Bachelorscriptie informatica


Radboud Universiteit Nijmegen

# Hoe accuraat kun je de populariteit van muziek voorspellen met behulp van social media?

3 april 2015

*Auteur:*
Dion van de Vooren
s4256468

*Begeleider:*
Tom Heskes

*Tweede lezer:*
Bram den Teuling

# 1 Achtergrond

Dit onderzoek is gebaseerd op de Nederlandse top 40. De top 40 zoals hij nu is wordt samengesteld door SoundAware (onder toezicht van een bestuurslid van de Stichting Nederlandse Top 40) met behulp van de volgende factoren (Stichting Nederlandse Top 40, 2014):

- Airplay
  Er wordt gekeken naar wat de bekende Nederlandse radiostations draaien. Elk nummer word vermenigvuldigd met de luistercijfers van dat uur, waarbij de luistercijfers van de dienst Kijk- en Luisteronderzoek (KLO) stammen. De radiostations die gevolgd worden zijn:
  - Radio 538
  - 3FM
  - Sky Radio
  - Q-music
  - Slam FM
  - 100% NL
- Streaminggegevens
  De gegevens zijn afkomstig van Spotify.
- Muziekonderzoek
  Elke week word er door DVJ Insights een onderzoek uitgevoerd waaruit moet blijken wat in Nederland op dit moment populaire muziek is. Hierbij gaat het alleen om nummers die niet ouder zijn dan 10 weken, omdat oudere nummers beter beoordeeld worden (Stichting Nederlandse Top 40, 2014).

Mensen delen vaak informatie op het internet. De gemiddelde Twitter gebruiker stuurt bijna 2 berichtjes per dag de wereld in, dat zijn in totaal ongeveer 500.000.000 tweets (Twitter, 2015). Die data is openbaar toegankelijk en bevat vaak een mening van de auteur. Dit betekent dat de mening van mensen eigenlijk al voor het oprapen ligt.
Dat is een enorme schat aan data waar ook al veel mee gedaan is. Er is bijvoorbeeld vaak geprobeerd met behulp van social media dingen te voorspellen, en dat ging vaak redelijk goed (zie hoofdstuk 2).

# Inhoudsopgave

## 2 Introductie

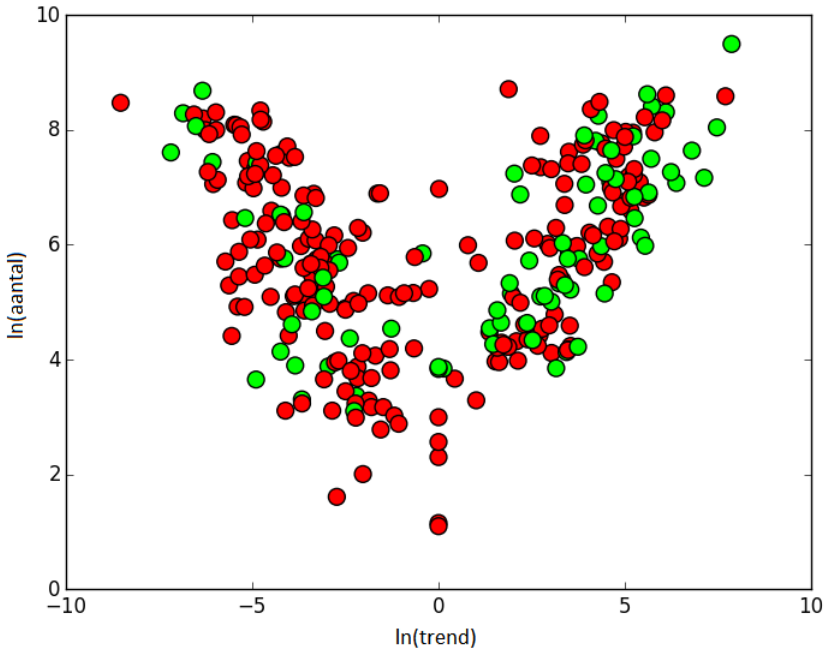
Het is erg lastig te bepalen welke muziek op dit moment populair is. Dit komt doordat er veel verschillende muzieksmaken zijn (Ter Bogt et al., 2003), en er ook nog verschillende manieren zijn om er naar te luisteren. Het kan zo zijn dat de groep die meer van popmuziek houdt vaak muziek downloadt, maar de typische rock-fan via Spotify luistert. Hierdoor geeft het kijken naar bijvoorbeeld alleen het aantal downloads geen goed beeld van de populariteit van dat nummer.
Bepalen wat op dit moment populair is, doet de Stichting Nederlandse Top 40 door een lijst van de 40 meest populaire nummers te maken. Hiervoor kijken ze naar waar mensen op Spotify veel naar luisteren, naar wat de radiostations draaien en doen ze ook een marktonderzoek (hoofdstuk 1). De interessantere vraag is wat er in de toekomst populair gaat worden. Als men weet wat er in de toekomst populair is, kunnen bijvoorbeeld artiesten en producers die data gebruiken om de muziek meer op de doelgroep toe te spitsen. Ook voor radiostations en andere bedrijven die met muziek werken is die data handig. Het zou ook gebruikt kunnen worden op de top 40 samen te stellen.
In dit onderzoek wordt duidelijk hoe nauwkeurig de populariteit voorspeld kan worden. Als maatstaf voor populariteit gebruik ik de top 40, de data die ik voor het voorspellen gebruik komt van Twitter. Er zal opgehelderd worden hoe nauwkeurig voorspeld kan worden, of een nummer in de top 40 een week later een hogere of lagere positie in die lijst heeft. Dit maakt het een binair classificatie probleem. De classificatie algoritmen zijn K-Nearest-Neighbour en Random Forest, dus het wordt ook duidelijk welk van die twee algoritmen voor dit soort problemen het meest geschikt is. Nadat je deze scriptie gelezen hebt, weet je of Twitter de potentie heeft om te voorspellen of de populariteit van muziek een week later gestegen of gedaald is.
Er is al vaker onderzoek gedaan naar de voorspellende gave van social media (Asur and A. Huberman, 2010; Bothos et al., 2010). Dat er veel informatie uit te halen valt, bewijst een onderzoek waarbij de sterkte van een band tussen 2 mensen bepaald wordt (Gilbert and Karahalios, 2009). Ze hebben vastgesteld dat dat met een nauwkeurigheid van meer dan 85% kan. Ook is geprobeerd met social media de uitslag van een verkiezing to voorspellen (Tumasjan et al., 2010). Het lastige aan het voorspellen van muziek is, dat er zoveel verschillende mensen zijn met verschillende smaken (Ter Bogt et al., 2003). De top 40 bestaat dus eigenlijk uit een mix van dat soort smaken. Voorspellen hoe die mix eruit gaat zien is nog niet gedaan met behulp van social media.

