Sentimenteren

Het bepalen en toepassen van sentiment

oktober 2011

Versienummer: 1.0

Robbert Segers

Studentnummer: 1517861

E-mailadres: robbert.segers@gmail.com

full service internetbureau.

Colofon

Titel: **Sentimenteren**

Ondertitel: **Het bepalen en toepassen van sentiment**

Auteur: Dhr. Robbert Segers

Opdrachtgever: Oxin Full Service Internetbureau

Projectbegeleider: Dhr. Henk Venema

Plaats: Wijk bij Duurstede

Datum: oktober 2011

Opleidingsbegeleider: Dhr. Pim Schonk

Download een digitaal exemplaar op:

www.robbertsegers.nl/afstuderen

Robbert Segers / *1517861*

T 06 - 51 43 99 51 **E** robbert.segers@gmail.com

Samenvatting

Aanleiding

Sinds kort is er binnen de onderneming behoefte aan een nieuw en innovatief product dat er voor zorgt dat Oxin een breder aanbod van producten en services kan aanbieden. Oxin houdt zich al enige tijd bezig met sociale media maar heeft hier nog geen toepassingen voor kunnen vinden. Na het bestuderen van de mogelijkheden is gebleken dat er een groeiende behoefte is aan het snel, automatisch en tegen lage kosten bepalen van sentiment.

Onderzoeksmodel

Voor dit onderzoek is er een selectie gemaakt van drie verschillende methoden waarmee het mogelijk is om sentiment te bepalen. Deze methoden zijn onderzocht en de resultaten zijn geanalyseerd en vergeleken. Vervolgens zijn er twee producten op basis van sentiment onderzocht waarvan er één is uitgewerkt in een business model.

Conclusies

Sentiment is een gevoel (emotie) dat betrekking heeft op iets of iemand anders. Met andere woorden: sentiment is de algemene gemoedstoestand over een bepaald onderwerp. Voor vele personen, bedrijven en instanties is sentiment een belangrijke indicator. Op basis van het sentiment worden grote beslissingen genomen. Wanneer er bijvoorbeeld een negatief sentiment heerst over een bepaald product kan er voor worden gekozen om nieuwe campagnes te starten om dit negatieve imago te verbeteren. Sentiment wordt ook gebruikt om bijvoorbeeld de peilingen van politieke partijen te meten, om boekverkopen of filmopbrengsten te voorspellen maar ook zijn er onderzoeken uitgevoerd om aan de hand van sentiment de beurskoers te voorspellen.

Sentiment kan op verscheidene manieren bepaald worden. De meest gangbare manier om sentiment te bepalen is aan de hand van enquêtes. Deze enquêtes zijn duur en tijdrovend. Vanwege de globalisering en de technologische vooruitgang is het tegenwoordig ook mogelijk om op een geautomatiseerde wijze sentiment te bepalen met behulp van computersystemen. Deze computersystemen zijn dusdanig ingericht dat ze teksten taalkundig kunnen ontleden. Ze zijn in staat om de onderwerpen en werkwoorden te identificeren. Maar het belangrijkste is dat ze ook in staat zijn om een tekst (of een gedeelte van een tekst) positief of negatief te classificeren. Door honderden berichten over een bepaald onderwerp door deze computersystemen te laten analyseren kan het sentiment over dit onderwerp vastgesteld worden.

Voor dit onderzoek zijn er drie methoden voor het bepalen van sentiment getest en vergeleken. Elke methode heeft zijn eigen benadering, werkwijze en mogelijkheden. De drie onderzochte methoden zijn: de Bayesian methode, OpinionFinder en Open Amplify. Open Amplify scoort van de drie geteste methoden op alle vergelijkingen het beste. Open Amplify heeft veruit de meeste mogelijkheden, het identificeert als beste de

onderwerpen en acties, kan door de hoge volatiliteit het beste veranderingen in sentiment waarnemen en is met de hoogste nauwkeurigheid als beste in staat om sentiment te bepalen.

Op basis van sentiment, dat is bepaald door Open Amplify zijn er twee mogelijke producten onderzocht die Oxin zou kunnen aanbieden om een breder aanbod te krijgen van producten en services.

Voor het eerste onderzochte product is op basis van sentiment de beurs voorspeld. In de periode van januari 2001 tot en met december 2010 heeft de markt gemiddeld 20 procent rendement weten te behalen. De theorie die is gebaseerd op basis van sentiment heeft in dezelfde periode 135% rendement behaald. Deze theorie zou aan een geïnteresseerde partij verkocht kunnen worden of de maandelijkse resultaten zullen in een abonnementsvorm aangeboden kunnen worden.

Voor het tweede product is er gekeken of sentiment kan worden aangeboden als marketing tool. Oxin verzamelt alle nieuws- en Twitterberichten over een bepaald onderwerp. Alle verzamelde berichten worden verwerkt met Open Amplify en de resultaten worden via een webportal toegankelijk gemaakt. De webportal zal voornamelijk gericht zijn op het sentiment en het verschil van sentiment over een bepaalde periode. Er kan eenvoudig een vergelijking worden gemaakt tussen diverse vormen van sentiment. De resultaten kunnen opgeslagen of geëxporteerd worden. Om toegang te krijgen tot de webportal kunnen bedrijven kiezen uit drie verschillende abonnementsvormen.

Hoewel de resultaten van het eerste product verrassend goed zijn is er toch voor gekozen om het tweede product ‘sentiment als marketing tool’ uit te werken in een business model. Oxin is tenslotte geen financiële instelling maar een full service internetbureau dat veel meer affiniteit heeft met marketing dan met finance. Uit het business model is naar voren gekomen dat dit product zeker door Oxin aangeboden zou kunnen worden. Het product is realistisch, heeft een financiële levensvatbaarheid en de investering van rond de 20.000 euro zal na ongeveer tien maanden zijn terug verdiend.

Inhoudsopgave

Voorwoord

Het rapport dat voor u ligt is het resultaat van vijf maanden onderzoek naar het bepalen van sentiment over Twitterberichten en nieuwsartikelen. Gedurende het gehele project zijn de mogelijkheden onderzocht hoe het bepalen van sentiment als product of dienst aangeboden kan worden. Dit onderzoek is uitgevoerd in het kader van de afstudeeropdracht ter afronding van de studie Digitale Communicatie aan de Hogeschool Utrecht.

Bij de totstandkoming van dit rapport heb ik de nodige steun gehad vanuit diverse hoeken. Twee mensen wil ik via deze weg persoonlijk bedanken. Samen met dhr. Stef Dekker, BSc. heb ik onderzoek gedaan naar de mogelijkheden om aan de hand van sentiment de beurs te voorspellen. Dhr. Henk Venema (programma manager ING) heeft me als projectbegeleider gedurende het gehele project begeleid en op weg geholpen.

Er rest mij niets anders dan u veel plezier te wensen met het lezen van dit rapport.

Wijk bij Duurstede, oktober 2011.

Robbert Segers

1. Inleiding

Het automatisch bepalen en toepassen van sentiment is een ware hype. Vele onderzoekers houden zich met deze materie bezig. Deze onderzoekers willen allemaal weten hoe sentiment bepaald kan worden en wat de resultaten er van zijn. Mede dankzij diverse onderzoeken die hebben aangetoond dat op basis van sentiment significant positieve resultaten behaald kunnen worden op de beurs heeft het bepalen van sentiment een ware boost gekregen. Het bepalen van sentiment heeft hiermee ook de aandacht van Oxin getrokken. In dit rapport wordt voor Oxin de volgende vraag beantwoord: Hoe kan sentiment worden bepaald en op welke wijze kan dit als product worden aangeboden?

Leeswijzer

De opbouw van dit rapport is als volgt: In het eerstvolgende hoofdstuk wordt het onderzoeksontwerp met daarin onder andere het projectkader, de probleemstellingen, de doelstellingen en de onderzoeksvragen nader toegelicht. In dit hoofdstuk moet tevens duidelijk worden welke weg is bewandeld gedurende dit onderzoek om uiteindelijk aan de doelstelling te kunnen voldoen en om de onderzoeksvragen te beantwoorden.

In hoofdstuk drie wordt vervolgens het theoretisch kader beschreven met daarin onder andere een uitleg van wat sentiment is, waarvoor het gebruikt kan worden, hoe het bepaald kan worden, welke van de onderzochte methoden het beste in staat is om sentiment te bepalen en tot slot wordt er gekeken hoe brondata verzameld kan worden.

Na het theoretisch kader wordt er in hoofdstuk vier gewerkt aan twee mogelijke producten op basis van sentiment die door Oxin aangeboden kunnen worden.

In hoofdstuk vijf wordt één van deze twee producten gekozen en verder uitgewerkt in een business model.

Tot slot wordt er in hoofdstuk zes een toekomstperspectief van sentiment beschreven en in hoofdstuk zeven staan de conclusies van dit onderzoek.

In de bijlagen wordt er extra informatie gegeven over OpinionFinder en Open Amplify en worden de technische details beschreven voor de realisatie van dit onderzoek.

2. Onderzoeksontwerp

2.1 Projectkader

Oxin Full Service Internetbureau (Oxin) houdt zich sinds 2007 bezig met het ontwikkelingen van op maat gemaakte websites en -applicaties voor het MKB [1]. Naast de ontwikkeling van deze websites en -applicaties kan Oxin bedrijven van het begin tot het eind ondersteunen met alle internetprojecten. De huidige expertises van Oxin zijn:

- Conceptontwikkeling;
- Het ontwerpen van websites & -applicaties;
- Het ontwikkelen van websites & -applicaties;
- Zoekmachine optimalisatie[2];
- Social media marketing;
- Het leveren van CMS[3];
- Het versturen van digitale nieuwsbrieven[4].

2.2 Probleemstelling

Sinds kort is er binnen de onderneming behoefte aan een nieuw en innovatief product dat er voor zorgt dat Oxin een breder aanbod van producten en services kan aanbieden. Oxin houdt zich al enige tijd bezig met sociale media maar heeft hier nog geen toepassingen voor kunnen vinden. Na het bestuderen van de mogelijkheden is gebleken dat er een groeiende behoefte is aan het snel, automatisch en tegen lage kosten bepalen van sentiment voor o.a. producten en merknamen. Met name door globalisering is er steeds meer nieuws digitaal beschikbaar waardoor het steeds moeilijker wordt voor grote bedrijven om nieuws te interpreteren. Dit creëert echter ook mogelijkheden, bijvoorbeeld: de opkomst van Twitter geeft Oxin een platform om direct en vrijwel kosteloos de meningen van vele klanten te peilen over bepaalde producten en merknamen.

