Relatie twittersentiment en SocialMedia aandeelprijzen

Dashboard ten behoeve van het relateren van het twittersentiment aan de socialmedia aandeelprijs

Jeroen Meijaard

Inhoudsopgave

1 Inleiding

De afgelopen jaren hebben social media een grote groei doorgemaakt. Steeds meer mensen maken gebruik van social media om onder andere hun mening te uiten over een varieteit aan onderwerpen. Na het lezen van de master thesis van Guangxue Cao [1] Assessed by Social Media Sentiment, august 2014, ben ik geinteresseerd geraakt in de relatie tussen Twitter sentiment en koersprijzen. Hoewel de profesionele aandeelhandelaren uitgebreide tools hebben om trends te spotten, lopen consumenten steeds verder achter.

2 Doel

2.1 Visualisatie

Het doel van de visualisatie is om de gebruiker een beeld te geven van de relatie tussen het Twitter sentiment en de ontwikkeling van de koersprijs van het aandeel Google. Om dit inzicht te kunnen geven is er een dashboard gemaakt die een dit in kaart brengt.

Eén van deze visualisaties is een grafiek die zowel de koersprijs van het aandeel Google als het Twitter sentiment van dat moment weergeeft. Zo kan de gebruiker kijken of veranderingen in de koersprijs gerelateerd zijn aan het Twitter sentiment. Vervolgens kan aan de hand van een visualisatie, die de correlatie laat zien, worden laten zien of een koersprijsverandering verband houdt met het Twitter sentiment op dat moment. De histogram laat de spreiding zien van het gemiddelde sentiment per gekozen tijdsperiode. Daarnaast bevat de de visualisatie ook een informatiegedeelte waar informatie over de geselecteerde tijdsperiode in de grafieken te vinden is.

Om eventueel voordeel te behalen uit koersfluctuaties wordt er in de visualisatie per dag de datapunten per vijf minuten bekeken. Echter kan door de opzet van de transformatie van de data, gemakkelijk over een grotere tijdsperiode gekeken worden.

2.2 Datacollectie

Het Twitter sentiment is verkregen uit waarneming binenn de totale twitter feed in de maand januari 2012. Deze is van Archive [2] onttrokken. De data bestaat uit JSON bestanden per minuut. Vervolgens zijn deze bestanden onderverdeeld in mappen naar maand, dag, uur. In elke file van een minuut staat elke regel één tweet in JSON geschreven.

De aandeelprijzen zijn gedownload in een tekstbestand van de historische database van Finam [3]. In het bestand zijn per 5 minuten verscheidene prijzen (openingens, sluitings,laagste en hoogste prijs) en de datum met tijdsperiode weergeven.

[1] Cao, G. (2014), The Impacts of Information on Stock Prices, Universiteit van Amsterdam
[2] https://archive.org/details/archiveteam-twitter-stream-2012-01
[3] http://www.finam.ru/analysis/profile041CA00007/default.asp

3 Transformatie

De transformatie is één van de belangrijkste gedeeltes van dit project, aangezien het veelvuldig gebruikt moet kunnen worden. Het transformatie proces kan worden onderverdeeld in de verwerking van de ruwe data in python en de vervolgverwerking in javascript.

3.1 Ruwe data verwerking

Allereerst wordt de data ingeladen in python. In eerste instantie is er geprobeerd met de libary OS.path.walk [4] de data in zijn geheel in te laden. Echter kon vanwege de grote hoeveelheid aan twitterdata die in geladen dient te worden kan de het geheugen van de computer overbelast worden. Daarom is er een functie gebouwd die uitgekozen aantal files per run inlaadt en verwerkt. De functie creert hierbij de path namen die op basis van uur,dag en maandnummers zijn opgeslagen.

Nadat ook de aandeelprijzen zijn ingeladen, wordt de Twitter data gefilterd op de opgegeven zoekwoorden. Deze zoekwoorden zijn afkomstig uit de scriptie van Guangxue Cao [5].