# 3 Data

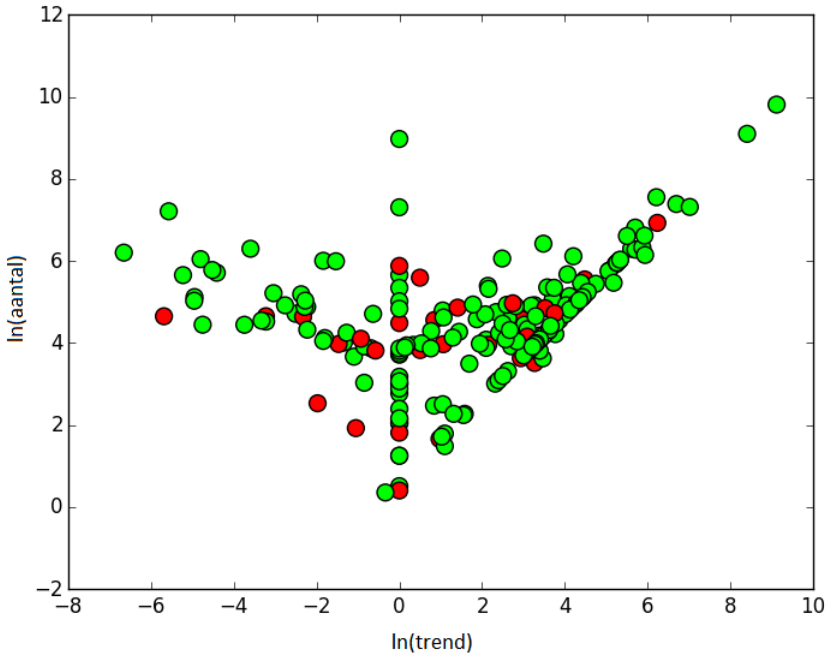
Er is data vanaf week 38 in 2014 data beschikbaar. In dit onderzoek is de data tot en met week 2 in 2015 gebruikt. Die data bevat, per nummer uit de top 40, hoeveel tweets er per dag verstuurd zijn die de titel en artiest van dat nummer bevatten. Ook is de inhoud van de tweets opgeslagen, maar daar is in dit onderzoek niets mee gedaan. Op het moment van schrijven zijn er ongeveer 600 datapunten beschikbaar. Daar komen nog ongeveer 250 datapunten uit de tipparade bij. De tipparade is een lijst van nummers waarvan wordt verwacht dat ze in populariteit stijgen.
Een nadeel van de data uit de top 40 is, dat er alleen data van nummers verzameld wordt, die al in de top 40 (of tipparade) staan. Elk datapunt krijgt het label stijgen of dalen. Een nummer daalt als de positie lager is dan in de week ervoor, óf uit de top 40 verdwijnt. Het stijgt als de positie de komende week hoger is, maar niet als het de top 40 binnenkomt, omdat er dan nog geen data over beschikbaar is. Dit zorgt ervoor dat er meer nummers in de data zitten die het label dalen hebben dan die met het label stijgen.

## 3.1 Tipparade

In de tipparade is het andersom, hier zijn er meer nummers met het label stijgen. Dit komt doordat de tipparade zo samengesteld wordt, dat er geen nummer is dat de week erna een lagere positie heeft en dus allen kan stijgen of uit de lijst verdwijnen. Dit zorgt ervoor dat er weinig nummers zijn die het label dalen hebben, zie figuur 1 en 2 (rood betekent dalen, groen stijgen).




Figuur 1: De top 40, data vanaf 13-10-2014 tot 26-01-2015. Elk punt staat voor een nummer in een bepaalde week. Een rode stip betekent dat het nummer de week erna daalde, een groene dat het de week erna steeg in de top 40. De Y-as (ln(aantal)) geeft de logaritme van het aantal tweets over dat nummer in die week weer. De X-as (ln(trend)) staat voor de logaritme van de trend van het aantal tweets, hoe dat berekend wordt staat in hoofdstuk 3.4.2.




Figuur 2: De tipparade, data vanaf 13-10-2014 tot 26-01-2015. Elk punt staat voor een nummer in een bepaalde week. Een rode stip betekent dat het nummer de week erna daalde, een groene dat het de week erna steeg in de top 40. De Y-as (ln(aantal)) geeft de logaritme van het aantal tweets over dat nummer in die week weer. De X-as (ln(trend)) staat voor de logaritme van de trend van het aantal tweets, hoe dat berekend wordt staat in hoofdstuk 3.4.2.

## 3.2 Herkomst

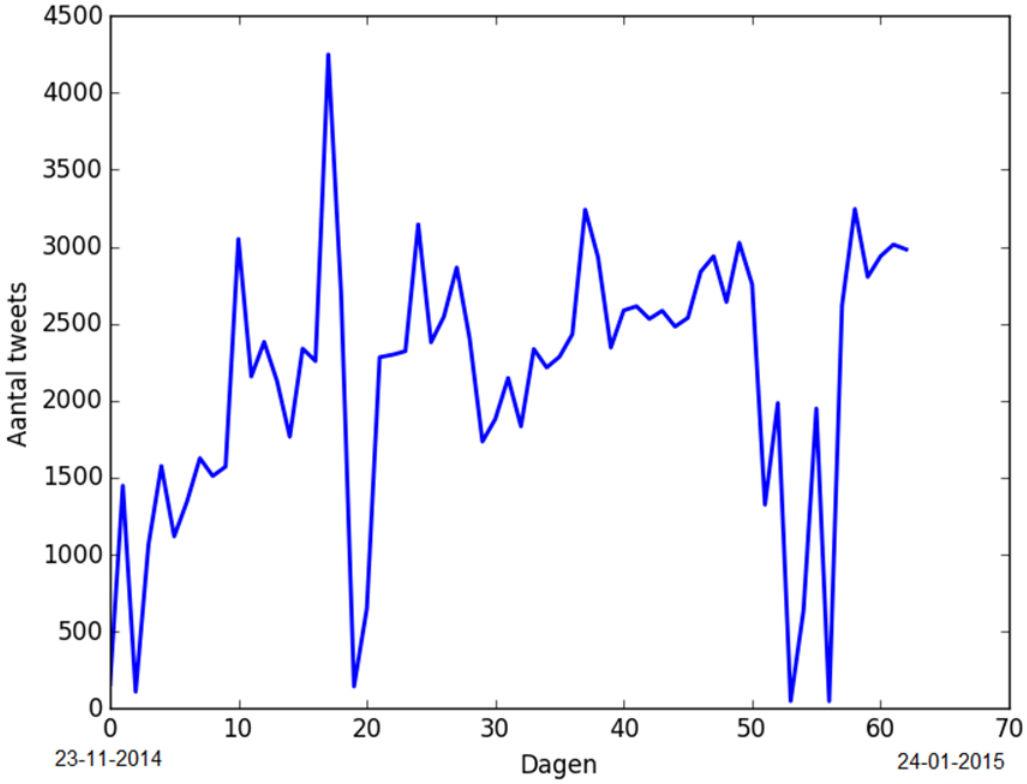
De gebruikte data werd verzameld door Orikami (Orikami, 2014). Elke week werd er een nieuwe top 40 van de officiële website gehaald (`http://www.top40.nl/`). Op basis daarvan is er gezocht naar tweets die een van die nummers uit de top 40 bevatten. Dat gebeurde op basis van titel en artiest. Hetzelfde geldt voor de tipparade, 30 nummers die niet in de top 40 staan maar wel kanshebber zijn om erin te komen. De volgende data is beschikbaar (vanaf week 38 in 2014):