Het bepalen van sentiment en de mogelijkheden om sentiment als product of dienst aan te bieden heeft de interesse gewekt en er toe geleid dat Oxin meer onderzoek wil doen naar dit onderwerp en de mogelijkheden ervan.

---

[1] MBK staat voor midden- en kleinbedrijf, bedrijven met maximaal 250 medewerkers.

[2] Zoekmachine optimalisatie is het geheel aan activiteiten die worden uitgevoerd om een website hoog te laten scoren in de organische (gratis) zoekresultaten van een zoekmachine.

[3] CMS staat voor content management system. Een CMS maakt het mogelijk om zonder technische kennis de inhoud van een website te beheren.

[4] Een digitale nieuwsbrief is een e-mail dat wordt verstuurt vanuit een bedrijf naar één of meerdere relaties.

2.3 Doelstellingen

Dit project heeft als doel sentiment te bepalen van merknamen en producten op een geautomatiseerde wijze, dit moet gedaan worden aan de hand van beschikbare bronnen zoals Twitter en nieuwsartikelen. Hiernaast moet duidelijk worden hoe dit als product kan worden aangeboden zodat Oxin hier positieve bedrijfsresultaten mee weet te behalen.

Deze afstudeeropdracht heeft diverse doelstellingen:

- Oxin een duidelijk inzicht geven in wat de mogelijkheden zijn met betrekking tot het bepalen van sentiment over geschreven tekst;
- Oxin een overzicht geven van mogelijkheden van welke producten er aangeboden kunnen worden op basis van sentiment;
- Het doel van de afstudeeropdracht is om het voor Oxin duidelijk te maken hoe het nieuwe product op basis van sentiment kan worden aangeboden en wat de levensvatbaarheid is van het product.

2.4 Onderzoeksvragen

Voor dit onderzoek zijn er een tweetal centrale onderzoeksvragen geformuleerd.

1. Hoe kan aan de hand van Twitterberichten en nieuwsartikelen sentiment worden bepaald?
2. Hoe kan sentiment als product worden aangeboden?

Voor het beantwoorden van deze hoofdvragen zijn de volgende deelvragen geformuleerd:

1. Wat is sentiment?
2. Waarom is het interessant om sentiment te bepalen?
3. Wat zijn bestaande methoden om sentiment te bepalen over geschreven tekst?
4. Welke methode is het beste geschikt om sentiment te bepalen over geschreven tekst?
5. Welke producten kunnen er worden aangeboden op basis van sentiment?

Beantwoording van de hierboven vermelde hoofd- en deelvragen zal er toe leiden dat de doelstellingen van dit onderzoek worden bereikt.

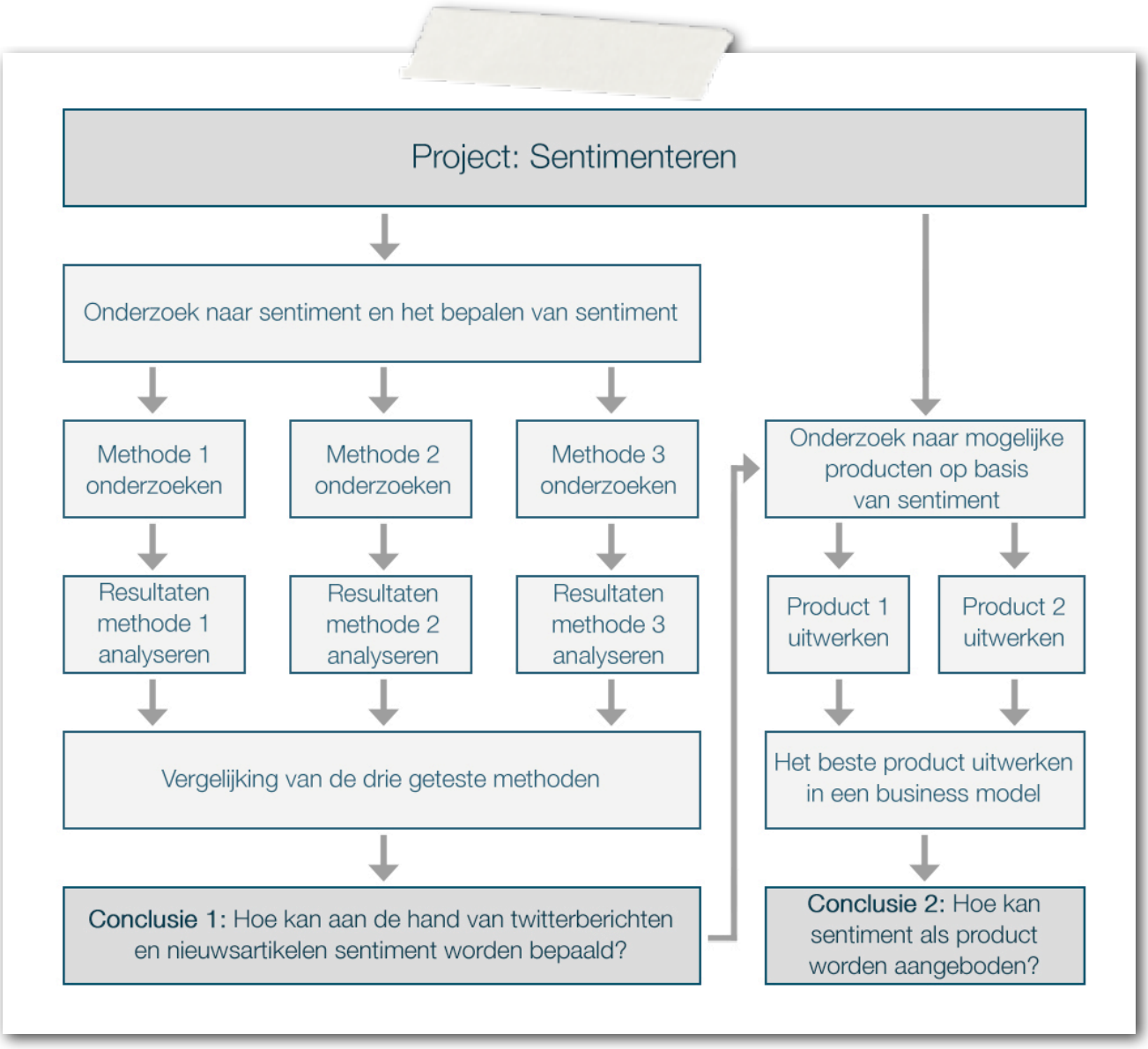
2.5 Afbakening

Zowel de omvang als de diepgang van dit project is zeer omvangrijk. Om niet vast te lopen in de omvang van het project is er besloten om maximaal drie verschillende methoden voor het bepalen van sentiment te onderzoeken en vergelijken. De brondata voor het bepalen van sentiment is beperkt tot Twitterberichten en nieuwsartikelen. Er is onderzoek gedaan naar twee mogelijke producten die op basis van sentiment door Oxin kunnen worden aangeboden. Één van deze producten is uitgewerkt in het business model. Het product dat is uitgewerkt in het business model is (nog) niet technisch gerealiseerd. Deze realisatie valt buiten de scope van dit project.

2.6 Onderzoeksmodel

In de voorbereiding van dit onderzoek bleek al dat er meerdere methoden zijn voor het bepalen van sentiment over geschreven tekst. Voor dit onderzoek is er een selectie gemaakt van drie verschillende methoden waarmee het mogelijk is om sentiment te bepalen. Deze methoden zijn onderzocht en de resultaten zijn geanalyseerd en vergeleken. Hiermee is de eerste onderzoeksvraag beantwoord.

Tijdens het gehele onderzoek naar de mogelijkheden van het bepalen van sentiment is er gekeken naar de mogelijkheden met betrekking tot het aanbieden van een product of dienst op basis van sentiment. De input van het sentiment van deze producten zijn Twitterberichten en nieuwsartikelen die worden verwerkt door één of meerdere van de onderzochte methoden. Om de tweede onderzoeksvraag te kunnen beantwoorden zijn er twee producten op basis van sentiment onderzocht waarvan er één (het beste product dat bij Oxin past) is uitgewerkt in een business model.



Afbeelding 1: Het onderzoeksmodel

3. Theoretisch kader

3.1 Wat is sentiment

Sentiment is een erg breed begrip en heeft verschillende definities. De definitie die wordt aangehouden in dit rapport is: *"Sentiment is een gevoel (emotie) dat betrekking heeft op iets of iemand anders."*. Met andere woorden: sentiment is de algemene gemoedstoestand over een bepaald onderwerp. Ieder onderwerp heeft zijn of haar eigen sentiment, zo kan bijvoorbeeld de president een positief sentiment hebben en het café een negatief sentiment. Sentiment kan worden bepaald over bedrijven, producten, marktsegmenten, personen, evenementen, politieke partijen, televisieprogramma's etcetera.

Sentiment kan worden uitgedrukt in positief, neutraal of negatief. Maar ook in een mate van positiviteit dat bijvoorbeeld varieert van -1.00 tot 1.00 waarbij -1.00 staat voor zeer negatief en 1.00 voor zeer positief.

Momenteel wordt sentiment voornamelijk bepaald aan de hand van enquêtes. Deze enquêtes zijn tijdrovend en brengen hoge kosten met zich mee. Een bijkomend nadeel van enquêtes is dat het moeilijk te controleren is of ze naar waarheid worden ingevuld.

Sentiment wordt altijd gemeten over een bepaalde periode. Één waarde van sentiment van één periode is vaak niet interessant. Het gaat om de verandering van sentiment over een bepaalde periode. Bedrijven willen bijvoorbeeld weten of het sentiment van hun product positiever is geworden na het starten van de marketingcampagnes.

Voornamelijk de laatste twee jaar zijn de technieken steeds verder door ontwikkeld zodat een computer door middel van geschreven tekst kan bepalen wat het sentiment van een onderwerp is. Met geavanceerde taaltechnieken worden zinnen aan de hand van zinconstructies ontleed. Deze nieuwe en opkomende markt heeft voornamelijk als voordeel dat er geen dure en tijdrovende enquêtes afgenomen hoeven te worden en dat de resultaten snel beschikbaar zijn. Deze sentiment analyse wordt ook wel opinion mining genoemd en bepaalt per bericht of het bericht positief, neutraal of negatief is. Door een groot aantal berichten over een bepaald onderwerp te analyseren kan het sentiment van dat onderwerp worden vastgesteld. Door dit sentiment vervolgens dagelijks te bepalen kan het verschil in sentiment over een bepaalde periode worden gemeten.

