Reddiment: Reddit Sentiment-Analyse

Tobias Bauer
t.bauer@oth-aw.de

Fabian Beer
f.beer1@oth-aw.de

Daniel Holl
d.holl1@oth-aw.de

Ardian Imeraj
a.imeraj@oth-aw.de

Konrad Schweiger
k.schweiger@oth-aw.de

Philipp Stangl
p.stangl1@oth-aw.de

Wolfgang Weigl
w.weigl@oth-aw.de

I. Einleitung

Die GameStop-Aktie (NYSE: GME) hatte Anfang 2021 eine längere Periode mit schnellen, drastischen Kursschwankungen erlebt. Der Subreddit `r/wallstreetbets`, ein Unterforum von Reddit, spielte dabei eine Rolle. `r/wallstreetbets`, auch bekannt als WallStreetBets oder WSB, ist ein Subreddit, in dem über Aktien- und Optionshandel spekuliert wird. Der Subreddit ist bekannt für seine profane Art und die Vorwürfe, dass Nutzer/innen Wertpapiere manipulieren. Im Fall der GameStop-Aktie wurde der Subreddit von Wang und Luo [1] in Bezug auf Vorhersagen durch Sentiment-Analyse betrachtet. Häufig wird Sentiment-Analyse dazu verwendet, um zu bestimmen, ob ein Text negative, positive oder neutrale Emotionen enthält.

In den weiteren Abschnitten des Konzeptpapiers stellen wir *Reddiment* vor – ein webbasiertes Dashboard zur Sentiment-Analyse von Subreddits. In Abschnitt II wird eine Auswahl verwandter Arbeiten vorgestellt. In Abschnitt III werden die Anforderungen in Form von User Stories wiedergegeben. Abschließend wird kurz auf die Methoden in Abschnitt IV eingegangen, gefolgt vom Literaturverzeichnis am Ende.

II. Verwandte Arbeiten

Lubitz [2] hat sich beispielsweise mit möglichen Treibern von Kapitalmärkten beschäftigt. Anhand von Sentiment-Analyse wurden Finanznews verglichen, die auf Reddit und in der Financial Times erschienen sind. Die Vorhersagekraft von Beiträgen auf Reddit in Bezug auf künftige Marktbewegungen sei Lubitz zufolge etwas besser als bei der Analyse von Zeitungen.

Die kommerzielle Plattform Brandwatch [3] ermöglicht die Analyse des Volumens der Gespräche bis hin zum Sentiment über ein Dashboard. So kann man sehen, welche Themen häufig besprochen werden oder welcher Subreddit aktiv ist. Es lassen sich Abfragen zu beliebigen Begriffen erstellen und mithilfe von booleschen Operatoren können diese Abfragen umfassend und spezifisch sein.

III. Anforderungen

In der Anforderungsanalyse wurden drei Stakeholder identifiziert: Benutzer, Entwickler und DevOps-Engineer. Deren Anforderungen werden in diesem Abschnitt in Form von User Stories beschrieben.

A. Selektion von Subreddits

Als Benutzer möchte ich Subreddits auswählen können, damit ich basierend auf der Selektion spezifischere Operationen ausführen kann. Akzeptanzkriterien sind:

- Eingabefeld für Subreddits.
- Subreddits werden als Namen oder URL des Subreddits eingegeben.
- Button zur Bestätigung der Selektion.

B. Eingrenzung der Suche

Als Benutzer möchte ich Schlagwörter spezifizieren können, welche den Suchauftrag eingrenzen. Akzeptanzkriterien sind:

- Eingabefeld für Schlagwörter.
- Schlagwörter können als beliebige Strings eingegeben werden.
- Suchauftrag agiert ausschließlich mit diesen Schlagwörtern.
- Zeitliche Eingrenzung der Suche durch vordefinierte Zeitintervalle (z.B. 7 Tage).

C. Visuelle Darstellung der Erwähnung

Als Benutzer möchte ich einen Graph der Erwähnungsrate der gewählten Begriffe, damit ich die Zunahme bzw. Abnahme des Interesses daran visuell erfassen kann. Akzeptanzkriterien sind:

- Zeit wird über die x-Achse aufgetragen.
- Visualisiertes Zeitintervall kann aus vordefinierten Zeitintervalle (z.B. 7 Tage) ausgewählt werden.
- Erwähnungsrate der eingegebenen Begriffe wird auf der y-Achse aufgetragen.
- Erwähnungsrate wird als absoluter Wert dargestellt.
- Graph muss so skaliert sein, dass das Graph-Feld ausreichend genutzt wird (Die Platznutzung liegt bei mindestens 90%).