- De top 40 lijsten
- De tipparades

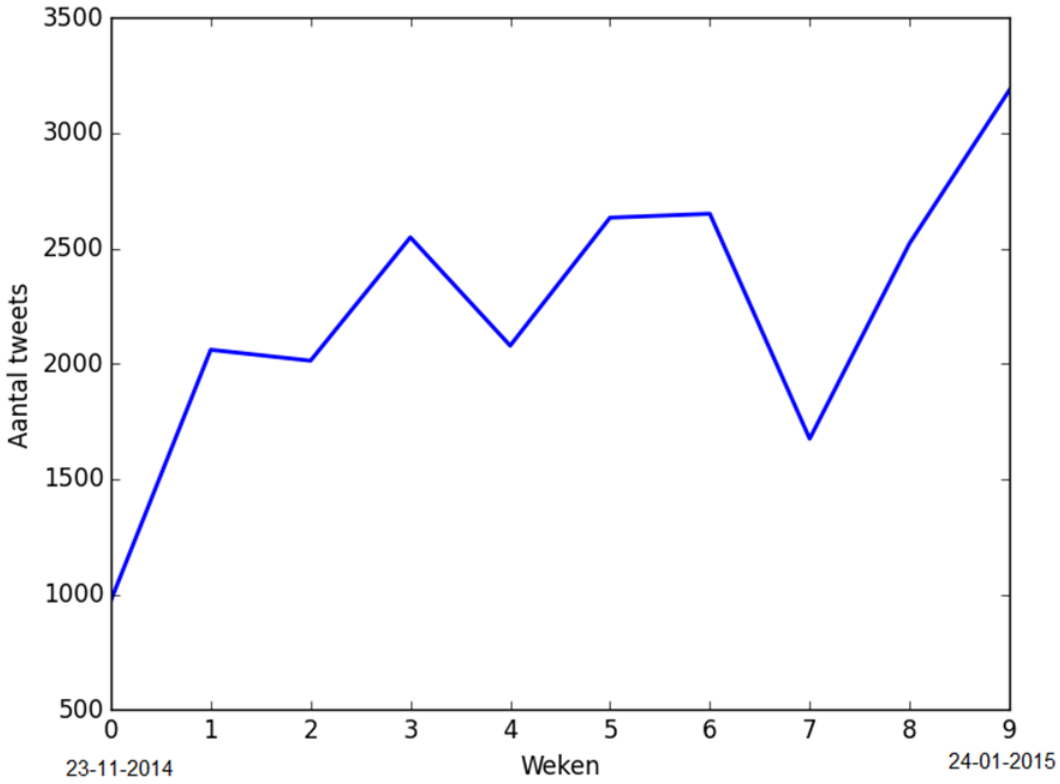
- Alle gevonden tweets die een nummer uit de top40/tipparade bevatten
- Alle nummers die sinds het verzamelen in de top40/tipparade stonden

### 3.3 Data per week

Het aantal tweets is beschikbaar op dagniveau. Ik heb ervoor gekozen om voor alle features het gemiddelde aantal tweets per week te gebruiken, met andere woorden het weekniveau te gebruiken. Hierdoor maakt het niet uit als het zo zou zijn dat mensen op bijvoorbeeld zondag meer tweeten dan op andere dagen. Als er voor een bepaalde week geen zeven dagen beschikbaar zijn, wordt het gemiddelde van de beschikbare dagen genomen, en niet die van de hele week. In dat geval zouden de overige dagen 0 tweets hebben, wat het gemiddelde naar beneden haalt.
In figuur 3 en 4 is het verschil tussen weken en dagen goed te herkennen. De grafiek die de tweets per dag weergeeft, is veel onrustiger, waardoor het lastig is te zien of het aantal tweets nu stijgt of daalt.




Figuur 3: Het aantal tweets per dag over het nummer "Take me to church"van Hozier. De Y-as geeft het aantal tweets weer die op die dag verzonden zijn en waar de titel en artiest in voorkwamen. De X-as geeft het aantal dagen sinds het begin van de meting aan. Het is goed te zien dat deze grafiek zeer onrustig is vergeleken met de grafiek die het aantal tweets op weekbasis laat zien (figuur 4).




Figuur 4: Het aantal tweets per week over het nummer "Take me to church"van Hozier. De Y-as geeft het aantal tweets weer die in die week gemiddeld per dag verzonden zijn en waar de titel en artiest in voorkwamen. De X-as geeft het aantal weken sinds het begin van de meting aan. Het is goed te zien dat deze grafiek rustig verloopt vergeleken met de grafiek die het aantal tweets op dagbasis laat zien (figuur 3).

### 3.4 Samenstelling

Na het verwerken van alle data bevat een datapunt de volgende elementen:

- klasse label
- features
- titel/artiest

- week
- positie

### 3.4.1 Klasse label

Elk datapunt krijgt het label **stijgen** of **dalen**, afhankelijk van de situatie (tabel 1).

| Lijst | Situatie | Klasse label |
|---|---|---|
| Top 40 | De week erna een hogere positie | **stijgen** |
| Top 40 | De week erna een lagere positie | **dalen** |
| Top 40 | De week erna niet meer in de lijst | **dalen** |
| Tipparade | De week erna een hogere positie | **stijgen** |
| Tipparade | De week erna niet meer in de lijst, ook niet in de top 40 | **dalen** |
| Tipparade | De week erna in de top 40 | **stijgen** |

Tabel 1: De situaties met bijbehorend klasse label

Het is niet voorgekomen dat er een nummer in de tipparade stond, dat de week erna een lagere positie in de tipparade had. Dus of het kreeg een hogere positie, of het ging door naar de top 40, of het verdween helemaal uit de lijst.
Als een nummer 2 weken achter elkaar op de zelfde positie blijft, wordt het uit de data verwijderd, omdat het niets toevoegt (hoofdstuk 4.1.1).

### 3.4.2 Features

De gebruikte features zijn het aantal tweets en de trend van het aantal tweets, gebaseerd op tweets over de hele wereld. Daaraan worden het aantal en de trend van de Nederlandse tweets nog toegevoegd.

**Aantal**
Deze waarde is de logaritme van het aantal tweets in de laatste week. De logaritme wordt toegepast, zodat het verschil tussen 100 en 1100 tweets groot is, maar tussen 18000 en 19000 niet, ondanks dat het verschil bij beide 1000 tweets is.

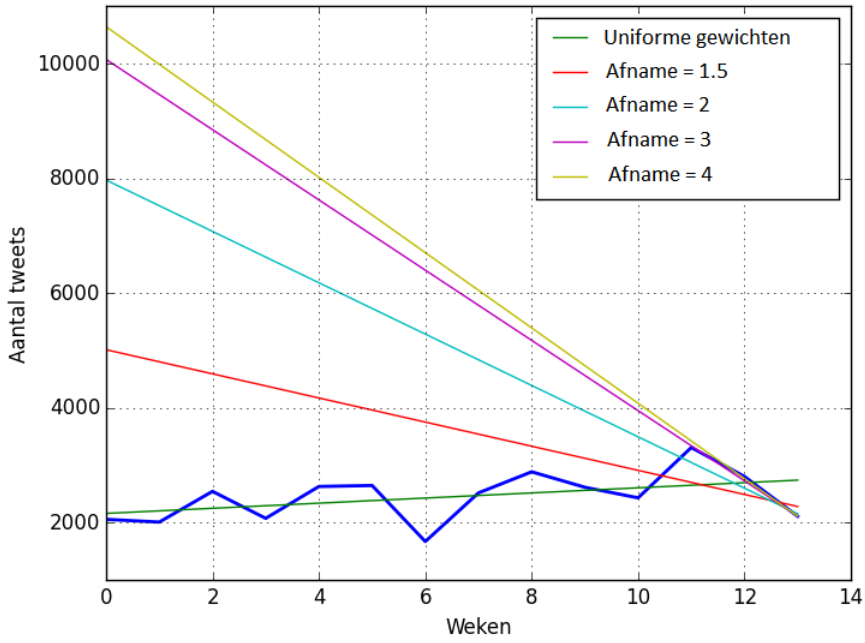
**Trend**
De trend geeft aan of het aantal tweets de laatste weken aan het stijgen of aan het dalen is. Om dit aan te geven heb ik *curve fitting* toegepast.

