Stimmungen in deutschsprachigen Nachrichten, Blogs und dem DAX

Von Robert Remus und Khurshid Ahmad

Zeitreihen, d. h., zeitäbhängige Folgen von Datenpunkten, sind ein wichtiges Instrument für die Analyse von Wetterphänomenen und Börsenkursen. Zeitreihenanalysen werden aber auch zunehmend auf Textkorpora angewendet, im vorliegenden Fall auf deutschsprachige Zeitungsartikel und Blogbeiträge mit inhaltlichem Schwerpunkt auf der deutschen Finanzwirtschaft. Die Autoren untersuchen dabei die Eigenschaften in den Veränderungen der Frequenzen positiv und negativ konnotierter Benennungen und vergleichen diese mit den Eigenschaften der Kursverläufe des DAX 30.

Schlagwörter

Sentiment Analysis, positiv und negativ konnotierte Benennungen, Wechselwirkungen zwischen natürlich-sprachlichem Text und Marktindikatoren

Einführung

Noch bis vor kurzem nahm man an, dass Entscheidungen im Bereich der Finanzinstrumente von rational handelnden Individuen getroffen werden. Rational handelnd meint dabei den Ausschluss einer Beeinflussung des Handelns durch Gefühle und Stimmungen, beispielsweise beim Kauf oder Verkauf von Aktien. Finanzexperten behaupteten, dass sogar im Falle irrationalen Handelns einiger weniger Individuen der Einfluss dieser durch das rationale Handeln vieler ausgeglichen würde. Im Gegensatz zu dieser Annahme steht der Zusammenbruch des weltweiten Finanzsystems 2008. Die Süddeutsche Zeitung schrieb am 4. Oktober 2008 in ihrem mit „Lehren aus der Finanzkrise: Die Mär vom Markt“ überschriebenen Artikel: „Der Homo oeconomicus entpuppt sich plötzlich als ein Wesen, das nicht rational handelt, sondern Gefühle zeigt.“ Gefühle und Stimmungen spielten und spielen offenbar ein wichtige Rolle im Finanzsystem und haben auch zu dem Fiasko geführt, dessen Auswirkungen bis zum heutigen Tag andauern.

Im Zeichen dieser Geschehnisse werden die Wechselwirkungen zwischen Stimmungen, die in Nachrichten und Blogs durch Individuen oder Gruppen von Individuen ausgedrückt werden, und den Kursverläufen bedeutender Marktindikatoren untersucht und bemessen.

Fallstudie

Stimmungen werden in natürlichsprachlichen Texten zum Ausdruck gebracht und können annäherungsweise durch eine Frequenzanalyse positiv und negativ konnotierter Benennungen, unabhängig ihres Kontextes, erfasst werden.

Diese Annahme bleibt selbstverständlich kritisch zu hinterfragen, stellt jedoch den Ausgangspunkt für tiefgreifendere Überlegungen dar. Diese Studie vergleicht die Stimmungsverläufe in deutschsprachigen Zeitungsartikeln und Blogbeiträgen mit inhaltlichem Schwerpunkt auf der deutschen Finanzwirtschaft mit den Kursverläufen eines deutschen Aktienindexes, dem DAX 30. Ziel der Untersuchungen ist es, Grundlagen für die Bemessung des Einflusses von Zeitungsartikeln und Blogbeiträgen auf die Preisentwicklung von Aktien und umgekehrt zu schaffen. Die für unsere Untersuchungen notwendigen Ressourcen, ein Textkorpus, ein Wörterbuch positiv bzw. negativ konnotierter Benennungen und ein Referenzpunkt in Form eines Aktienindexes, werden im Folgenden beschrieben.

