaus mit nicht verifizierten Käufen nicht zu funktionieren.

Die Texte – gerade wenn sie von Spammern mehrfach verwendet werden – sind wenig konkret, z. B.: „Alles bestens und schnell wie immer gelaufen – würde ich immer wieder wiederholen. Die Ware ist OK", „also die lieferung ist schnell und unkompliziert. die ware ist top und es gibt keine beanstandungen. da würde ich wieder bestellen. :-)". Häufig beziehen sich die Spammer auf die Lieferung, wie im oben genannten Beispiel, und nicht auf das Produkt selbst, weil sie dann für jedes Produkt eine eigene Bewertung schreiben müssten. Manche versuchen jedoch, auch diesen Prozess zu automatisieren, was im folgenden Fall schiefgegangen ist, weil die Variablen im Text geblieben sind:

„Ich kann das oben angegebene Produkt $article_name vorbehaltlos empfehlen. Als ich $article_medium endlich erwerben konnte, war ich mehr als positiv überrascht. Ich werde auch in Zukunft $article_name immer wieder konsumieren und habe gleich noch einmal zugegriffen, da auch der Preis $article_price für das Produkt $article_name sehr gut ist. Ich freue mich schon auf weitere sehr gute Angebote von $article_manufacturer."

Gefälschte Texte sind im Durchschnitt kürzer als echte Bewertungen (27,5 Tokens pro Bewertung). Viele Spammer reagieren auf die Anforderungen von Amazon nach einer Mindestlänge eines Reviews von 20 Wörtern mit Tricks, wie sinnlose Sätze, Wiederholungen und Wörtern mit Leerzeichen zwischen den Buchstaben:

- „alles war gut, ich habe leider keine weitere Lust noch mehr dazu zu schreiben mit recht freundlichen grüßen danke !!!"
- „gefällt mir, sieht gut aus, ist sehr praktisch,einfach gut,gefällt mir, sieht gut aus, ist sehr praktisch,einfach gut,gefällt mir, sieht gut aus, ist sehr praktisch,einfach gut"
- „Hab den Anhänger damals für ne Freundin bestellt – hat Ihr gefallen – e m p f e h l e n s w e r t"

Eine automatische Klassifikation als gefälschte Bewertung erscheint mit maschinellen Lernverfahren möglich bei einer deutlichen Erweiterung des Korpus und dem Aufbau eines Vergleichskorpus mit echten Bewertungen.

Fazit

Meinungen über Produkte, Reisen, Filme, Bücher, Dienstleistungen und vieles mehr werden heutzutage in unzähligen Foren geäußert. Es ist üblich, sich die Meinung der anderen Kunden durchzulesen, bevor man etwas kauft oder eine Dienstleistung in Anspruch nimmt. Firmen nutzen diese Information, um sich auf Kundenwünsche einzustellen oder Missstände schnell zu beheben.

Unter den Reviewern tummeln sich aber vermehrt Fälscher. Ähnlich wie bei E-Mail-Spam geht das vom kleinen Betrug mit gefälschten Meinungsäußerungen bis hin zu kriminellen Machenschaften.

In diesem Beitrag wurde im ersten Schritt die Information vorgestellt, die für die Identifikation von gefälschten Bewertungen zur Verfügung steht. Neben dem Text und seiner Überschrift haben wir Meta-Daten wie z. B. das Datum des Postings und Produktinformationen wie z. B. den Verkaufsrang.

Letztlich handelt es sich um eine klassischen Klassifikationsaufgabe, um solche Einträge als echt oder gefälscht identifizieren zu können. Für das Englische und das Chinesische gibt es in der Forschungsliteratur verschiedene Ansätze für diese Klassifikation, die im Abschnitt „Fake-Erkennung als Klassifikationsaufgabe" vorgestellt wurden. Die Forschungen haben gemein, dass sie als Basis ein Textkorpus mit gefälschten Bewertungen benötigen. Von der Qualität dieses Korpus hängt ab, wie gut die vorgestellten Methoden funktionieren.