Daarmee heb ik een lineaire vergelijking gegenereerd, waarbij de logaritme van de gradiënt van deze lijn de feature is. Indien de gradiënt negatief is, wordt die waarde met -1 gemultipliceerd, daar de logaritme van genomen, en de uitkomst weer negatief gemaakt. Bij het genereren van deze lijn zijn er gewichten aan elke week toegekend, zodat de laatste week/weken meer mee tellen dan de eerste weken. De laatste week kreeg gewicht 1, elke andere week kreeg het gewicht van de week erna, gedeeld door een constante. Die constante staat voor de mate van afname. Een hoge constante betekent dus, dat de gewichten sterker variëren. In tabel 2 is te zien welke invloed de afname op de accuratesse heeft.

| Afname | Accuratesse met KNN als classifier |
|---|---|
| 1 | 70% |
| 1.5 | 71% |
| 2 | 71% |
| 3 | 72% |
| 4 | 69% |
| 5 | 69% |

Tabel 2: De accuratesse met verschillende waardes voor de afname. Uitgevoerd op data van 03-11-2014 tot 04-03-2015. De data bevatte geen tipparade en er is geen preprocessing op uitgevoerd.

Voor dit onderzoek zal ik de waarde drie gebruiken om de gewichten voor de trend te berekenen. Figuur 5 toont wat het effect van verschillende gewichten is.




Figuur 5: Vergelijking tussen verschillende gewichten, data vanaf 23-11-2014 tot 05-03-2015 over het nummer "Take me to church"van Hozier. De blauwe lijn geeft het aantal tweets in een bepaalde week weer, de andere lijnen staan voor de trendlijn met verschillende gewichten. De waarde van de afname geeft aan hoe zwaar de weken voor de laatste week meetellen in de berekening van de trend. De uiteindelijke trend feature is de logaritme van de gradiënt van een lijn. Indien het een daling is, dus de gradiënt negatief is, wordt die waarde met -1 gemultipliceerd, daar de logaritme van genomen, en de uitkomst weer negatief gemaakt.

### 3.4.3 Week

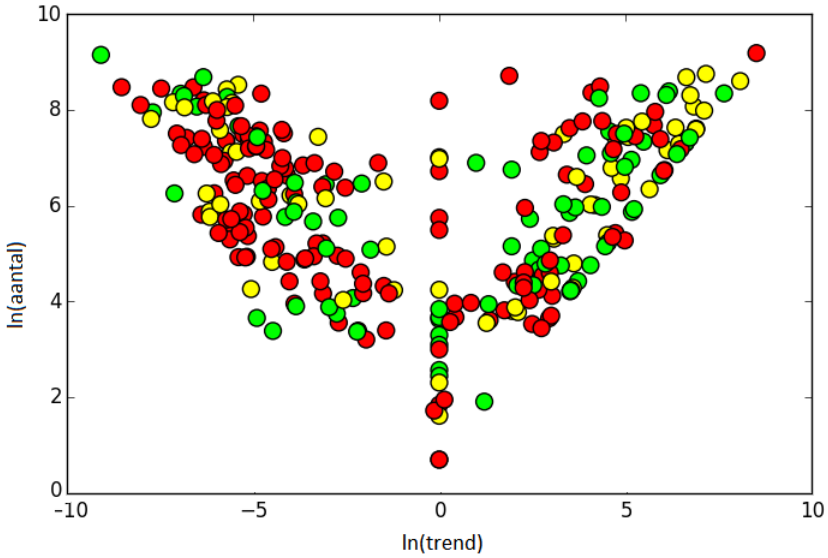
Als een nummer meerdere weken in de top 40 staat, staat het ook meerdere keren in de data. Een nummer uit week $x$ heeft dus alleen data gebaseerd op wat er die week bekend was. Als hetzelfde nummer er in week $x+1$ ook nog in staat, zijn het twee afzonderlijke datapunten. Een van die twee datapunten heeft dan een week meer data.

# 4 Methoden

Ik heb alle code in Python geschreven. Voor K-Nearest-Neighbours, Random Forest en normalization heb ik bestaande code gebruikt en komt van scikit-learn (Scikit-learn, 2015).

## 4.1 Pre-processing

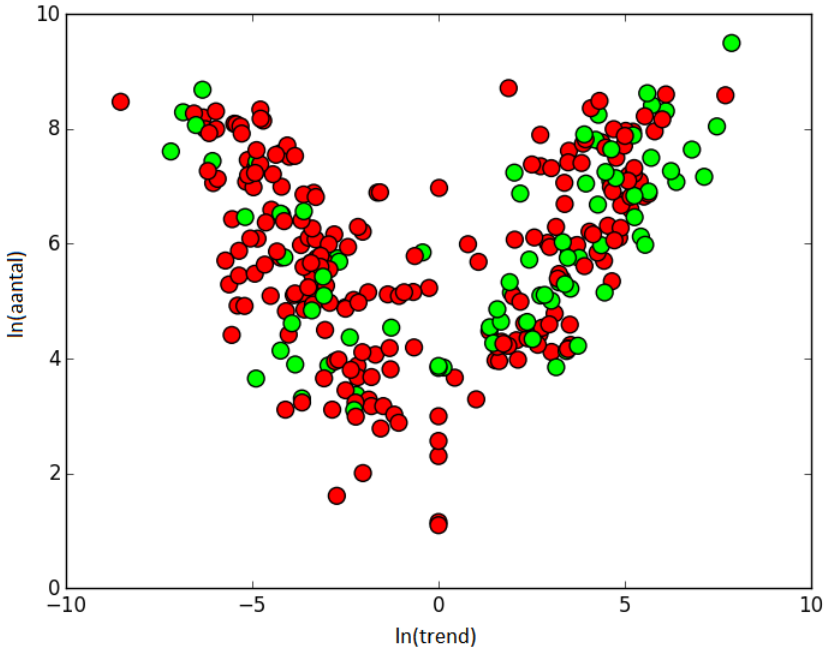
Doordat er niet zeer veel data beschikbaar is, zijn veel pre-processing algoritmen ongeschikt, doordat ze data verwijderen. Figuur 6 laat zien hoe de data er zonder pre-processing eruitziet. Op de assen zijn twee features weergegeven, zie hoofdstuk 3.4.2. Een groen datapunt betekent dat het nummer de week erna een betere positie in de top 40 heeft, geel betekent dat de positie gelijk blijft, en rood betekent een slechtere positie in de top 40 de week erna.




Figuur 6: De data zonder preprocessing, vanaf 13-10-2014 tot 26-01-2015. Elk punt staat voor een nummer in een bepaalde week. Een rode stip betekent dat het nummer de week erna daalde, een gele dat het een week later dezelfde positie heeft en een groene stip dat het de week erna steeg in de top 40. De Y-as (ln(aantal)) geeft de logaritme van het aantal tweets over dat nummer in die week weer. De X-as (ln(trend)) staat voor de logaritme van de trend van het aantal tweets, hoe dat berekend wordt staat in hoofdstuk 3.4.2.