Om een computer te laten bepalen of een bericht positief of negatief is, is een hele opgave. Het door computers laten bepalen van sentiment is daardoor ook nog niet helemaal perfect. Zo heeft de computer veel moeite met sarcasme en dubbelzinnigheid. Tevens staan er vaak meerdere meningen in een tekst. Bijvoorbeeld: *"Dit is een goed boek maar ik had meer verwacht van deze auteur"*.

Voor het bepalen van sentiment speelt de hoeveelheid brondata een grote rol. Met bijvoorbeeld drie geanalyseerde berichten over de president die allemaal positief zijn kan niet worden gezegd dat het sentiment van de president over het algemeen positief is. Om dit vast te stellen zullen er minstens tientallen berichten

geanalyseerd moeten worden. Hoe meer brondata beschikbaar is hoe zekerder en nauwkeuriger de sentiment analyse wordt. Vanwege de globalisering is er steeds meer brondata digitaal beschikbaar. Met bijvoorbeeld Twitter kunnen vrijwel kosteloos duizenden berichten per dag worden verzameld voor de sentiment analyse.

Het automatisch bepalen van sentiment over geschreven tekst biedt bedrijven de mogelijkheid om op de hoogte te blijven van hoe er over hen wordt gesproken, door wie en op welke platformen.

3.2 Toepassingsgebieden voor sentiment

Voor vele personen, bedrijven en instanties is sentiment een belangrijke indicator. Op basis van het sentiment worden grote beslissingen genomen. Wanneer er bijvoorbeeld een negatief sentiment heerst over een bepaald product kan er voor worden gekozen om nieuwe campagnes te starten om dit negatieve imago te verbeteren.

In de huidige vorm heeft sentiment meerdere toepassingsgebieden, de belangrijkste en meest voorkomende toepassingsgebieden worden hieronder weergegeven:

- Voor bedrijven & marketeers[5] is sentiment belangrijk om te bepalen hoe een product of bedrijf in de markt ligt;
- Voor marketeers wordt sentiment over een bepaalde periode gebruikt om te meten wat de impact van de uitgezette campagne is;
- In toenemende mate worden alle online geschreven berichten geanalyseerd om inzicht te verkrijgen in de impact van de communicatie uitingen.
- Op basis van sentiment worden concurrenten geanalyseerd. Zodra concurrenten met vrijwel hetzelfde product een veel positiever sentiment hebben moet er actie ondernomen worden om het eigen sentiment te verbeteren.
- Onder andere de peilingen van politieke partijen worden gemeten aan de hand van sentiment;
- Sentiment wordt ingezet om filmopbrengsten en boekenverkopen te voorspellen;
- Sentiment wordt gebruikt om de stemming op de beurs te bepalen. Een positieve stemming kan zorgen voor een stijging van de markt en vice versa.

[5] Marketeer is een persoonsbenaming voor iemand die in de marketing werkt.

3.3 Methoden voor het bepalen van sentiment

Er zijn meerdere methoden beschikbaar voor het bepalen van sentiment over geschreven tekst. Er zijn standalone- en web-based programma's die het bepalen van sentiment mogelijk maken. Tevens zijn er grote verschillen in de mogelijkheden van de beschikbare programma's. De eenvoudige programma's zijn alleen in staat om sentiment te bepalen van de gehele tekst. Uitgebreide programma's identificeren alle onderwerpen en acties en bepalen afzonderlijk van elk geïdentificeerd onderwerp het sentiment.

Voor dit onderzoek is er gekozen om drie verschillende mogelijkheden voor het bepalen van sentiment te onderzoeken en te vergelijken. Elk type programma heeft zijn eigen benadering, werkwijze en mogelijkheden.

1. Het eerste type programma moet standalone werken en een tekst alleen positief of negatief classificeren;
2. Het tweede type programma moet standalone werken. Naast het bepalen van sentiment moet dit type programma ook in staat zijn om de onderwerpen en acties te identificeren;
3. Het derde type programma moet web-based zijn. Naast het bepalen van sentiment moet dit type programma ook in staat zijn om de onderwerpen en acties te identificeren.

Voor elk van deze type programma's is gezocht naar de beste mogelijkheid die al veelvuldig ingezet wordt en die zich reeds heeft bewezen voor zijn mogelijkheden en nauwkeurigheid met betrekking tot het bepalen van sentiment. De onderstaande methoden zijn geselecteerd voor dit onderzoek:

1. Bayesian methode (2001);
2. OpinionFinder (2005);
3. Open Amplify (2011).

Veel spamfilters[6] maken gebruik van de **Bayesian methode** voor het identificeren van spam.[7] Als een bepaald stuk tekst veel voorkomt in spamberichten maar niet in legitieme e-mailberichten, dan is het redelijk om aan te nemen dat dit bericht waarschijnlijk spam is. Deze methode kan ook toegepast worden om een tekst positief of negatief te classificeren.

Met behulp van **OpinionFinder** heeft Johan Bollen het voor elkaar gekregen om sentiment te bepalen over Twitterberichten. Met deze gegevens kan Johan Bollen met 87.6% procent zekerheid voorspellen of de Dow Jones over vier dagen gaat stijgen of dalen.[8]

Open Amplify wordt door een aantal grote bedrijven ingezet om het sentiment van hun merknaam te monitoren. Een aantal van deze bedrijven zijn: Shell, Honda, Volvo, E-on, PlayStation3 en T-Mobile.[9]

[6] Een spamfilter zorgt ervoor dat ongewenste e-mail wordt herkend en verwijderd.

[7] Jonathan A. Zdziarski (2005), Bayesian Content Filtering And The Art Of Statistical Language Classification.

[8] Johan Bollen, Huina Mao en Xiao-Jun Zeng (2010). Twitter mood predicts the stock market.

[9] Open Amplify (2011), Customers. Verkregen op 22 juli 2011 via www.openamplify.com/customers.

3.3.1 Bayesian methode

De eerste methode die is onderzocht is de Bayesian methode. Deze methode is gebaseerd op het principe dat gebeurtenissen in de meeste gevallen met elkaar samenhangen en dat de kans dat iets in de toekomst zal gebeuren kan worden bepaald uit eerdere gebeurtenissen.

De Bayesian methode stelt ons in staat om aan de hand van twee tekstbestanden een tekst positief of negatief te laten classificeren. Deze twee bestanden bestaan uit meerdere teksten waarvan vooraf is bepaald of ze positief of negatief zijn.

3.3.1.1 De tekstbestanden

Door gebruik te maken van twee tekstbestanden die in 2005 zijn opgesteld door Bo Pang & Lillian Lee[10] kan de Bayesian methode getest worden. De tekstbestanden bestaan beiden uit 5331 tekstfragmenten (zinnen) van www.rottentomatoes.com[11]. Wanneer een tekstfragment op de website van Rotten Tomatoes als ‘fresh’ is bestempeld is het in het bestand geplaatst met positieve teksten (rt-polarity-pos.txt) en wanneer een tekstfragment als ‘rotten’ is bestempeld staat het in het bestand met de negatieve teksten (rt-polarity-neg.txt).

Omdat in dit document niet beide bestanden in zijn geheel opgenomen kunnen worden staan hieronder drie willekeurige positieve en drie negatieve fragmenten uit de bestanden rt-polarity-pos.txt en rt-polarity-neg.txt.

Voorbeelden van zinnen uit het bestand rt-polarity-pos.txt:

- An utterly compelling 'who wrote it' in which the reputation of the most famous author who ever lived comes into question;
- Illuminating if overly talky documentary;
- A masterpiece four years in the making.

Voorbeelden van zinnen uit het bestand rt-polarity-neg.txt:

- The code talkers deserved better than a hollow tribute;
- Skip the film and buy the philip glass soundtrack cd;
- Feels like a cold old man going through the motions.

[10] Bo Pang & Lillian Lee zijn o.a. auteur van het boek: Opinion Mining And Sentiment Analysis (2008)

[11] Op de website www.rottentomatoes.com zijn filmtrailers en filmreviews te vinden. Gebruikers kunnen op deze website in combinatie met hun eigen review een film als ‘fresh’ of als ‘rotten’ bestempelen, hiermee geven ze aan of een film goed of slecht is.

3.3.1.2 Analyse van de tekstbestanden

Zowel de positieve als de negatieve tekstfragmenten zijn niet 100% nauwkeurig. Dit komt omdat de tekstfragmenten door user-generated content als 'fresh' of 'rotten' wordt bestempeld en dit kan tegenstrijdig zijn met de tekst. Een grove schatting na handmatige controle is dat ongeveer 90% van de tekstfragmenten in het juiste bestand zitten.

3.3.1.3 Werking programma

De werking van het programma is redelijk eenvoudig. Het programma begint met het bepalen van de kans dat een tekst positief of negatief is op basis van het aantal zinnen in het positieve en het negatieve tekstbestand. De kans op positief is meestal 50%. Dit komt omdat er meestal evenveel positieve als negatieve zinnen staan in de bestanden rt-polarity-pos.txt en rt-polarity-neg.txt. Nadat de kansfactor is vastgesteld wordt de tekst waarvan bepaald moet worden of het positief of negatief is verwerkt. Dit gebeurt door de tekst te splitsen op woorden en vervolgens kijkt het script in welk bestand de meeste matches voorkomen op die woorden. Als er meer positieve dan negatieve matches zijn wordt de tekst positief geclassificeerd en vice versa.

Voorbeeld werking programma

Ter illustratie worden hieronder twee zinnen door de Bayesian methode verwerkt. Beide zinnen komen van de website www.rottentomatoes.com en één is als 'fresh' bestempeld en de andere zin is als 'rotten' bestempeld.

De zin die als 'fresh' is bestempeld: '*Spiritually moving, visually daring and emotionally stirring, Avatar is classic storytelling at its very best.*' wordt door de bayesian methode als positief geclassificeerd.

De zin die als 'rotten' is bestempeld: *'Everything about the story, the setting, the dialog, and the parts that aren't purely visual is awful.'* wordt door de Bayesian methode als negatief geclassificeerd.