Vervolgens wordt ten behoeve van het berekenen van het sentiment de libary Textblob [6] gebruikt. De libary wordt gebruikt om natuurlijke tekst (common natural language processing) te verwerken om sentiment analyses uit te voeren. Daarna worden de uitkomsten naar aparte csv weggeschreven en wordt door gebruik te maken van de Dateutil [7] en Calendar [8] libary's de eindtijd van de tijdsperiode mee geven aan elke regel. Zodoende kan aan de hand van deze unieke identificatie met de libary Pandas [9]de csv bestanden worden samengevoegd.

3.2 Vervolg bewerking

Wanneer het csv bestand is ingeladen in javscript wordt de data omgezet in een object van objecten. Waarbij elk object een datapunt is met de volgende attributen per tijdsperiode: Sluitingsprijs, hoogste prijs, laagste prijs, openingsprijs, ticker, volume, eindtijd en sentiment.

Op basis van de dag worden vervolgens per dag een object opgeslagen in een nieuw object. Aangezien de datum van de dag zelf de key is, kan makkelijk de data per dag uitgelezen worden in de visualisaties. Om de datums met elkaar te vergelijken is er gebruikt gemaakt van de moment.js [10] libary die een datum in een moment object verandert. Vervolgens kan ook aan de hand van de ingebouwde functies "volgende dagën "dag terug"de data voor update functies van de visualisaties makkelijk worden ontrokken.

De grafiek die de fluctuaties in sentiment en koersprijs laat zien, wordt ingeladen op basis van de laatste prijs van de tijdsperiode en het bijhorende sentiment. Voor de data van de cirkeldiagram worden de negatieve en positieve sentimenten gescheiden op basis van een counter. De transformatie van de data voor de histogram is bereikt door de sentimenten per dag af te ronden naar één decimaal achter de komma. Daarna wordt per unieke sentiment bekeken hoe vaak het voortkomt in de geselecteerde data voor de desbetreffende dag. Tot slot wordt er in de scatterplot de procentuele verandering in de koersprijs berekend en afgezet tegen de fluctuaties in het sentiment.

3.3 Visualisatie ondersteuning

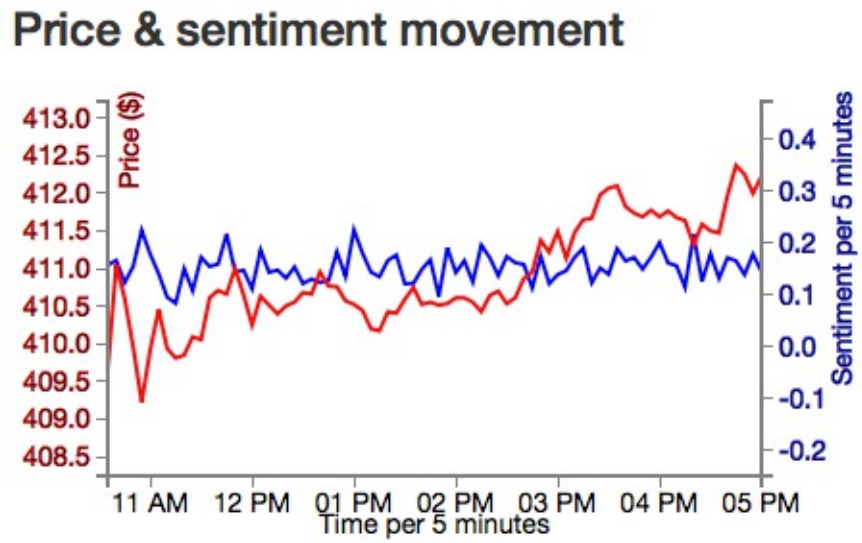
De visualisatie ondersteunt in principe elke twitterinput op basis van json regels en prijsinformatie van tekstbestanden. Ook is de transformatie zo gebouwd dat alle tijdsperiodes ingelezen worden in python. Waarna de javascript code deze opdeeld in tijdsvakken van een dag. Hierdoor wordt de mogelijkheid geboden om een grote verscheidenheid aan datasets te testen.