### 4.1.1 Gelijkblijvers verwijderen

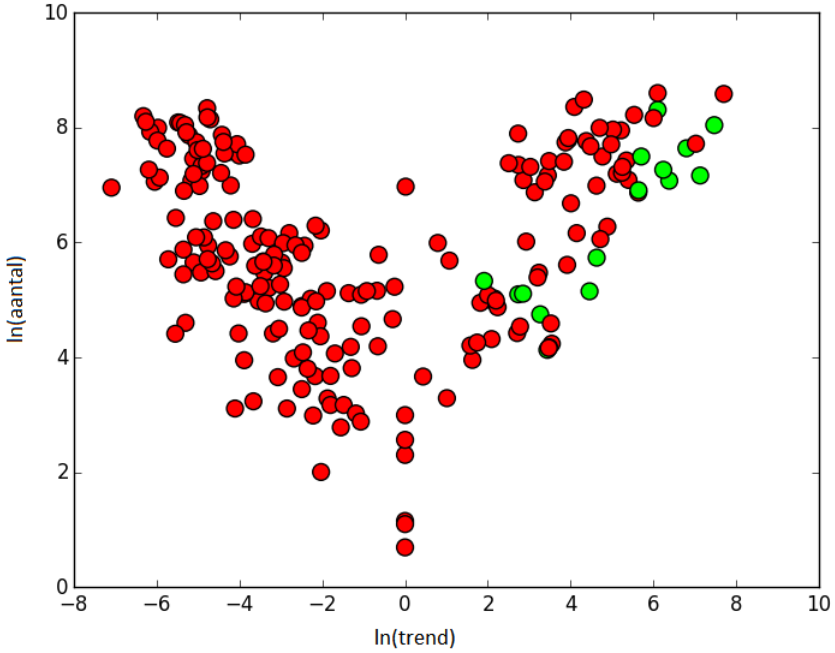
Er zijn veel nummers, die twee weken achter elkaar op dezelfde positie in de top 40 staan, en dus niet als stijger of als daler geclassificeerd kunnen worden. Om het probleem eenvoudiger te maken, worden ze uit de data verwijderd (figuur 7). Hierdoor is het een binair probleem, wat de accuratesse ten goede komt, zonder dat er cruciale data verloren gaat. Doel van dit onderzoek is immers de stijgers en dalers identificeren, niet de nummers waarvan de populariteit gelijk blijft.




Figuur 7: De data nadat de gelijkblijvers eruit zijn gehaald, vanaf 13-10-2014 tot 26-01-2015. Elk punt staat voor een nummer in een bepaalde week. Een rode stip betekent dat het nummer de week erna daalde, een groene dat het de week erna steeg in de top 40. De Y-as (ln(aantal)) geeft de logaritme van het aantal tweets over dat nummer in die week weer. De X-as (ln(trend)) staat voor de logaritme van de trend van het aantal tweets, hoe dat berekend wordt staat in hoofdstuk 3.4.2.

### 4.1.2 Wilson's Algorithm

Wilson's Algorithm is een algoritme dat data verwijdert. Het past voor elk datapunt in de trainset K-Nearest-Neighbour toe, en kijkt vervolgens of het correct geclassificeerd zou worden. Is dat niet het geval, wordt het uit de data verwijderd (Wilson and Martinez, 2000).
Het resultaat hiervan is dat alleen die datapunten overblijven, die heel duidelijk tot een bepaalde klasse behoren. Hierdoor is er minder ruis in de dataset. Een nadeel is echter dat het veel data weggooit, weergegeven in figuur 8.



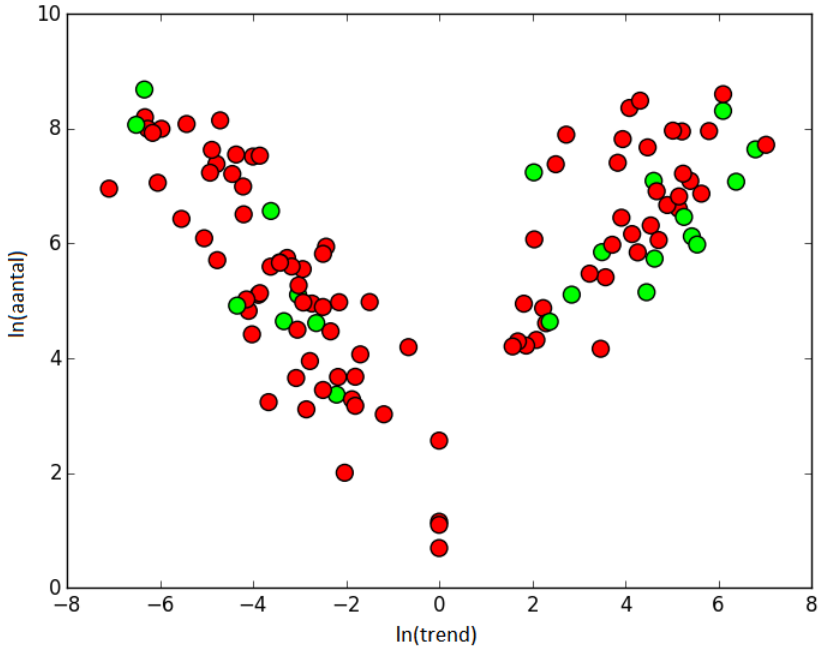

Figuur 8: De data na het toepassen van Wilson's Algorithm, vanaf 13-10-2014 tot 26-01-2015. Elk punt staat voor een nummer in een bepaalde week. Een rode stip betekent dat het nummer de week erna daalde, een groene dat het de week erna steeg in de top 40. De Y-as (ln(aantal)) geeft de logaritme van het aantal tweets over dat nummer in die week weer. De X-as (ln(trend)) staat voor de logaritme van de trend van het aantal tweets, hoe dat berekend wordt staat in hoofdstuk 3.4.2. Er zijn voornamelijk rode punten, omdat er voor het preprocessen meer nummers waren, die de week erna daalden en Wilson's algoritme dit juist versterkt.

### 4.1.3 Normalisatie

Normalisatie zorgt ervoor dat de waarden van alle features tussen de 0 en 1 liggen. Dit is vooral bij K-Nearest-Neighbour nodig, omdat het de dichtst bijzijnde datapunten gebruikt. Als bijvoorbeeld de waarden van feature $X$ tussen de 0 en 1 liggen, en die van feature $Y$ tussen de 0 en 1000, zou $X$ veel minder belangrijk zijn dan $Y$.

### 4.1.4 Alleen grote stijgers/dalers gebruiken

Er zijn nummers die een week later licht zijn gestegen of gedaald, zeg één tot drie posities. De populariteit van die nummers is minder veranderd dan die van nummers die meer dan tien posities gedaald/gestegen zijn. Door alleen die nummers in de data op te nemen, die een x aantal posities in de top 40 gezakt of juist gestegen zijn, blijven er alleen datapunten over waarbij de populariteit echt veranderd is.
Het idee is dat het de grens tussen de twee klassen duidelijker wordt. Ook dit is een algoritme dat het aantal datapunten vermindert, en is dus alleen bruikbaar als er genoeg data beschikbaar is. In figuur 9 is te zien hoe de data met alleen de grote stijgers en dalers eruit ziet.




Figuur 9: De data van 13-10-2014 tot 26-01-2015 met alleen nummers erin die 5 of meer posities gestegen of gedaald zijn. Het doel hiervan is dat het de grens tussen de twee klassen duidelijker wordt. Elk punt staat voor een nummer in een bepaalde week. Een rode stip betekent dat het nummer de week erna daalde, een groene dat het de week erna steeg in de top 40. De Y-as (ln(aantal)) geeft de logaritme van het aantal tweets over dat nummer in die week weer. De X-as (ln(trend)) staat voor de logaritme van de trend van het aantal tweets, hoe dat berekend wordt staat in hoofdstuk 3.4.2.