Textkorpus

Um Veränderungen im Stimmungsverlauf zu beobachten, werden Quellen benötigt, die regelmäßig veröffentlichen. Der Textkorpus besteht daher aus Artikeln der Süddeutschen Zeitung, die in den Sektionen Wirtschaft und Geld platziert wurden (insgesamt 3,91 Millionen Wörter), und Beiträgen in Weblogs bzw. Blogs zum Thema „Deutscher Finanzmarkt“ (insgesamt 0,43 Millionen Wörter). Die Zeitungsartikel und Blogbeiträge wurden zwischen 2006 und 2008 veröffentlicht.

Zeitungstexte stellen dabei im Wesentlichen eine indirekte Meinungswiedergabe dar, die redaktionell gefiltert und zeitlich leicht verzögert ist. Blogbeiträge hingegen sind ein Medium der direkten Meinungswiedergabe, die in der Regel

ungefiltert und unmittelbar zur Verfügung steht. Es ist zu erwarten, dass sich die unterschiedliche Natur dieser Medien, auch im Hinblick auf Sprachstil und Terminologie, in den Ergebnissen niederschlägt.

Wörterbuch

Unser deutschsprachiges Wörterbuch basiert auf dem englischsprachigen Wörterbuch General Inquirer (GI) [1]. Dieses Wörterbuch enthält 1914 positive Benennungen und 2293 negative Benennungen. Diese Benennungen wurden per Google translate (http://translate.google.com) halb automatisch, halb manuell ins Deutsche übersetzt, überprüft und anschließend um ihre Flexionsformen erweitert. Um der Domäne der Finanzwirtschaft terminologisch gerecht zu werden, wurden darüber hinaus domänenspezifische Benennungen wie z. B. *Finanzkrise* und *Bankrott* hinzugefügt.

Die Häufigkeiten des Auftretens dieser Benennungen in den oben beschriebenen Textkorpora wurden jeweils täglich und automatisch ausgezählt. Sie folgen Zipf-ähnlichen Verteilungen [2], d. h., wenige Benennungen finden sehr häufig Verwendung, einige Benennungen treten weniger häufig auf und viele Benennungen sind sehr selten, die sogenannten *hapax legomena*.

Jährlich betrachtet bleiben die Häufigkeiten der positiven Benennungen ungefähr konstant: Sie steuern ca. 4% der Gesamtanzahl aller Wörter des Nachrichten-Textkorpus bei. Die Häufigkeiten der negativen Benennungen steigen zwischen 2006 und 2008 dagegen kontinuierlich (siehe Tabelle 1).

Ähnlich verhält es sich für die Häufigkeiten in unserem Blog-Textkorpus. Hier steuern die positiven Benennungen im Schnitt 5,6%, die negativen Benennungen im Schnitt 2,1% aller Wörter bei.

Aktienindex

Untersucht wurde der Deutsche Aktienindex 30 (DAX 30), der die 30 größten und meistgehandelten deutschen Unternehmen umfasst, die an der Frankfurter Börse geführt werden, und der daher als einer der maßgebenden Marktindikatoren Deutschlands angesehen werden kann. Da der Wert des DAX 30 im Tagesverlauf variiert, wurden für den Zeitraum zwischen 2006 und 2008 die jeweiligen Werte der Tagesabschlüsse verwendet.

Benennungen	2007	2008	2009
positiv konnotiert	4,35%	4,28%	4,36%
negativ konnotiert	2,62%	2,76%	3,14%

Tabelle 1: Häufigkeiten positiv und negativ konnotierter Termini.

Return-Serien

Ein Return r_t wird definiert als der natürliche Logarithmus aus dem Verhältnis eines Preises p zum Zeitpunkt t und Zeitpunkt t-1, d. h.

$$\mathbf{r_t = ln(p_t / p_t\text{-}1)}$$

Beispielsweise ergibt sich der Return des DAX 30 für zwei aufeinanderfolgende Tage aus dem natürlichen Logarithmus des Verhältnisses seines Wertes an einem Tag und dem des vorherigen Tages. Der Return ist somit ein dimensionsloses Maß der Veränderung einer Größe und ermöglicht daher den Vergleich sonst nicht vergleichbarer Zeitreihen. Darüber hinaus lässt sich mittels der Returnwerte die Volatilität v bestimmen. Sie wird definiert als die Standardabweichung der zurückliegenden Returnwerte über eine bestimmte Zeitdauer n. Die Volatilität ist somit ein Maß der Fluktuation einer Größe.