[4] https://docs.python.org/2/library/os.path.html
[5] Cao, G. (2014), The Impacts of Information on Stock Prices, Universiteit van Amsterdam
[6] http://textblob.readthedocs.org/en/dev/
[7] https://pypi.python.org/pypi/python-dateutil
[8] https://docs.python.org/2/library/calendar.html
[9] http://pandas.pydata.org/
[10] http://momentjs.com/

4 Design

4.1 Visuele coderingen

Zoals in onderstaande grafiek te zien is, hebben beide y assen een ordinale kleurcodering toegekend gekregen op basis van de corresponderende lijn. Hierdoor is het voor de gebruiker direct zichtbaar welke waardes bij de lijn horen. Daarnaast is er voor gekozen om beide assen gelijk te zetten aan de minimale en maximale waardes van de lijnen. Hiervoor is gekozen omdat de gebruiker eerder fluctuaties kan waarnemen.



Ook is er volgens Tufte's principe aan het maximaliseren van contrasten gedacht. Dit is gedaan door de de kleuren rood en blauw te gebruiken in de lijngrafiek. Hierdoor zijn de lijnen goed van elkaar te onderscheiden.

De intergriteit van de weergeven data in de visualisatie komt in deze visualisatie goed naar voren doordat de resultaten overzichtelijk zijn af te lezen en er manipulaties hebben plaats gevonden die de gebruiker anders kan interpreteren. Enige uitzondering hierop is dat de y assen bij de lijngrafiek niet op nul beginnen, wegens eerder genoemde reden. Hierdoor kunnen de veranderingen in koersprijs en sentiment in ander perspectief worden bekeken dan wanneer dit niet was gebeurd. Hiervoor is gekozen omdat anders de kleine verschillen niet af te lezen waren uit de grafiek en met dit in gedachten zijn de assen duidelijk aangegeven, waardoor het voor de gebruiker ook direct duidelijk is dat de y assen niet op nul beginnen.

Omdat het doel van het dashboard is om de gebruiker zo objectief mogelijk van informatie te voorzien is er gelet op het minimaliseren van "chart junk". Daarom is er voor gekozen om de sorteerfunctie in d histogram, uit de visualisatie te halen omdat het weinig toevoegd aan de getoonde informatie.

Wat betreft "data densityïn deze visualisatie, is er gekeken naar de hoeveelheid grafieken die naast elkaar opgenomen kunnen worden per rij. Hierdoor wordt de ruimte op het scherm zo goed mogelijk benut.

4.2 Gebuikers interactie

Voorheen bevonden de knoppen voor het selecteren van de datum zich boven het dashboard.

Om de gebruiker op intuitie door de dagen te laten scrollen, is er voor gekozen om de knoppen voor het selecteren van de dagen naast rondom de datum te plaatsen. Hierdoor ziet de gebruiker direct de verandering.

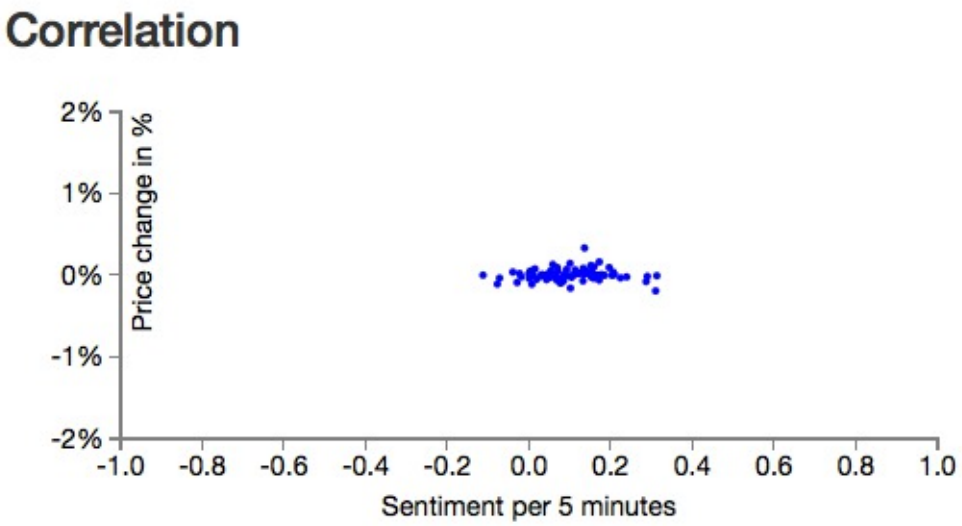
5 Implementatie

Bij de implementatie van de visualisaties is vooral gekeken wat nodig is voor de gebruiker om zo snel mogelijk inzicht te krijgen in de data. Daarom is er voor het selecteren van de columns de css van libary bootstrap [11] gebruikt.