Deze twee voorbeelden heeft de bayesian methode goed geïnterpreteerd. In de volgende paragraaf wordt de nauwkeurigheid van de Bayesian methode getest. In deze test komt naar voren hoe vaak de Bayesian methode het goed heeft.

3.3.1.4 Nauwkeurigheidstest

Voor deze nauwkeurigheidstest bestaan beide tekstbestanden uit 5.331 tekstfragmenten. Deze tekstbestanden zijn voor deze test afgekapt tot 5.000 tekstfragmenten. De overige (2*331 stuks) tekstfragmenten zijn getoetst op nauwkeurigheid door ze opnieuw door de Bayesian methode te laten classificeren. Uit deze test kwam de volgende nauwkeurigheid naar voren:

- Bij de positieve tekstfragmenten heeft de Bayesian methode een nauwkeurigheid van 76%;
- Bij de negatieve tekstfragmenten heeft de Bayesian methode een nauwkeurigheid van 82%.

3.3.1.5 Voordelen van de Bayesian methode

- De Bayesian methode kan self learning gemaakt worden door tekstfragmenten die door de methode als positief zijn bestempeld toe te voegen aan het positieve tekstbestand (rt-polarity-pos.txt) en tekstfragmenten die negatief worden bestempeld toe te voegen aan het negatieve tekstbestand (rt-polarity-net.txt). Hierdoor kan er een steeds nauwkeurigere sentiment analyse worden gemaakt. Probleem is dat dit alleen werkt als vooraf met zekerheid is vastgesteld dat de bestanden rt-polarity-pos.txt en rt-polarity-neg.txt goed zijn opgebouwd. Als dit niet het geval is zou het self learning mechanisme averechts werken en zou het script steeds slechter worden;
- Een groot voordeel van de Bayesian methode is dat het script redelijk eenvoudig is waardoor er een duidelijk inzicht van de werking is;
- De Bayesian methode heeft zich al bewezen, deze techniek wordt veel gebruikt voor het identificeren van spam. Als een bepaald stuk tekst veel voorkomt in spam maar niet in legitieme e-mailberichten, dan is het redelijk om aan te nemen dat dit bericht waarschijnlijk spam is.

3.3.1.6 Nadelen van de Bayesian methode

- De Bayesian methode heeft het gemiddeld 21% van de keren niet goed. Hierbij wordt geen rekening gehouden met neutrale tekst;
- In dit geval werkt de Bayesian methode alleen bij movie reviews. Dit komt omdat dit het onderwerp is van de opgebouwde tekstbestanden. Voor ieder onderwerp moeten er nieuwe tekstbestanden worden aangemaakt, het is erg lastig om dit nauwkeurig te doen. Als een zin voor een movie review positief is wil dit niet zeggen dat dit ook voor elk ander onderwerp zo is;
- De Bayesian methode is niet in staat om neutrale tekst te identificeren;
- De sentiment analyse is afhankelijk van de taal waarin de positieve en negatieve bestanden zijn opgebouwd. Indien dit Engels is kan er alleen sentiment worden bepaald over engelse tekst;
- De Bayesian methode bepaald alleen of een tekst positief of negatief is. Hij identificeert niet de onderwerpen, werkwoorden etcetera;
- Het is erg lastig om de bestanden rt-polarity-pos.txt en rt-polarity-neg.txt nauwkeurig op te bouwen.

3.3.2 OpinionFinder

De tweede methode voor het bepalen van sentiment dat voor dit onderzoek is onderzocht is OpinionFinder. OpinionFinder is een verzameling van tools dat in 2005 is samengesteld en ontwikkeld door de universiteit van Pittsburgh in samenwerking met de universiteit van Utah en Cornell.[12] OpinionFinder kan documenten automatisch verwerken en kan bepalen of een zin subjectief of objectief is. OpinionFinder kan de onderwerpen en werkwoorden bepalen en ook welke woorden van een document positief en welke negatief zijn.

OpinionFinder kan als één pakket worden gedownload, iedere tool moet echter afzonderlijk op Linux[13] worden geïnstalleerd. Enige ervaring met Linux is hiervoor vereist. Technische informatie over de installatie van OpinionFinder staat in bijlage 3.

OpinionFinder werkt als het ware als één grote pijplijn (zie de flowchart in afbeelding 2). Iedere tekst die door OpinionFinder wordt verwerkt doorloopt deze stappen. Het eerste gedeelte zorgt ervoor dat de tekst goed voorbereid wordt voor de verwerking. Het tweede gedeelte bepaalt de werkwoorden, de onderwerpen, de subjectiviteit of de objectiviteit en het positieve of negatieve sentiment. De tien stappen die OpinionFinder doorloopt zijn:

1. Preprocessing;
2. Sentence Splitting and POS Tagging;
3. Stemming;
4. Feature Finder;
5. Shallow Parsing;
6. SourceFinder;
7. Direct Subjective Expression and Speech Event Classifier;
8. Subjectivity Classifier;
9. Polarity Classifier;
10. SGML markup.

3.3.2.1 Input en output

De input van OpinionFinder zijn standaard tekstbestanden met in ieder bestand een eigen tekst. Er kunnen meerdere teksten per keer verwerkt worden. Een documentenlijst (files.doclist) houdt alle verwijzingen naar ieder tekstbestand bij. Vervolgens hoeft maar één keer het document ‘files.doclist’ aangeroepen te worden.

In bijlage 1.1 wordt een voorbeeld van de in- en output van OpinionFinder gegeven. Voor ieder origineel tekstbestand is de output van OpinionFinder een nieuw tekstbestand in SGML/XML[14].

[12] OpinionFinder (2005). Documentation for OpinionFinder 1.5 (opinionfinderv1.5.readme).

[13] Linux is een gratis besturingssysteem dat oorspronkelijk is ontwikkeld door Linus Torvalds met ondersteuning van ontwikkelaars van over de hele wereld. Linux wordt geprezen voor het gebruik als besturingssysteem voor servers.

[14] SGML/XML staat voor Standard Generalized Markup Language / eXtensible Markup Language en is een standaard voor het uitwisselen van data.

3.3.2.2 Werkwijze OpinionFinder

In deze paragraaf worden de tien stappen van OpinionFinder beschreven. Alle fasen die OpinionFinder doorloopt en de tools die worden gebruikt komen aan bod. In bijlage 1.2 staan een aantal voorbeelden van de top 30 meest voorkomende onderwerpen, DSESE's [15] en voorbeelden van polariteit[16].

Stap 1: Voor het verwerken van een document maakt OpinionFinder als eerste gebruik van 'Preprocessing'. In deze fase worden eventueel voorkomende XML[17] en HTML[18] verwijderd van het originele tekstbestand. Uiteraard kan er geen sentiment worden bepaald over XML en HTML. Door dit te verwijderen is het originele tekstbestand klaar voor de verdere verwerking en kan OpinionFinder niet meer struikelen over deze codes.

Stap 2: Vervolgens maakt OpinionFinder gebruik van 'OpenNLP 1.3.0'. Deze tool zorgt ervoor dat zinnen worden gescheiden en vervolgens wordt d.m.v. 'part-of-speech-tagging' de lexicale categorie toegekend aan woorden in de zin. 'Fietsen' kan bijvoorbeeld een zelfstandig naamwoord zijn, maar ook een werkwoord. Deze woordsoorten worden met 'part-of-speech-tagging' geïdentificeerd.

OpenNLP heeft nog meerdere mogelijkheden in huis maar daar maakt OpinionFinder geen gebruik van.

Stap 3: Het derde proces wat OpinionFinder doorloopt is 'Stemming'. Voor dit proces maakt OpinionFinder gebruik van de tool 'SCOL version 1k' (ontwikkeld in Perl[19]). Deze tool zorgt ervoor dat alle werkwoorden worden omgezet naar de stam van het desbetreffende werkwoord. Bijvoorbeeld: 'fietsen' wordt 'fiets', 'lopen' wordt 'loop'. De output van het proces stemming wordt gebruikt bij de stappen 7, 8 en 9.

Stap 4: Na het proces 'Stemming' te hebben afgerond komt OpinionFinder in het vierde proces: 'Feature Finder'. Tijdens dit proces gaat OpinionFinder op zoek naar aanwijzingen voor subjectieve zinnen en ook naar aanwijzingen voor positieve en negatieve uitdrukkingen in de tekst. Deze aanwijzingen worden gebruikt voor de fasen 'Direct Subjective Expression and Speech Event Classifier', 'Subjectivity Classifier' en 'Polarity Classifier'. 'Feature Finder' is ontwikkeld in de programmeertaal Perl.

Stap 5: De vijfde fase waar OpinionFinder in terecht komt is 'Shallow Parsing'. 'Shallow Parsing' zorgt ervoor dat de werkwoorden en zelfstandige naamwoorden in de tekst worden geïdentificeerd.

Dit gebeurt door middel van de tool: SUNDANCE (Sentence UNDerstanding ANd Concept Extraction), ontwikkeld in Python door de NLP laboratorium op de universiteit van Utah.

'Shallow Parsing' wordt in de wetenschap ook wel aangeduid als 'Chunking' of 'Light Parsing'.

---

[15] De afkorting DSESE staat voor 'Direct Subjective Expression and Speech Event'. Zie stap 7 voor meer informatie.

[16] Polariteit is de aanduiding voor de plus- en minpool, waarbij plus staat voor positief en min voor negatief. In dit rapport wordt met polariteit de mate van positiviteit / negativiteit aangeduid van een tekst.

[17] XML staat voor eXtensible Markup Language en is een standaard voor het uitwisselen van data tussen computers.

[18] HTML is een afkorting voor: 'Hyper Text Markup Language'. Het is een programmeertaal waarmee websites ontwikkeld kunnen worden.

[19] Perl is een programmeertaal en de afkorting staat voor 'Practical Extraction and Report Language'.

Stap 6: De onderwerpen van de tekst worden in stap 6 geïdentificeerd. Dit gebeurt met de tool 'SourceFinder'. 'SourceFinder' is in 2005 ontwikkeld door NLP groups op de universiteit Cornell en de universiteit van Utah in de programmeertaal Python. De ontwikkelaars van 'SourceFinder' claimen dat hun tool een nauwkeurigheid heeft van 81,3%[20].

In de uiteindelijke output van OpinionFinder kunnen de onderwerpen herkent worden aan de SGML tags: <MPQASRC></MPQASRC>. Bijvoorbeeld: <MPQASRC>My mom</MPQASRC>.