In einer vorangegangenen Studie wurde vorgeschlagen [3], den Einfluss positiver bzw. negativer Stimmung in Nachrichten auf die Finanzmärkte anhand von Häufigkeiten positiv bzw. negativ konnotierter Benennungen und Variationen zentraler Marktindikatoren zu bemessen. Motiviert ist dieser Vorschlag dadurch, dass sich die Returnwerte von Aktien- und Rohstoffpreisen, Wechselkursen und Marktindikatoren wie dem DAX 30 auf einem gewissen Level statistischer Be-

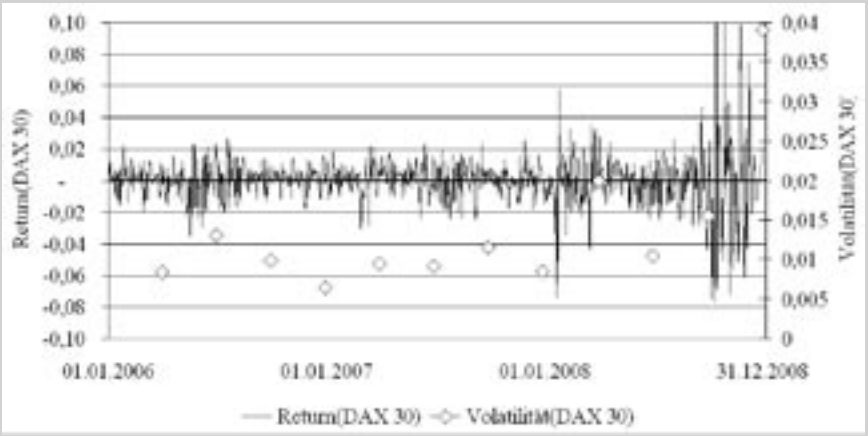

Abbildung 1: Returnwerte und die vierteljährliche Volatilität des DAX 30

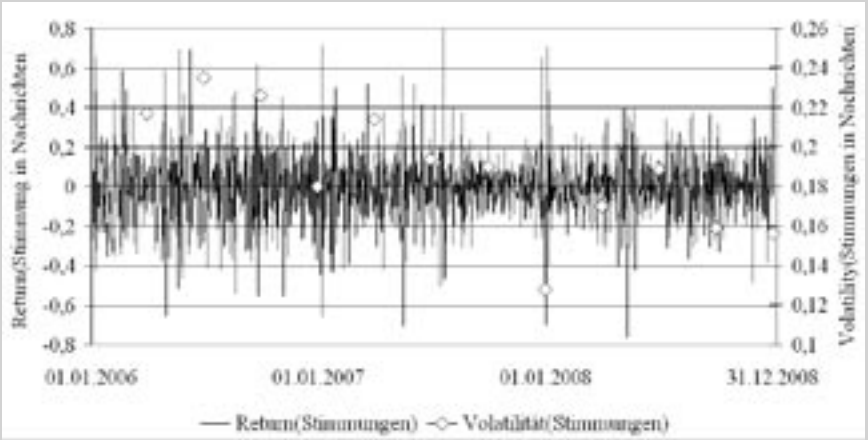

Abbildung 2: Returnwerte und die Volatilität der Häufigkeiten positiv und negativ konnotierter Benennungen in Zeitungsartikeln