## 4.2 Classifiers

### 4.2.1 K-Nearest-Neighbour

K-Nearest-Neighbour werkt als volgt. Voor elk te classificeren nummer $m$, kijkt het algoritme welke van de datapunten, waarvan het label al bekend is (datapunten uit de trainset), het meest op $m$ lijken. Hiervoor berekent het de afstand tussen de features van $m$ en die van de datapunten uit de trainset. Een korte afstand betekent dat ze op elkaar lijken. De kortste $k$ afstanden worden geselecteerd en de bijbehorende datapunten zijn de buren van $x$, waarbij $k$ zelf gekozen mag worden. Het label dat door de meeste buren vertegenwoordigd wordt, is het label dat aan $x$ gegeven wordt.

Ik heb hiervoor gekozen omdat deze methode zeer eenvoudig is, maar toch relatief snel resultaten oplevert (Lim et al., 2000). Voor $k$ heb ik verschillende waarden gekozen, afhankelijk van welke preprocessing methoden toegepast worden (tabel 3). De optimale waarde van $k$ is afhankelijk van welke preprocessing methode gebruikt wordt (hoofdstuk 4.1), omdat de meeste pre-processing methoden die geïmplementeerd zijn heb data verwijderen. Doordat er niet veel data beschikbaar is, mag $k$ niet te hoog zijn. In dat geval zou het vaak het label toekennen, dat het vaakst vertegenwoordigd wordt.
De beste waarden voor $k$ zijn negen voor de data zonder preprocessing, en vijf voor de data nadat Wilson's algoritme toegepast is, of alleen de grote stijgers en dalers bekeken worden (tabel 3).

| Waarde van $k$ | Zonder pre-processing | Wilson's Algorithm | Grote stijgers/dalers |
|---|---|---|---|
| 3 | 63.2% | 69.9% | 63.1% |
| 5 | 65.6% | 70.2% | 65.7% |
| 7 | 66.3% | 67.3% | 64.6% |
| 9 | 66.5% | 68.5% | 64.6% |
| 11 | 65.8% | 67.4% | 63.8% |
| 13 | 65.9% | 67.0% | 64.0% |

Tabel 3: De gemiddelde accuratesse met K-Nearest-Neighbour met verschillende waarden voor $k$, data van 03-11-2014 tot 04-03-2015 en zonder tipparade

### 4.2.2 Random Forest

Het Random Forest algoritme creëert meerdere beslissingsbomen, waar ook toeval een rol in speelt. Daarna laat het algoritme alle bomen het gewenste element classificeren. Het label dat door de meeste bomen gekozen wordt, is het uiteindelijke resultaat. Ik heb ervoor gekozen 100 bomen te genereren. Hierdoor is het algoritme snel, en heeft meer bomen gebruiken geen duidelijk effect meer op de nauwkeurigheid (Liaw, Andy and Wiener, Matthew, 2002). Deze methode heeft meerdere voordelen. Zo traint de classifier zeer snel, en duurt het evalueren ook niet lang. Daarbuiten is het ook uiterst geschikt voor voorspellingen (Breiman, 2001; Ho, 1995).

## 4.3 Validatie

Bij het genereren van de train- en testsets is het van belang dat een nummer niet in de trainset én in de testset terecht komt. Een nummer staat er meerdere keren in, als het ook meerdere weken in de top 40 heeft gestaan. Hierdoor staat het week $x$ in de data, maar ook voor week $x+1$ en $x+2$. Deze datapunten lijken veel op elkaar, maar zijn niet hetzelfde. Het nummer uit week $x$ heeft namelijk alleen die data die in week $x$ bekend was. Door deze nummers óf in de trainset óf in de testset te stoppen, word het resultaat niet vervalst.

### 4.3.1 Cross-validation

Er is genoeg data beschikbaar om 5-fold cross-validatie toe te passen. Hierdoor krijgen we een beter beeld van hoe nauwkeurig de voorspelling is. Daarbuiten is ook gelijk de variantie beschikbaar, waarmee we kunnen zien of het gemiddelde van de vijf classificaties representatief is, zonder het algoritme vaak uit te voeren.

# 5 Resultaten

Normalisatie heb ik niet toegepast, omdat de waarden van alle features al tussen de -10 en 10 liggen. Ik heb na diverse testen dus ook geen verschil in nauwkeurigheid opgemerkt. Bij alle testen zijn wel de gelijkblijvers eruit gehaald, omdat steeds een beter resultaat opleverde, zonder echt iets uit te hoeven rekenen.
De accuratesse per dataset en classificatiemethode wordt in tabel 4 en 5 weergegeven. In de tabellen betekent ww dat de tweets van over de hele wereld, en nl alleen de tweets afkomstig uit Nederland gebruikt worden. In de kolom baseline staat hoeveel procent van de data tot één bepaalde klasse behoort, dus de accuratesse als alleen maar hetzelfde label voorspeld wordt.

| | **K-Nearest-Neighbour** | | | |
|---|---|---|---|---|
| | Zonder pre-processing | Wilson's Algorithm | Grote stijgers/dalers | Baseline |
| Zonder tip, ww | 68.6% | 69.3% | 66.1% | 71.3% |
| Zonder tip, nl | 72.4% | 72.5% | 68.3% | 71.2% |
| Met tip, ww | 57.6% | 62.1% | 54.9% | 51.1% |
| Met tip, nl | 72.5% | 66.4% | 56.0% | 51.4% |

Tabel 4: De gemiddelde accuratesse met K-Nearest-Neighbour als classifier, data van 03-11-2014 tot 04-03-2015

| | **Random forest** | | | |
|---|---|---|---|---|
| | Zonder pre-processing | Wilson's Algorithm | Grote stijgers/dalers | Baseline |
| Zonder tip, ww | 63.9% | 71.2% | 68.7% | 71.3% |
| Zonder tip, nl | 71.1% | 71.0% | 69.2% | 71.2% |
| Met tip, ww | 66.6% | 62.1% | 53.7% | 51.1% |
| Met tip, nl | 71.1% | 66.2% | 55.0% | 51.4% |

Tabel 5: De gemiddelde accuratesse met Random Forest als classifier, data van 03-11-2014 tot 04-03-2015

## 5.1 K-Nearest-Neighbour

Welke waarde K heeft is belangrijk, omdat het moet schalen met de hoeveelheid data. Als die waarde te hoog is, zou het altijd hetzelfde label opleveren, doordat er te weinig datapunten met het andere label zijn. Zo heb ik na het

toepassen van Wilson's algorithm $k$ de waarde vijf gegeven (zie figuur 8). $k$ was ook vijf nadat alleen de grote stijgers en dalers geselecteerd waren. Voor de tipparade en de data zonder preprocessing heb ik $k$ de waarde negen gegeven, dat gaf de beste resultaten (tabel 3).

### 5.1.1 Zonder preprocessing

KNN uitgevoerd op data waarop geen preprocessing was uitgevoerd, gaf een nauwkeurigheid van ongeveer 66.5% voor de wereldwijd verzamelde tweets, en 72.5% voor de tweets uit Nederland. In beide gevallen lag de baseline, dus de accuratesse als alleen maar dalen voorspeldt wordt, hoger, respectievelijk 72% en 73%. Dit betekent dat op deze manier de populariteit van muziek niet goed genoeg voorspeld kan worden.
Met de data van de tipparade erbij ziet het er echter anders uit. Hier ligt de accuratesse lager (62.5% wereldwijd en 66% Nederlands), echter ligt de baseline ook een stuk lager, rond de 53%. Om deze data voor zinvolle toepassingen te gebruiken, is een hogere accuratesse noodzakelijk.