Stap 7: Na 'SourceFinder' komt OpinionFinder bij 'Direct Subjective Expression and Speech Event (DSESE) Classifier'. De DSESE Classifier is ontwikkeld door Eric Breck in de programmeertaal Python.

Zoals de naam al aangeeft identificeert de DSESE Classifier subjectieve uitdrukkingen zoals: 'fears' en 'is happy' en gesproken & schriftelijke gebeurtenissen zoals: 'said' en 'according to'.

De DSESE's worden in de volksmond ook wel acties genoemd. Ter verduidelijking wordt hieronder een voorbeeld gegeven van drie verwerkte zinnen. De geïdentificeerde acties in deze zinnen zijn 'said', 'hate', 'thought' en 'hoped'.

- Jill **said**, "I **hate** Bill.";
- John **thought** he won the race;
- Mary **hoped** her presentation would go well.

In de uiteindelijke output van OpinionFinder kunnen de acties herkent worden aan de SGML tags: <MPQASD></MPQASD>. Bijvoorbeeld: <MPQASD>love</MPQASD>.

De DSESE Classifier is door de ontwikkelaars getest op 98 documenten waaruit een nauwkeurigheid kwam van 82%[21].

Stap 8: Na de 'DSESE Classifier' komt OpinionFinder in de 'Subjectivity Classifier'. Tijdens deze fase worden door middel van twee tools de zinnen die met behulp van 'OpenNLP 1.3.0' zijn geïdentificeerd subjectief of objectief geclassificeerd.

Dit zijn de twee tools waar de 'Subjectivity Classifier' van gebruik maakt:

- Riloff and Wiebe, 2004;
- Wiebe and Riloff, 2005.

[20] OpinionFinder (2005). SourceFinder.readme, dit bestand wordt meegeleverd met OpinionFinder.

[21] OpinionFinder (2005). Speech_DirSubj.readme, dit bestand wordt meegeleverd met OpinionFinder.

De eerste tool (Riloff and Wiebe, 2004) classificeert volgens de ontwikkelaars iedere zin objectief of subjectief met een gemiddelde nauwkeurigheid van 74%[22]. Dit hebben de ontwikkelaars getest door 9732 zinnen handmatig te bepalen of ze subjectief of objectief zijn en vervolgens gekeken hoe vaak hun tool dit goed heeft.

De tweede tool classificeert een zin alleen objectief of subjectief als hij dit zeker denkt te weten. Als deze tool het niet zeker weet classificeert hij de zin als 'unknown'. Wanneer deze tool een zin subjectief classificeert heeft hij dit 91,7% van de keren goed. Wanneer deze tool een zin objectief classificeert heeft hij een nauwkeurigheid van 83%[23].

Er kan dus gezegd worden dat wanneer tool 2 een zin objectief of subjectief classificeert en tool 1 zegt hetzelfde dat het dan zo goed als zeker is dat de 'Subjectivity Classifier' het goed heeft.

In de uiteindelijke output van OpinionFinder is dit de SGML output wat aangeeft of een zin subjectief of objectief is: <MPQASENT autoclass1=tool1 autoclass2=tool2 diff=0.0></MPQASENT>.

De 'diff' waarde geeft aan met welke zekerheid de 'Subjectivity Classifier' het goed heeft. Hoe hoger de 'diff' waarde hoe nauwkeuriger de uitkomst is.

Stap 9: Nadat OpinionFinder heeft bepaald welke zinnen van de ingevoerde tekst objectief of subjectief zijn komt OpinionFinder bij de 'Polarity Classifier' (ontwikkeld in de programmeertaal Python).

De 'Polarity Classifier' classificeert alle woorden in de tekst met hun contextuele polariteit. Eenvoudig gezegd houdt dit in dat de 'Polarity Classifier' aangeeft welke woorden in de tekst positief en welke woorden negatief zijn. Bijvoorbeeld: 'Happy' wordt positief geclassificeerd en 'Hate' wordt negatief geclassificeerd.

Het is echter niet zo dat een bepaald woord altijd positief of altijd negatief geclassificeerd wordt. Afhankelijk van de context van de tekst kan een woord de ene keer positief worden geclassificeerd en de andere keer negatief. Bijvoorbeeld het woord 'Good' is in de test uit bijlage 1.2 (output voorbeelden) 226.364 keer positief geclassificeerd maar ook 28.511 keer negatief.

De 'Polarity Classifier' werkt door middel van twee verschillende classificatiemethodes. De eerste methode herkent de sentimentele uitdrukkingen. De tweede methode bepaalt vervolgens of deze uitdrukkingen positief of negatief zijn.

De ontwikkelaars van deze tool geven aan dat de 'Polarity Classifier' een nauwkeurigheid heeft van 73,9%[24].

[22] OpinionFinder (2005). Subjectivity.readme, dit bestand wordt meegeleverd met OpinionFinder.

[23] OpinionFinder (2005). Subjectivity.readme, dit bestand wordt meegeleverd met OpinionFinder.

[24] OpinionFinder (2005). Polarity.readme, dit bestand wordt meegeleverd met OpinionFinder.

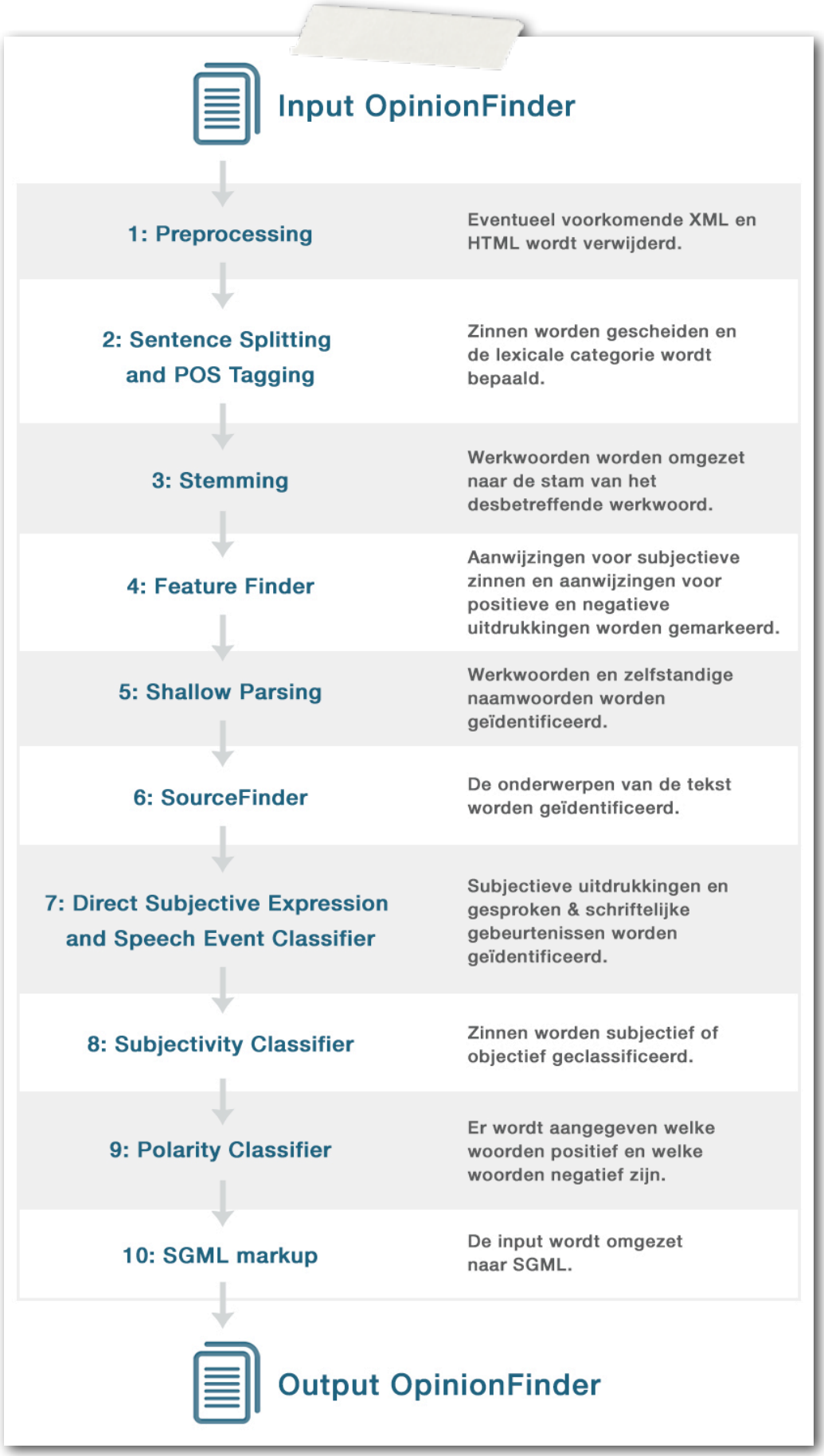
Stap 10: De laatste stap wat OpinionFinder doorloopt is 'SGML markup'. In deze stap wordt de input van OpinionFinder omgezet naar SGML (Standard Generalized Markup Language) waarin alle zinnen, onderwerpen, DSESE's en polariteiten van de vorige stappen zijn aangegeven. SGML is een platform onafhankelijke ISO-standaard voor de syntaxis voor markup talen. In bijlage 1.1 staat een voorbeeld van de output van OpinionFinder.

3.3.2.3 Verwerkingstijd

De verwerkingstijd van OpinionFinder is afhankelijk van de configuratie en de capaciteit van de server. OpinionFinder vereist pure rekenkracht. Met de gebruikte configuratie: een Penium(R) Dual-Core CPU @ 2,5 GHz en 3 Gb werkgeheugen verwerk OpinionFinder ongeveer 10.000 Twitterberichten in 45 minuten.