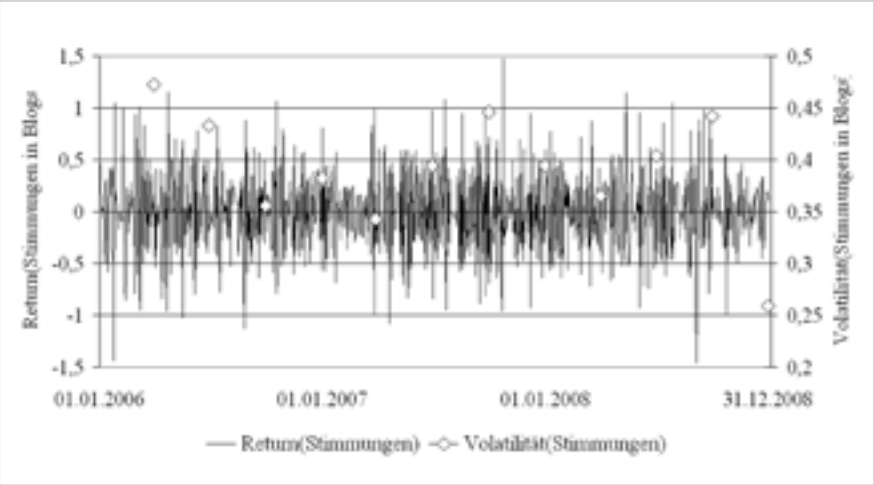

Abbildung 3: Returnwerte und die Volatilität der Häufigkeiten positiv und negativ konnotierter Benennungen in Blogbeiträgen

schreibung ähnlich verhalten. Weisen nun Stimmungsschwankungen in Nachrichten ein ähnliches Verhalten auf, so ist eine Grundlage geschaffen, die es ermöglicht, die Verteilung positiv und negativ konnotierter Benennungen mit der Verteilung von Preisen, Preisentwicklungen usw. zu vergleichen und ggf. zu korrelieren.

Abbildung 1 zeigt die Returnwerte und die vierteljährliche Volatilität des DAX 30. Während die Schwankungen in den Jahren 2006 und 2007 gering ausfallen, zeigen sie sich im Jahr 2008 deutlicher, besonders im dritten und vierten Quartal - dem Zeitpunkt des Zusammenbruchs des weltweiten Finanzsystems. Vergleichend dazu sind in Abbildung 2 und 3 die Returnwerte und die Volatilität der Häufigkeiten positiv und negativ konnotierter Benennungen in Zeitungsartikeln und Blogbeiträgen dargestellt. Diese Returnwerte gestalten sich deutlich sprunghafter und sind weniger klar interpretierbar. Auffällig ist aber, dass sich die Volatilität des DAX 30 und die der Benennungshäufigkeiten in etwa reziprok verhält. Während die Volatilität des DAX 30 vom 1. Quartal 2006 bis zum 1. Quartal 2008 niedrig ist, fällt sie im gleichen Zeitraum für die Benennungshäufigkeiten hoch aus. Mit dem sprunghaften Anstieg der Volatilität des DAX 30 in der zweiten Jahreshälfte 2008 fällt gleichzeitig die Volatilität der Benennungshäufigkeiten, d. h. der Schwankungen positiv und negativ konnotierter Benennungen.

Das Verhalten von Returnwerten ist Untersuchungsgegenstand der Ökonometrie. Hier wird argumentiert, dass die Wahrscheinlichkeit eines Preisanstiegs, also eines positiven Returns, ungefähr der Wahrscheinlichkeit eines Preisabfalls, also eines negativen Returns, entspricht. Dementsprechend seien die Returnwerte ähnlich einer Gaußschen Glockenkurve, also normalverteilt. Dies ist aber nicht grundsätzlich der Fall, wie im vorliegenden Beitrag noch exemplarisch gezeigt wird.