Daher haben wir uns entschieden ein solches Korpus für das Deutsche zu initialisieren. Dafür wurde zunächst ein XML-Annotationsschema entworfen, um die notwendigen Informationen für die Klassifikation bereitzustellen. Im nächsten Schritt erfolgte unter Rückgriff auf das deutsche Amazon-Portal der Aufbau eines kleinen Korpus. Der erste Eindruck bestätigt die Erkenntnisse anhand der englischen und chinesischen Korpora zum großen Teil. Als eine interessante Abweichung im deutschen Amazon-Portal ergab sich aber, dass in unserem Korpus viele gefälschte Bewertungen als „verifizierter Kauf" klassifiziert sind.

Unsere Beobachtungen zeigen: Es ist möglich, den Prozess der Klassifikation gefälschter Bewertungen mit automatischen Methoden zu unterstützen. Dafür ist es notwendig, ein hinreichend großes Textkorpus zu erstellen. Das vorgestellte Annotationsschema enthält die notwendigen Informationen dafür. Im nächsten Schritt werden wir das Korpus erweitern, ein Vergleichskorpus mit nicht gefälschten Bewertungen erstellen und die Methoden zur Erkennung implementieren. Bei den Methoden wird es interessant sein zu sehen, wie linguistische Analysen und Analysen der Meta-Daten zusammenspielen.

Literatur

Banerjee, S., A. Y. Chua und J. J. Kim. „Using supervised learning to classify authentic and fake online reviews.“ Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication. 2015. 88.

Conrady, Roland. „Customer Reviews: kaufentscheidend, glaubwürdig, strategierelevant? Eine empirische Studie der ITB und der Fachhochschule Worms.“ ITB Berlin Kongress. 2014.

Hooi, B. et al. BIRDNEST: Bayesian Inference for Ratings-Fraud Detection. arXiv preprint arXiv:1511.06030, 2016.

Jindal, Nitin und Bing Liu. „Opinion Spam and Analysis.“ Proceedings of the 2008 International Conference on Web Search and Data Mining. 2008. 219–230.

Kincaid, J.P, R.P. Fishburne, R.L. Rodgers und Chisson, B.S. „Derivation of new readability formulas for Navy enlisted personnel.“ Research Branch Report 8–75, U. S. Naval Air Station, Memphis, 1975.

Li, H., Z. Chen, A. Mukherjee, B. Liu und J. Shao. „Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns.“ Proceedings of The 9th International AAAI Conference on Web and Social Media (ICWSM-15), May 2015: 26–29.

Liu, Bing. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.

Mukherjee, A. „Detecting Deceptive Opinion Spam using Linguistics, Behavioral and Statistical Modeling.“ ACL-IJCNLP 2015. 2015. 21.

Ott, M., Y. Choi, C. Cardie, und J. T. Hancock. „Finding deceptive opinion spam by any stretch of the imagination.“ Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011. 309–319.

Sandulescu, V. und M. Ester. „Detecting Singleton Review Spammers Using Semantic Similarity.“ Proceedings of the 24th International Conference on World Wide Web. 2015. 971–976.

Shojaee, S., A. Azman, M. Murad, N. Sharef und N. Sulaiman. „A Framework for Fake Review Annotation.“ Herausgeber: IEEE Computer Society. Proceedings of the 2015 17th UKSIM-AMSS International Conference on Modelling and Simulation. 2015. 153–158.

Wang, G., Xie, S., B. Liu und P. S. Yu. „Identify online store review spammers via social review graph.“ ACM Trans. Intell. Syst. Technol., September 2012.

Ye, J., S. Kumar und L. Akoglu. „Temporal Opinion Spam Detection by Multivariate Indicative Signals.“ Herausgeber: arXiv preprint arXiv:1603.01929. 2016.

Prof. Dr. Melanie Siegel
h_da, Hochschule Darmstadt
Fachbereich Media
Medien-Campus der h_da
Max-Planck-Straße 2
64807 Dieburg
melanie.siegel@h-da.de

Frau Prof. Dr. Melanie Siegel ist studierte Computerlinguistin und Sprachtechnologin und seit 2012 Professorin an der h_da, Hochschule Darmstadt im Fachbereich Media. Ihre fachlichen Schwerpunkte sind Sprachtechnologie, Maschinelle Übersetzung, Syntax und Semantik der japanischen Sprache, Ontologien, Informationsextraktion, Sentimentanalyse und Technische Dokumentation.