3.3.2.4 Flowchart

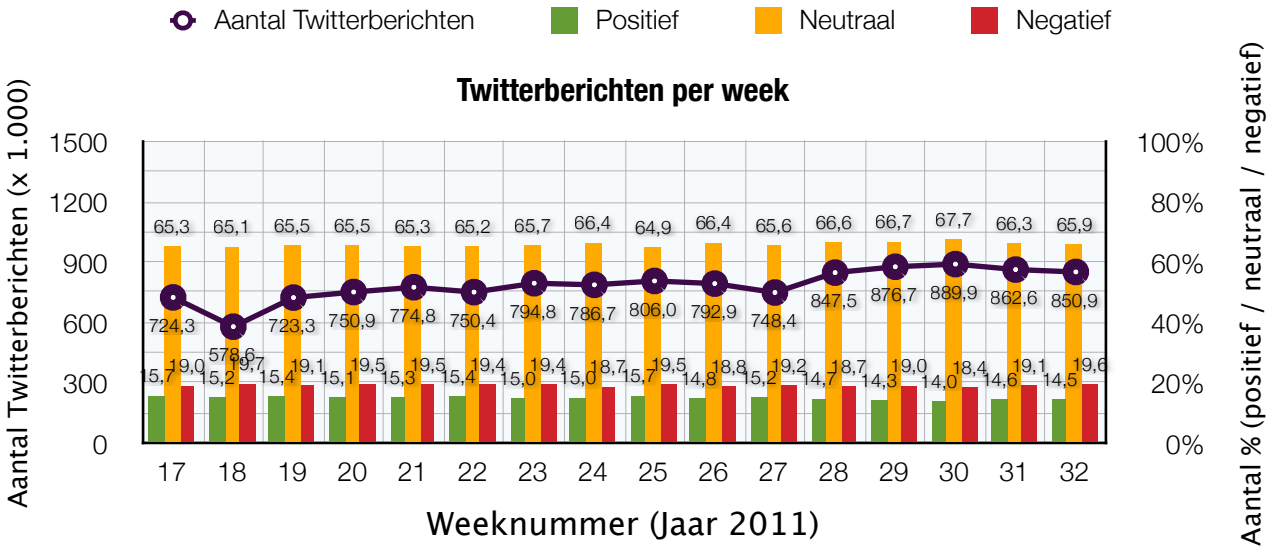
Zoals eerder vermeld werkt OpinionFinder als het ware als een grote pijplijn met allemaal processen die één voor één worden doorlopen. In afbeelding 2 wordt de flowchart van OpinionFinder weergegeven met alle processen die door OpinionFinder worden doorlopen plus een korte omschrijving van die processen.



Afbeelding 2: Flowchart OpinionFinder

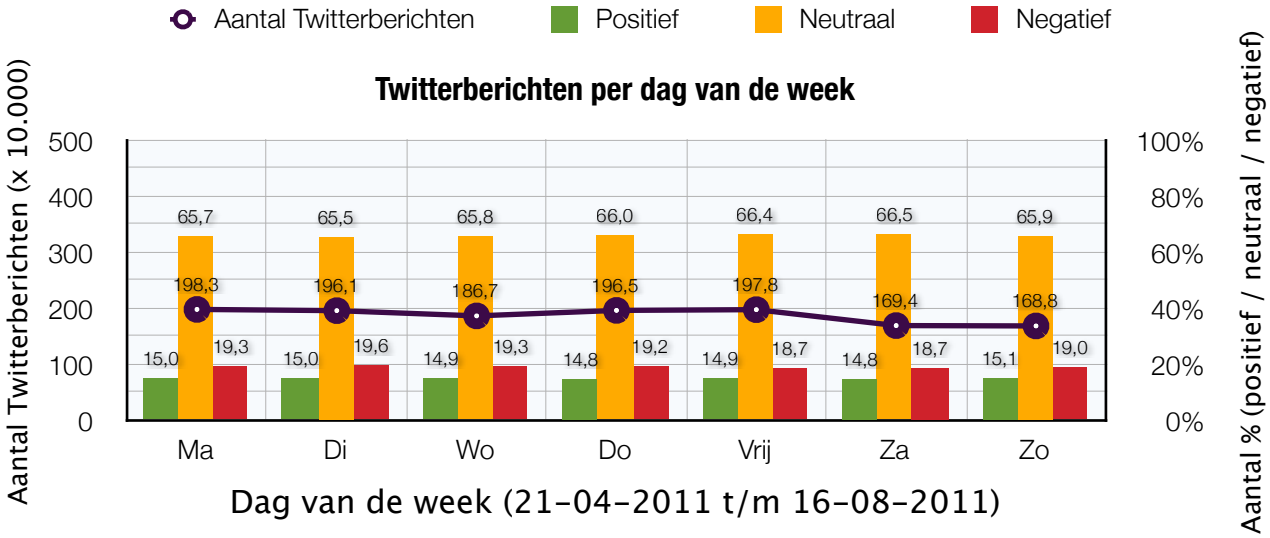
3.3.2.5 Bepalen van sentiment aan de hand van de output van OpinionFinder

Er zijn meerdere testen uitgevoerd voor het bepalen van sentiment met de output van OpinionFinder. Deze testen zijn uitgevoerd op de 10.850.000 willekeurig Twitterberichten die reeds zijn verwerkt door OpinionFinder. De uitkomst van deze testen worden in de vier onderstaande afbeeldingen weergegeven en geven allemaal een verandering van sentiment weer over een bepaalde periode.



Afbeelding 3: wekelijkse verandering van sentiment

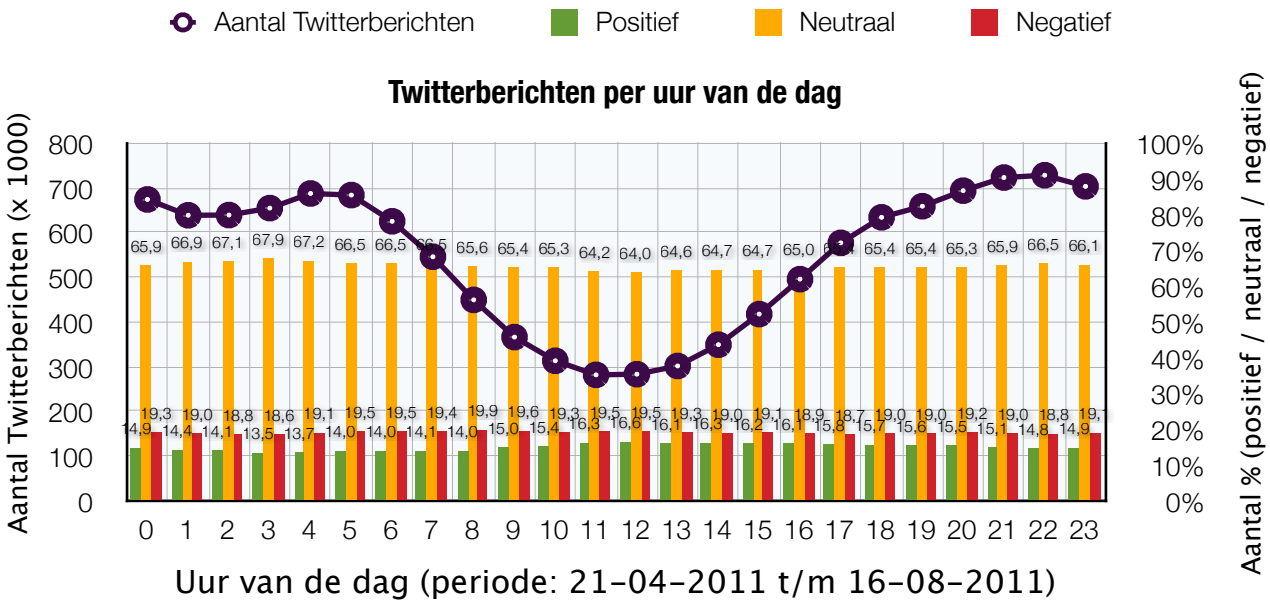
De verandering van sentiment over de willekeurig verwerkte Twitterberichten per week is nagenoeg stabiel. Dit komt waarschijnlijk doordat het volume en de diversiteit van deze Twitterberichten erg groot is.



Afbeelding 4: verandering van sentiment per dag van de week

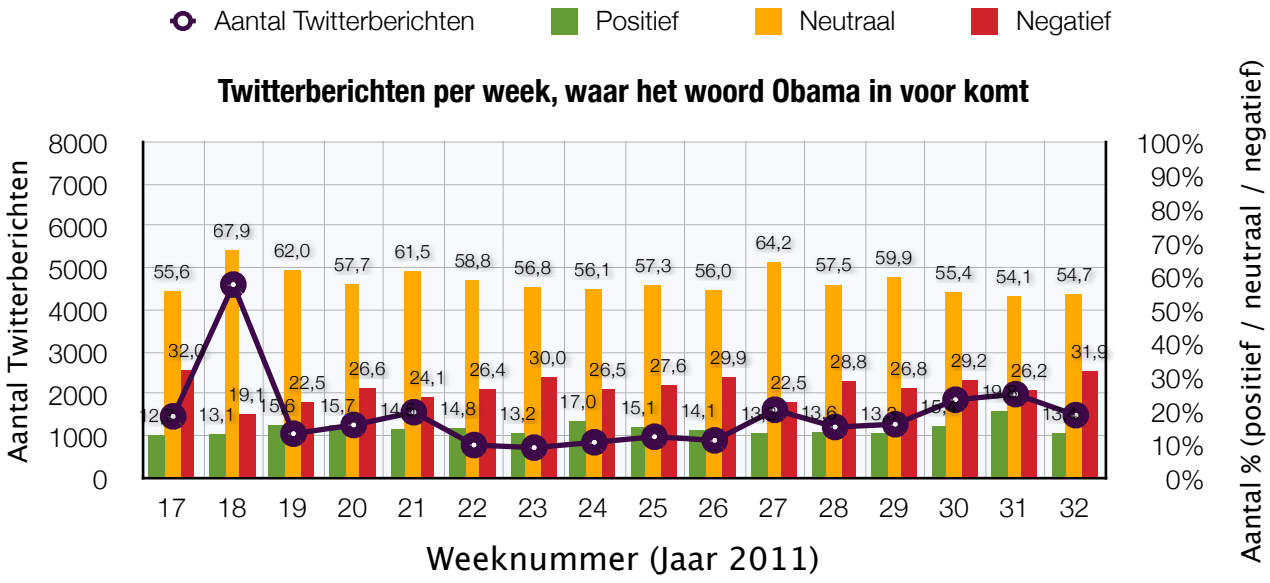
Zodra het sentiment van de willekeurig verwerkte Twitterberichten per dag van de week wordt weergegeven is te zien dat het sentiment wederom erg stabiel is. Op vrijdag en zaterdag is het sentiment iets minder negatief. Dit komt hoogstwaarschijnlijk door het weekend. Over het algemeen zijn mensen in het weekend positiever[25].