Stephen J. Taylor schlägt vor [4], dass die verschiedenen Momente einer Verteilung von Returnwerten genutzt werden sollten, um stilisierten Variablen zusammenzustellen, die einen Vergleich verschiedener Datensätze bzw. Zeitreihen ermöglichen. Das erste Moment ist hierbei das Mittel der Verteilung der Zeitreihe, das zweite Moment ihre Standardabweichung, das dritte Moment ihre Schiefe und das vierte Moment ihre Wölbung. Eben jene stilisierten Variablen wurden für die Verteilungen der Zeitreihen der Frequenzen positiv und negativ konnotierter Benennungen berechnet und beschrieben.

In den Minima, Maxima und (skalierten) Mittelwerten ist erkennbar, dass die Veränderungen in positiver wie negativer Stimmung in deutschsprachigen Zeitungsartikeln im Allgemeinen höher ist als die Kursschwankungen des DAX 30. Die Veränderungen in positiver wie negativer Stimmung in deutschsprachigen Blog-Beiträgen wiederum sind allgemein höher als die in deutschsprachigen Zeitungsartikeln. Dies spiegelt sich auch deutlich in der (skalierten) Standardabweichung, d. h. der Volatilität, wieder. Die Wölbungswerte zeigen, dass alle Zufallsverteilungen außer „Blogs negativ“ hohe Peaks haben. Aus den Schiefewerten lässt sich ablesen, dass die Zufallsverteilungen „Nachrichten positiv“, „Nachrichten negativ“ und „Blogs positiv“ leicht rechtsschief sind, d. h., dass es mehr Anstiege als Abfälle in diesen Zeitreihen gibt. Für die Schiefewerte der Zufallsverteilungen „DAX 30“ und „Blogs negativ“ gilt das Gegenteil. Sie sind leicht links-schief, d. h. es gibt mehr Abfälle als Anstiege in diesen Zeitreihen.

Bemerkenswert ist weiterhin, dass die Veränderungen in der negativen Stimmung deutlich größer sind als die Veränderungen in der positiven Stimmung - dies lässt sich zum einen aus den Mittelwerten, zum anderen aus den Standardabweichungen der Zeitreihen ablesen.

Zeitreihe	Min	Max	$10^{\wedge 4}$ x Mittel	$10^{\wedge 2}$ x Std.-Abw.	Schiefe	Wölbung
DAX 30	-0,09	0,11	-2,03	1,61	0,15	11,14
Nachrichten positiv	-1,82	1,66	1,65	28,21	-0,07	5,45
Nachrichten negativ	-1,54	1,69	8,29	35,44	-0,03	1,45
Blogs positiv	-2,2	2,08	5,34	45,02	-0,02	1,81
Blogs negativ	-2,18	2,48	16,18	68,96	0,04	0,24

Tabelle 2: Stilisierte Variablen der den Zeitreihen zugrunde liegenden Zufallsverteilungen

Verteilung innerhalb	Normal		Blogs		Nachrichten	DAX 30
		positiv	negativ	positiv	negativ	
0 bis 0,25 Std.-Abw.	19,74	19,53	20,12	24,74	22,77	33,46
0 bis 0,5 Std.-Abw.	18,55	19,88	19,41	23,61	20,41	22,31
0 bis 1 Std.-Abw.	29,98	33,14	31,12	30,01	29,82	27,17
1 bis 1,5 Std.-Abw.	18,37	15,74	15,74	11,67	14,77	10,24
1,5bis 2 Std.-Abw.	8,81	7,34	8,76	4,70	7,15	3,02
2 bis 3 Std.-Abw.	4,28	3,55	4,38	3,57	4,14	1,44
3+ Std.-Abw.	0,27	0,83	0,47	1,69	0,94	2,36