### 5.1.2 Wilson's Algorithm

De voorspellingen van de data, nadat Wilson's algoritme erop toegepast is, lijken qua nauwkeurigheid erg op die van de data zonder preprocessing. Het grootste verschil is dat er bij de data zonder de tipparade, maar wel de tweets over de hele wereld een betere voorspelling is gedaan (ongeveer 70% tegenover 66.5% zonder preprocessing).
Ook hier is er een duidelijk verschil tussen de voorspellingen als er alleen naar Nederlandse tweets, of naar alle tweets gekeken wordt. Als alleen de Nederlandse tweets bekeken worden, ligt de accuratesse drie tot vier procentpunten hoger.

### 5.1.3 Grote stijgers/dalers

In deze set zijn alleen nummers opgenomen, die meer dan vier posities gedaald of gestegen zijn ten opzichte van de week ervoor. Bij minder dan vier had het weinig tot geen effect, en bij meer dan vier bleef er te weinig data over om er voorspellingen mee te doen.
Op basis van de wereldwijde tweets doet KNN het slecht, zowel met als zonder tipparade. Zonder tipparade scoort het ongeveer 66% (tegenover 71% met alleen Nederlandse tweets) en met tipparade 53.5% (tegenover 57% met alleen Nederlandse tweets).

## 5.2 Random Forest

Het aantal bomen in random forest 100. Hierdoor is het algoritme snel, en heeft meer bomen gebruiken geen duidelijk effect meer op de nauwkeurigheid (Liaw, Andy and Wiener, Matthew, 2002). Random forest doet het over het algemeen beter dan K-Nearest-Neighbour, maar scoort niet significant hoger. De accuratesse is nergens meer dan drie procent hoger dan die van KNN.
Zonder preprocessing en zonder de data van de tipparade ligt nauwkeurigheid bij ongeveer 67% en 71% voor respectievelijk wereldwijde en Nederlandse tweets. Die waarden zijn bijna hetzelfde als die van KNN, plus minus een procent.
Als de tipparade wél voor de classificatie gebruikt wordt, is de nauwkeurigheid bij de Nederlandse tweets weer vergelijkbaar, maar scoort KNN bij de wereldwijde tweets ongeveer drie procent beter.

### 5.2.1 Wilson's Algorithm

Als Wilson's algoritme toegepast wordt, doet random forest het beter dan KNN. Als er alleen naar de Nederlandse tweets gekeken wordt scoort random forest maximaal twee procent, bij de wereldwijde tweets maximaal drie procent hoger.
Net als bij KNN ligt de accuratesse net onder de baseline als er niet naar de tipparade gekeken wordt en anders meer dan tien procent hoger.

### 5.2.2 Grote stijgers/dalers

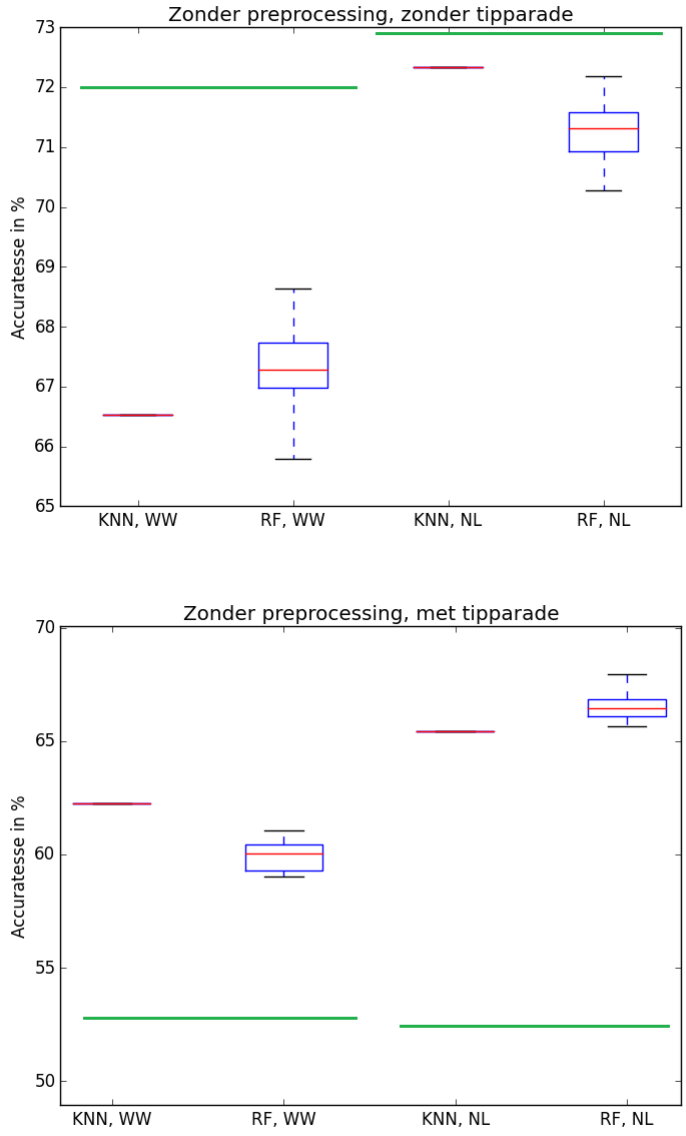
Als alleen naar de grote stijgers en dalers gekeken wordt, doet random forest het niet goed, en scoort onder de baseline als de data van de tipparade niet meegeneomen wordt. Bij de Nederlandse tweets iets onder, bij de wereldwijde tweets ver onder de baseline. Net als KNN doet random forest het ook slecht als de data van de tippparade wél bekeken wordt, namelijk minder dan 60%. Dat is te weinig om er iets mee te gaan doen.