[25] George Lowery (2011). Tweets study: People across the globe report similar, ever-changing moods. Verkregen op 30 september 2011, via Cornell University website: http://www.news.cornell.edu/stories/Sept11/TwitterMoods.html



Afbeelding 5: verandering van sentiment per uur van de dag

In afbeelding 5 worden de resultaten weergegeven van de derde test. In deze test is er gekeken naar op welk moment van de dag het meest positief of negatieve sentiment is. Zoals in de afbeelding wordt weergegeven is er vrijwel geen verschil in sentiment in de loop van de dag. Wel is te zien dat er een grote schommeling is van het aantal Twitterberichten dat op het moment van de dag wordt verstuurd. Zodra het in Nederland tussen de 7.00 en 17.00 uur is wordt er veel minder in de engelse taal gecommuniceerd via Twitter. Dit is te verklaren door het tijdsverschil. In Amerika is het dan nacht.



Afbeelding 6: de wekelijkse verandering van sentiment van Obama

In de afbeelding 6 wordt de verandering van sentiment per week weergegeven zodra er wordt gezocht op een bepaald onderwerp. In dit voorbeeld is het onderwerp ‘Obama’. In week 18 is er veel meer over ‘Obama’ geschreven, tevens is in deze week het sentiment beduidend minder negatief dan in de rest van de periode. Hoogstwaarschijnlijk is dit te verklaren door het feit dat Osama Bin Laden in die week is omgebracht door de Amerikanen.

In de vier bovenstaande afbeeldingen is de verandering van sentiment over een bepaalde periode erg stabiel tenzij er wordt gezocht op een onderwerp. Deze stabiliteit is te verklaren door het volume en de diversiteit van de 10.850.000 willekeurig verwerkte Twitterberichten.

3.3.2.6 Voordelen van OpinionFinder

- OpinionFinder heeft een hoge verwerkingssnelheid, per dag kan OpinionFinder ongeveer 320.000 Twitterberichten verwerken, dit zijn ongeveer 4 Twitterberichten per seconde;
- Na de installatie van alle tools van OpinionFinder werkt het erg stabiel. OpinionFinder is na het verwerken van 10.850.000 Twitterberichten geen één keer vastgelopen;
- Omdat OpinionFinder erg snel teksten kan verwerken is het uitermate geschikt voor de verwerking van Twitterberichten. Per dag kunnen er meer dan 100.000 Twitterberichten binnen komen. Deze berichten moeten dezelfde dag allemaal nog verwerkt kunnen worden. Als dit niet zou gebeuren zou de tool naar mate de tijd verstrekt alleen maar verder achter komen te lopen.

3.3.2.7 Nadelen van OpinionFinder

- OpinionFinder stelt niet in staat om eenvoudig teksten automatisch te verwerken. Er moet eerst zelf een programma worden ontwikkeld die ervoor zorgt dat de input voor OpinionFinder automatisch goed wordt neergezet, dat OpinionFinder vervolgens automatisch wordt aangeroepen en dat de output vervolgens op een juiste manier wordt uitgelezen en wordt opgeslagen in een centrale database;
- OpinionFinder kan alleen documenten verwerken die in de Engelse taal zijn geschreven. Voor dit onderzoek is niet getest wat de resultaten van OpinionFinder zijn als bijvoorbeeld Nederlandse teksten eerst zouden worden vertaald met bijvoorbeeld Google Translate alvorens de teksten zouden worden verwerkt met OpinionFinder. Er is hier bewust voor gekozen om de teksten niet te vertalen omdat Google heeft aangekondigd dat vanaf december de API van Google Translate niet meer gratis beschikbaar is[26];
- OpinionFinder 1.5 dat voor dit onderzoek is gebruikt bestaat sinds het jaar 2005 en wordt niet meer verder doorontwikkeld, de tool is inmiddels behoorlijk verouderd;
- Afhankelijk van de juiste fase waar OpinionFinder zich in bevindt is de nauwkeurigheid verschillend maar nooit 100% perfect. Volgens de ontwikkelaars ligt de nauwkeurigheid meestal rond de 80%;
- OpinionFinder is volgens de ontwikkelaars voor onderzoek gratis te gebruiken, voor commerciële doeleinden niet;
- De installatie van OpinionFinder is erg lastig, kennis van Linux, Python en PERL is vereist. Naast de bijgeleverde readme's is er weinig documentatie over OpinionFinder te vinden.

[26] Pricing and Terms of Service (z.d.). Verkregen op 5 augustus 2011, via http://code.google.com/apis/language/translate/v2/pricing.html

3.3.3 Open Amplify

De derde en tevens laatste methode voor het bepalen van sentiment die is onderzocht is Open Amplify. Deze software voor het analyseren van tekst identificeert net als andere soortgelijke programma's de woordsoorten (werkwoord, lidwoord, bijvoeglijk naamwoord, zelfstandig naamwoord etc.). Open Amplify heeft echter veel meer mogelijkheden dan alleen het identificeren van deze woordsoorten. Open Amplify heeft een heel geavanceerde technologie in huis voor het classificeren van indicatoren van emotionele betekenis.

Open Amplify is een webservice die eenvoudig te benaderen is via het internet. Zodra de API van Open Amplify wordt aangeroepen wordt de ingevoerde tekst volledig geanalyseerd. Hoe deze webservice exact de ingevoerde teksten analyseert is onbekend. Open Amplify is als het ware een 'black box'.

3.3.3.1 Hoe kan er verbinding gemaakt worden met de API van Open Amplify?

Alvorens er verbinding gemaakt kan worden met de Open Amplify API dient de gebruiker eerst de beschikking te hebben tot een API-key van Open Amplify. Deze API-key wordt na registratie op de website van Open Amplify via de e-mail toegezonden. Met de gratis registratie kunnen er 1000 teksten per dag worden verwerkt.

Zodra er een API-key beschikbaar is kan er eenvoudig verbinding worden gemaakt met de Open Amplify API d.m.v een POST of GET commando. Het maakt niet uit vanuit welke programmeertaal dit commando wordt verstuurd. Hierdoor kan de gebruiker van de API zelf de programmeertaal kiezen die voor hem het beste werkt. Nadat Open Amplify de tekst heeft geanalyseerd stuurt de API de taalkundige gegevens van de tekst in een van te voren aangegeven formaat terug naar de gebruiker. De belangrijkste formaten voor de Output die Open Amplify ondersteunt zijn: XML, JSON, CSV en RDF.

3.3.3.2 De mogelijkheden van Open Amplify

De belangrijkste mogelijkheden van Open Amplify worden in deze paragraaf beschreven. Het gaat hierbij voornamelijk om wat voor output de Open Amplify API genereert, hiervan zullen in deze paragraaf diverse voorbeelden worden weergegeven. De vijf belangrijkste onderdelen die zullen worden beschreven zijn:

1. De Topics Analysis;
2. De Actions Analysis;
3. De Styles Analysis;
4. De Demographics Analysis;
5. De Topic Intentions Analysis.

Tijdens het aanroepen van de Open Amplify API kan er aangegeven worden van welke mogelijkheden gebruik gemaakt moet worden. Zo kan a.d.h.v. de wensen van de gebruiker de optimale snelheid behaald worden met het analyseren van de teksten.

Om deze vijf mogelijkheden verder uit te leggen is er één willekeurige tekst verwerkt zodat bij iedere mogelijkheid een voorbeeld gegeven kan worden van de output. Zie bijlage 2.1 voor de gebruikte voorbeeld tekst.

3.3.3.2.1 De Topics Analysis

Kort samengevat identificeert de 'Topics Analysis' alle onderwerpen van de ingevoerde tekst en welke onderwerpen met elkaar in verband staan. Daarnaast geeft de 'Topics Analysis' onder andere ook aan wat de domeinen en locaties zijn van de tekst. Tot slot identificeert de 'Topics Analysis' ook de eigennamen. Deze vier mogelijkheden van de 'Topics Analysis' worden hieronder verder uitgewerkt.

1: De onderwerpen die Open Amplify identificeert worden gesorteerd op basis van hoe vaak het onderwerp in de tekst voor komt. Open Amplify geeft hiervoor een waarde mee, hoe hoger deze waarde is, hoe vaker het woord in de tekst voor komt.

De ontwikkelaars van Open Amplify zeggen dat Open Amplify echt de tekst begrijpt. Hierdoor kunnen ze naast het identificeren van de onderwerpen, locaties en domeinen ook aangeven welke onderwerpen met elkaar samenhangen en tevens ook per onderwerp een mate van polariteit (positiviteit / negativiteit) geven.

Een belangrijke functie die Open Amplify in huis heeft voor het analyseren van onderwerpen is co-referentie. Het komt er op neer dat Open Amplify meerdere onderwerpen aan elkaar weet te koppelen zodra ze naar elkaar verwijzen. Open Amplify weet:

- dat zodra er wordt geschreven over bijvoorbeeld 'Jack' en 'Jill', en vervolgens over 'He' en 'She'. Dat 'He' een verwijzing is naar 'Jack' en dat 'She' een verwijzing is naar 'Jill'. 'He' & 'Jack' en 'She' & 'Jill' zullen worden gezien als hetzelfde onderwerp;
- dat zodra er wordt geschreven over bijvoorbeeld 'President Obama', 'Obama' of 'President' dat het allemaal over hetzelfde onderwerp gaat;
- dat zodra er wordt geschreven over bijvoorbeeld 'Grooveproof Rappers Corporation' en vervolgens over de afkorting 'GRC' dat de schrijver van de tekst dan hetzelfde onderwerp bedoelt.

Deze co-referentie is erg belangrijk om de tekst echt te kunnen begrijpen. Zie het volgende voorbeeld: "Ik ben vandaag bij Hendrik op visite geweest. Wat is het toch een akelige vent!". Zonder co-referentie zou Open Amplify nooit kunnen weten dat het onderwerp 'Hendrik' negatief is.

Voor ieder onderwerp dat Open Amplify heeft geïdentificeerd komt Open Amplify met een waarde van polariteit. Deze waarde loopt van -1.00 tot 1.00 waarbij -1.00 staat voor negatief, 1.00 staat voor positief en 0.00 voor neutraal. Deze waarde van polariteit is nodig voor het bepalen van sentiment. Hoe open Amplify de waarde van polariteit bepaalt is niet bekend. Zoals eerder staat vermeld is Open Amplify als het ware een black box.