Tabelle 3: Verteilung der Zeitreihen

Nicht-normale Verteilung der Zeitreihen

Wie oben bereits angedeutet, ist spätestens seit [4] bekannt, dass die Return-Serien von Aktienindizes wie beispielsweise dem DAX 30 nicht normalverteilt sind. Der DAX 30 weist im Zeitraum zwischen 2006 und 2008 ca. 8,7 mal mehr Ausreißer auf, d. h. Werte, die über die dritte Standardabweichung einer vorhergesagten Normalverteilung hinausgehen. Diese Erkenntnis ist insofern bemerkenswert, als dass nach wie vor viele mathematische Modelle, die der Beschreibung und Vorhersage von Kursverläufen dienen, auf der Annahme basieren, die zugrunde liegenden Daten seien normalverteilt. Dies wurde für die Stimmungsverläufe in deutschsprachigen Zeitungsartikeln und Blogbeiträgen gezeigt. Zeitungstexte weisen im Mittel ca. 4,9 mal und Blogbeiträge im Mittel ca. 2,4 mal so viele Ausreißer auf wie von einer Normalverteilung vorausgesagt. Diese Ergebnisse konnte für alle Zeitreihen (abgesehen von „Blogs negativ“) in Testverfahren validiert werden.

Fazit

Die Studie hat gezeigt, dass Stimmungsschwankungen in deutschsprachigen Zeitungsartikeln und Blogbeiträgen einige Charakteristika mit den Verläufen einem der maßgebenden Marktindikatoren Deutschlands, dem DAX 30, gemein haben. Dies lässt hoffen, dass zumindest in begrenztem Maße durch die Beobachtung des Verhaltens des Einen über das Verhalten des Anderen gelernt werden kann und umgekehrt.

Darüber hinaus wurde unsere Erwartung bestätigt, dass Stimmungsschwankungen in einem ungefiltertem, unmittelbar zur Verfügung stehenden Medium wie beispielsweise den Blogs stärker zu Tage treten als in einem redaktionell gefiltertem und zeitlich leicht verzögert zugänglichem Medium wie den Nachrichten. Sowohl in Zeitungsartikeln als auch in Blogbeiträgen fallen die Schwankungen negativer Stimmung deutlicher aus als die Schwankungen positiver Stimmung.

Im Zuge der Studie wurde außerdem ein umfangreiches Wörterbuch positiv und negativ konnotierter Benennungen erstellt, das in überarbeiteter Form in Kürze über das Wortschatz-Projekt (http://wortschatz.uni-leipzig.de) der Abteilung für Automatische Sprachverarbeitug an der Universität Leipzig frei zur Verfügung stehen wird [5].

Literatur

[1] Stone, P. J. et al. (1966): The General Inquirer: A Computer Approach to Content Analysis. Boston: The MIT Press.

[2] Zipf, G. K. (1949): Human Behaviour and the Principle of Least Effort. Cambridge: Addison-Wesley Press.

[3] Ahmad, K. (2008): The „Return“ and „Volatility“ of Sentiments: An Attempt to Quantify the Behavior of the Markets? In: Proceedings of the 6th Intl. Conf. on Language Ressources and Evaluation.

[4] Taylor, S. J. (2005): Asset Price Dynamics, Volatility, and Prediction. Princeton: Princeton University Press.

[5] Remus, R. / Quasthoff, U. / Heyer, G. (2010): SentiWS - a publicly available German-language Resource for Sentiment Analysis. Erscheint 2010.

Prof. Khurshid Ahmad ist Inhaber des Lehrstuhls für Informatik am Trinity College Dublin. Zu den Schwerpunkten seiner Forschung gehören Stimmungsanalysen, Wissensrepräsentation durch Ontologien, Bild- und Videoannotation sowie die Modellierung kognitiver Prozesse.

Kontaktadresse
Trintiy College Dublin
Computer Science Department
kahmad@tcd.ie
www.cs.tcd.ie/Khurshid.Ahmad

Robert Remus studiert seit 2004 Linguistische Informatik an der Universität Leipzig, wirkte zwischenzeitlich in einem Projekt am Trinity College Dublin mit und schreibt gegenwärtig seine Diplomarbeit auf dem Gebiet der Sentiment Analysis.

Kontaktadresse
Universität Leipzig
Fakultät für Mathematik und Informatik
robert.remus@googlemail.com