D. Visuelle Darstellung des Sentiments

Als Benutzer möchte ich eine grafische Auswertung zum Sentiment, damit ich visuell erkennen kann, ob sich die Stimmung gegenüber den von mir gewählten Begriffen über die Zeit bessert oder verschlechtert. Akzeptanzkriterien sind:

- Dieser Graph soll auf denselben Achsen dargestellt werden wie der Graph zur Erwähnungsrate.
- Sentiment wird als absoluter Wert über die Zeit dargestellt.

E. Entwicklungsdatenbank

Als Entwickler möchte ich eine Entwicklungsdatenbank, damit ich die Funktionalität getrennt von der Produktionsumgebung umsetzen und evaluieren kann. Akzeptanzkriterien sind:

- Datenschema der Entwicklungsdatenbank ist mit dem der in der Produktion eingesetzten Datenbank identisch.
- Entwicklungsdatenbank wird mit realen Reddit-Daten populiert.

F. Testabdeckung

Als Entwickler möchte ich eine ausreichende Testabdeckung, damit Fehler frühzeitig erkannt werden. Akzeptanzkriterien sind:

- Backend-Code wird durch ein geeignetes Werkzeug getestet, beispielsweise `istanbul/nyc`.
- Testabdeckungsrate liegt bezüglich Backend-Routinen bei mindestens 50%.
- Frontend-Code wird durch ein geeignetes Werkzeug getestet, das abhängig von der Wahl des Frontend-Frameworks ausgewählt wird.

G. Cloud-Kompatibilität

Als DevOps-Engineer möchte ich eine Cloud-kompatible Anwendung für eine einfachere Bereitstellung und Skalierung der Anwendung. Akzeptanzkriterien sind:

- Alle Teile der Anwendung laufen einzeln in Docker-Containern.
- Grundlegende Secrets-Verwaltung zur sicheren Aufbewahrung von Zugangsinformationen ist eingerichtet.

H. (Optional) Bereitstellung in der Cloud

Als DevOps-Engineer möchte ich die Cloud-kompatible Anwendung bei einem Cloud-Anbieter bereitstellen. Akzeptanzkriterien sind:

- Prinzipielle Erreichbarkeit ist sichergestellt, sodass bei der Abschlusspräsentation eine Demonstration der Anwendung in der Cloud möglich ist.
- Dokumentation der eingeschränkte Erreichbarkeit der Anwendung in der Cloud.

I. (Optional) Verknüpfung mit Aktienkurs

Als Benutzer möchte ich eine grafische Darstellung eines wählbaren Aktienkurses, damit ich eine visuelle Möglichkeit bekomme, die vorhandenen Graphen mit dem Verlauf des Aktienkurses zu vergleichen. Akzeptanzkriterien sind:

- Eingabefeld für einen Aktiennamen.
- Eingabe des Aktiennamen als beliebiger String.
- Button zur Bestätigung des Aktiennamens.
- Fehlermeldung, wenn Aktienname nicht gefunden wurde.
- Darstellung des Aktienverlaufs über denselben Zeitbereich wie die Auswertung der Erwähnungsrate.

IV. Methoden

Für die Repräsentation der Anwendung im Client-seitigen Frontend wird ein geeignetes Framework verwendet. Zur Speicherung der Daten kommt voraussichtlich Elasticsearch zum Einsatz. Ein Node.js-Webserver mit ExpressJS im Server-seitigen Backend stellt die Funktionalität der im Frontend angebotenen Aktionen bereit. Die Kommunikation zwischen Frontend und Backend wird über eine RESTful-API abgewickelt. Um eine fehlerfreie Anwendung zu entwickeln, wird zum einen TypeScript als projektweite Programmiersprache verwendet. Dadurch sollen Fehler bereits zur Kompilierzeit identifiziert werden können. Zum anderen wird ein geeignetes Test-Framework für Unit-Tests verwendet.

Literatur

[1] C. Wang und B. Luo, "Predicting $ GME Stock Price Movement Using Sentiment from Reddit r/wallstreetbets," in *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, 2021, S. 22–30.

[2] M. Lubitz, "Who drives the market? Sentiment analysis of financial news posted on Reddit and Financial Times," *University of Freiburg Publications*, 2017.

[3] J. Boyd. "Reddit Top 6 Tools für die Subreddit Analyse." [Online]. (2022), Adresse: https://www.brandwatch.com/de/blog/reddit-top-6-tools-fuer-die-subreddit-analyse/.