# 6 Conclusie

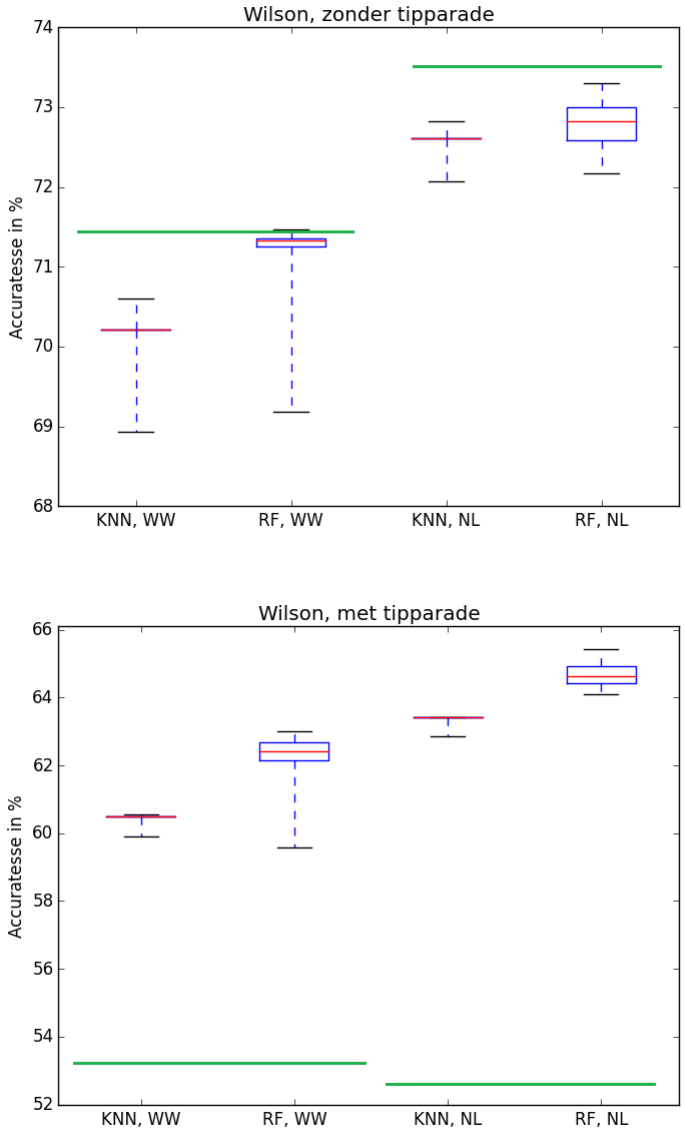
Een accuratesse van ongeveer 70% is veel te laag om er een nuttige toepassing voor te vinden, vooral als de baseline ook rond de 70% ligt. Wat dat betreft gaf de data met de tipparade en zonder preprocessing het beste resultaat, rond de 70% maar met een baseline van ongeveer 50%. Dit zou gebruikt kunnen worden om een richting aan te geven, maar zeker niet om een consumentenonderzoek te vervangen. Daarvoor is het niet nauwkeurig genoeg.
Van de twee classificatiemethoden werkt random forest voor deze toepassing vaak iets beter, ook al is het verschil niet groot. Van de twaalf verschillende tests deed random forest het zes keer duidelijk beter dan K-Nearest-Neighbour, en andersom deed KNN het maar twee keer beter. De overige vier keer deden ze het ongeveer even goed, met een marge van 1%.
Voor deze studie had ik voor ongeveer een half jaar aan data beschikbaar. Het resultaat zou waarschijnlijk nauwkeuriger worden als er jaren aan data beschikbaar is. Daardoor zou ook na het preprocessen genoeg data overblijven om classificatie succesvol toe te passen. Daarbuiten zijn er ook veel andere classificatiemethoden waarmee eventueel een beter resultaat kan worden behaald (Lim et al., 2000).
Ook zouden er andere sociale platformen gebruikt kunnen worden, zoals Facebook, Youtube, Spotify of last.fm. Hierdoor zou er meer en misschien nauwkeurigere data beschikbaar komen. In een ander onderzoek kan worden bepaald welke van die platformen het beste resultaat oplevert, en dat niet alleen op muziekgebied, maar ook bijvoorbeeld voor films. Die vertaalslag is niet moeilijk te maken, als er maar over getwitterd wordt.

# 7 Appendix

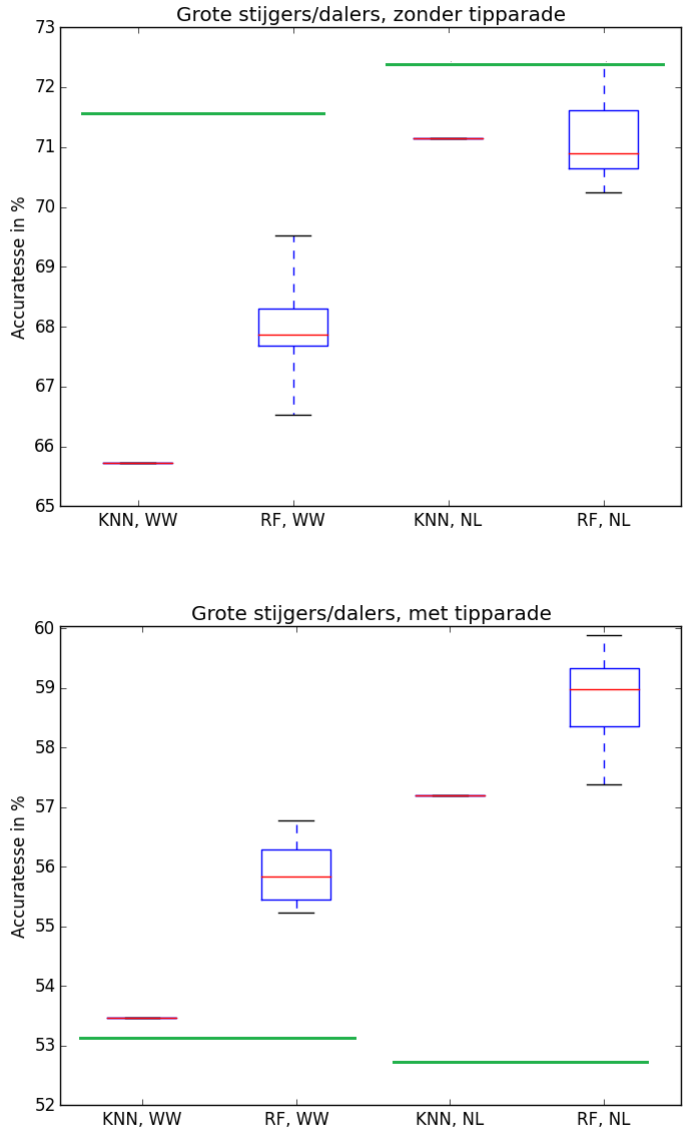
Alle boxplots zijn gemaakt op basis van data van 03-11-2014 tot 17-03-2015




Figuur 10: De resultaten zonder preprocessing, WW betekent dat alle tweets gebruikt, NL dat alleen de Nederlandse tweets bekeken werden. Iedere boxplot is gebaseerd op 20 metingen, bij elke meting werd 5-fold cross-validatie gebruikt. De groene lijn geeft de baseline weer, dus de accuratesse als alleen dalen of stijgen voorspeld wordt (afhankelijk van welke waarde hoger is).




Figuur 11: De resultaten met Wilson's algoritme, WW betekent dat alle tweets gebruikt, NL dat alleen de Nederlandse tweets bekeken werden. Iedere boxplot is gebaseerd op 20 metingen, bij elke meting werd 5-fold cross-validatie gebruikt. De groene lijn geeft de baseline weer, dus de accuratesse als alleen dalen of stijgen voorspeld wordt (afhankelijk van welke waarde hoger is).




Figuur 12: De resultaten met alleen grote stijgers en dalers, WW betekent dat alle tweets gebruikt, NL dat alleen de Nederlandse tweets bekeken werden. Iedere boxplot is gebaseerd op 20 metingen, bij elke meting werd 5-fold cross-validatie gebruikt. De groene lijn geeft de baseline weer, dus de accuratesse als alleen dalen of stijgen voorspeld wordt (afhankelijk van welke waarde hoger is).

## Referenties


Asur, S. and A. Huberman, B. (2010). Predicting the future with social media.

Bothos, E., Apostolou, D., and Mentzas, G. (2010). Using social media to predict future events with agent-based markets.

Breiman, L. (2001). Random forests.

Gilbert, E. and Karahalios, K. (2009). Predicting tie strength with social media.

Ho, T. K. (1995). Random decision forests.

Liaw, Andy and Wiener, Matthew (2002). Classification and regression by randomforest.

Lim, T.-S., Loh, W.-Y., and Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms.

Orikami (2014). `http://orikami.nl/`.

Scikit-learn (2015). `http://scikit-learn.org/stable/`.

Stichting Nederlandse Top 40 (2014). `http://www.top40.nl/samenstelling`.

Ter Bogt, T., Raaijmakers, Q., Vollebergh, W., Van Wel, F., and Sikkema, P. (2003). Youngsters and their musical taste: Musical styles and taste groups.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment.

Twitter, I. (2015). `https://about.twitter.com/company`.

Wilson, D. R. and Martinez, T. R. (2000). Reduction techniques for instance-based learning algorithms.