In eerste instantie is ervoor gekozen om de assen in de lijngrafiek van 0 tot het eind te laten lopen. Echter bleek al snel dat kleine veranderingen niet voor de gebruiker op te merken zijn. Daarom zijn de assen aangepast naar de minimale en maximale waarden, zodat deze veranderingen wel merkbaar zijn.

Ondanks dat de cirkeldiagram een laag "data-ink"ratio heeft is er in dit geval toch voor gekozen er één op te nemen. De reden daarvoor is dat de cirkeldiagram de visuele aantrekkelijkheid van de pagina ten goede komt. Daarnaast zal de gebruiker ook sneller het algehele sentiment verschil van dag tot dag waarnemen wanneer een andere dag wordt geselecteerd. Om de gebruiker wel precieze informatie te geven is aan de legenda van de cirkeldiagram het daadwerkelijke percentage per categorie toegevoegd.

De koersprijs veranderingen in de scatterplot zijn eveneens aangepast nadat er goed is gekeken naar de interpretatie door gebruikers. Waar vooraf was gekozen voor een nominale verandering weer te geven. Is uiteindelijk toch een procentuele verandering toegepast. Dit is met het oog op eventuele grote koerschommelingen in de toekomst of wanneer de gebruiker een grotere tijdsperiodes kiest dan 1 dag. Hierdoor vallen de datapunten dus beter met elkaar te vergelijken.



In de histogram zat aanvankelijk een sorteer functie, welke de kolommen eerst door elkaar husselde om deze vervolgens weer goed te zetten. Uiteindelijk is beslote deze transitie uit de visualisatie te halen omdat het geen waarde toevoegd aan de weergeven informatie.

Hoewel aanvankelijk zowel de histogram en de scatterplot op basis van de gehele dataset waren weergeven, is dit toch verandert. Door beide visualisaties per dag te laten varieren, is er op belangrijke dagen voor het geselecteerde aandeel, beter veranderingen te zien.

[11] http://getbootstrap.com

6 Validatie

6.1 Validatie design

De visualisatie is getest door voor meerdere datasets (apple en google) voor de maand januari 2012 te visualiseren. Vervolgens is door gebruik te maken van de feedback tijdens codereviews en presentaties, gekeken naar of de gebruiker de informatie uit de visualisatie kan interpreteren .

Zo zijn op basis van de feedback alle visualisaties interactief gemaakt, zodoende kan ook de correlatie en histogram per dag afgelezen worden. Ook zijn de y assen anders geschaald dan afhankelijk vastgesteld in het oorspronkelijke design. Vervolgens is in de correlatie grafiek de weergeven data op de y as aangepast van norminale koersprijs verandering naar procentuele.

Hoewel het dashboard in grote lijnen overeenkomt met het aanvankelijke design, is er helaas geen selectie mogelijk op verschillende aandelen en tijdsperiodes. Dit is nog een mogelijke toevoeging voor de toekomst.

6.2 Validatie algemeen

De kracht van de visualisatie ligt in de flexibiliteit van de dataverwerking. Zo kan een vergelijkbare dataset zonder aanpassing worden ingeladen. Deze dataset wordt dan direct getoont in de visualisatie.

Door feedback is de inlaad functie voor de ruwe data verandert zodat deze universeel is voor alle data in hetzelfde format.Daarnaast is ook de selectie functie voor het opdelen van de data in tijdperiodes universeel gemaakt in javascript. Dit biedt dan ook veel de mogelijkheden voor andere geintresseerden om eigen datasets in te laden en op deze code voort te bouwen.

Het gemaakte dashboard aan visualisaties kan in de toekomst worden uitgebreid met een knoppen om verschillende datasets te selecteren.