Zoals in bijlage 2.2 is te zien geeft Open Amplify naast de polariteit van ieder onderwerp ook beschrijvende informatie over het onderwerp (NER). Daarnaast geeft Open Amplify ook aan of het onderwerp begeleiding vraagt of begeleiding aanbiedt. Gevraagde begeleiding houdt in dat het nut of het doel van het onderwerp niet helemaal duidelijk is. Bijvoorbeeld in de zin 'Door wie is het boek geschreven?' is er gevraagde begeleiding over het onderwerp 'boek'. Er wordt extra informatie gevraagd over het onderwerp. Aangeboden begeleiding houdt in dat er extra informatie wordt gegeven over het onderwerp. Bijvoorbeeld in de zin 'Het boek is geschreven door Piet Hein' is er aangeboden begeleiding. Er wordt extra informatie gegeven over het onderwerp 'boek'. Open Amplify maakt onderscheid in de volgende mate van begeleiding: geen, een beetje of veel begeleiding.

Voor het bepalen van sentiment kan het erg handig zijn om te weten wat het onderwerp is van de tekst. Door de manier waarop Open Amplify de onderwerpen classificeert is het erg eenvoudig om het sentiment over een bepaald onderwerp te bepalen.

2. Een eigennaam is een bijzonder zelfstandig naamwoord dat wordt gebruikt om een bepaald persoon of zaak mee aan te duiden. Open Amplify identificeert alle eigennamen van de tekst. Voor iedere eigennaam worden er dezelfde gegevens als bij een onderwerp teruggegeven. Voor een voorbeeld van de output (structuur) van de eigennamen zie bijlage 2.2.

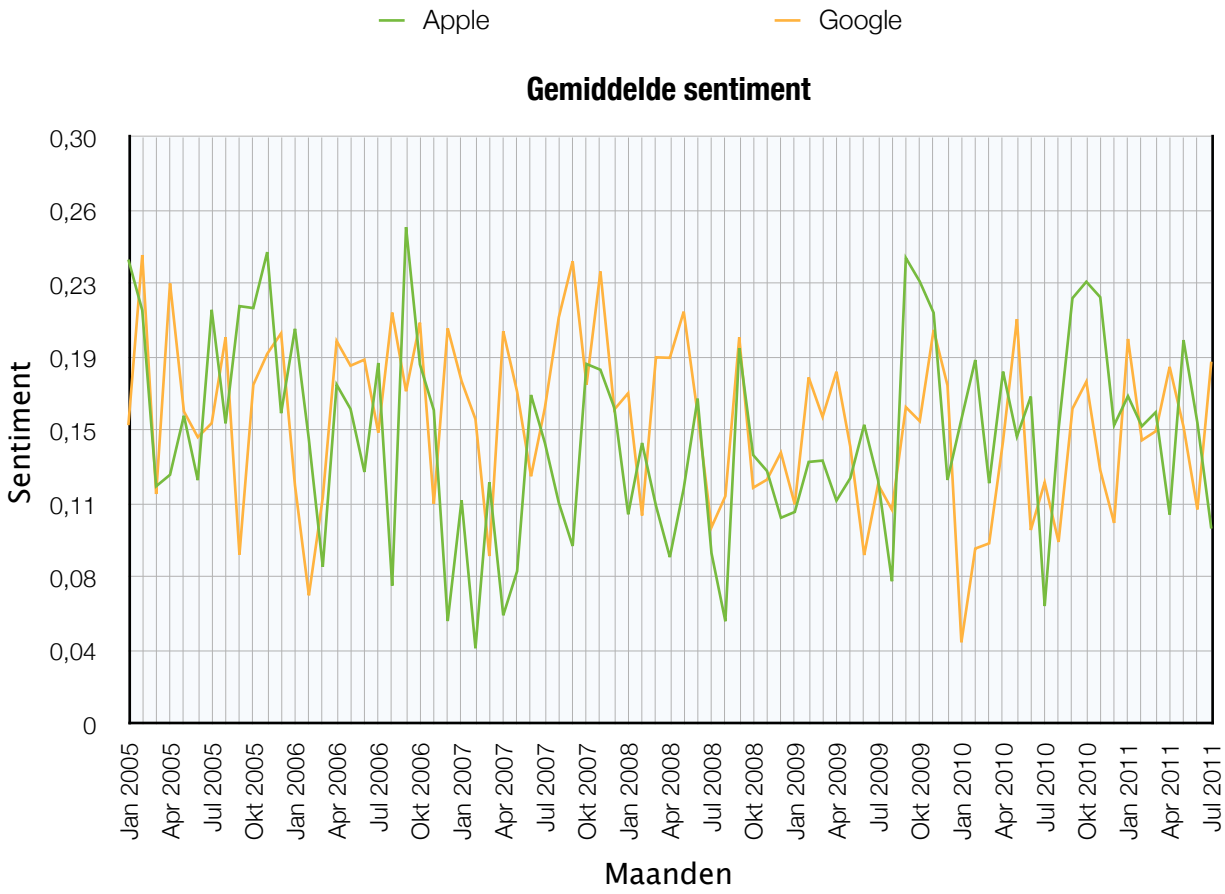
3: Het domein is het algemene onderwerp van de tekst. Om het domein te kunnen classificeren hoeft dit niet persé genoemd te worden in de tekst. Zodra Open Amplify bijvoorbeeld doorheeft dat de belangrijkste onderwerpen van de tekst 'katten' & 'honden' zijn dan classificeert Open Amplify het domein 'dieren' zonder dat dit woord direct in de tekst vermeld wordt. Open Amplify classificeert ook onderliggende domeinen. Deze onderliggende domeinen, ook wel subdomeinen genoemd gaan tot drie niveau's diep. In bijlage 2.3 wordt een voorbeeld van deze geïdentificeerde domeinen gegeven.

4: De locaties die Open Amplify classificeert komen rechtstreeks uit de lijst met onderwerpen. Echter voor het gebruiksgemak heeft Open Amplify ervoor gekozen om locaties niet tussen de rest van de onderwerpen weer te geven maar apart. Open Amplify is in staat om per tekst verscheidene locaties te identificeren. Net als bij de onderwerpen en domeinen wordt er ook voor iedere locatie een waarde meegegeven. Hoe hoger deze waarde, hoe zekerder Open Amplify van het resultaat is.

Een **duidelijk overzicht** van alle mogelijkheden van de 'Topics Analysis' wordt hieronder in hiërarchische weergave weergegeven.

- Domeinen;
  - Sub domeinen;
    - Sub domeinen;
- Onderwerpen;
  - Polariteit (waarde van positiviteit / negativiteit);
  - Begeleiding;
    - Aangeboden;
    - Gevraagd;
  - Beschrijvende informatie van het onderwerp (bijvoorbeeld: persoon, organisatie, land etc.);
- Eigennamen;
  - Polariteit (waarde van positiviteit / negativiteit);
  - Begeleiding;
    - Aangeboden;
    - Gevraagd;
  - Beschrijvende informatie van het eigennaam (bijvoorbeeld: persoon, organisatie, land etc.);
- Locaties.

Sentiment bepalen over de onderwerpen vindt plaats door de gemiddelde polariteit van een bepaalde periode te meten. In onderstaande grafiek wordt het verschil in sentiment weergegeven van de onderwerpen Apple en Google. Voor deze grafiek zijn er ongeveer 150.000 nieuwsartikelen verwerkt over deze onderwerpen. Sentiment met een gemiddelde waarde die hoger is dan 0 is positief, en lager dan 0 is negatief.



Afbeelding 7: gemiddelde sentiment van Apple en Google

Een opvallend detail is dat het gemiddelde sentiment per maand van beide onderwerpen in de periode januari 2005 tot en met juli 2011 nooit negatief is.

3.3.3.2.2 De Actions Analysis

De 'Actions Analysis' geeft alle acties terug die in de tekst voorkomen. De acties zijn niet direct de werkwoorden, een actie kan ook combinatie zijn van het werkwoord met het object waar het werkwoord naar verwijst. Bijvoorbeeld: 'Gooi de bal'. Dit is vaak veel nuttiger dan alleen een lijst met werkwoorden. Van iedere actie wordt tevens altijd het werkwoord als infinitief weergegeven, 'Said' wordt dus 'Say'.

Bij alle geïdentificeerde acties wordt extra informatie gegeven. Denk hierbij bijvoorbeeld aan een tijdsaanduiding, een mate van waarschijnlijkheid dat de actie wordt uitgevoerd en een actietype. Ook wordt er bij iedere actie net als bij de onderwerpen en eigennamen de mate van begeleiding aangegeven.

In bijlage 2.2 staat een voorbeeld van de output (structuur) van de 'Actions Analysis'. Om het overzichtelijk te houden wordt alleen de top vijf van acties weergegeven.

1: De decisiveness geeft aan wat de waarschijnlijkheid is dat de actie is of wordt uitgevoerd. Taalkundige technieken maken het mogelijk om dit door de computer te laten vaststellen. Bij iedere actie geeft Open Amplify deze mate van waarschijnlijkheid. Open Amplify maakt onderscheid in de volgende mate van waarschijnlijkheid: niet beschikbaar, laag, gemiddeld-laag, gemiddeld, gemiddeld-hoog of hoog. Hoe hoger de waarschijnlijkheid, hoe waarschijnlijker het is dat de actie is of wordt uitgevoerd.

2: De action type is een indeling van de actie in een bruikbare categorie. Acties zijn gebaseerd op werkwoorden, hierdoor is de range van deze acties is erg breed. Omdat de range van de verschillende acties erg breed is kan het erg moeilijk zijn om geautomatiseerde beslissingen te maken a.d.h.v. deze acties. Open Amplify classificeert daarom de verschillende acties in actie typen. Bijvoorbeeld: ‘construct’, ‘develop’, en ‘mock up’ worden allemaal verbonden aan de actie type ‘build’.

Dit zijn de typen acties waar Open Amplify iedere actie aan koppelt: Advise, assess, build, buy, choose, collaborate, communicate, create, help, lead_control, learn, request, sell, socialize, travel, use, vote, watch_attend, other

3: De temporality geeft een tijdsaanduiding van de actie weer. Een actie kan beter worden begrepen zodra er een tijdsaanduiding bij die actie is. Open Amplify maakt voor iedere actie onderscheid in de volgende tijdsaanduidingen: niet beschikbaar, verleden, recent, heden of toekomst. Naast deze verschillende tijdsaanduidingen kan Open Amplify ook expliciete tijdsaanduidingen voor de acties identificeren zoals ‘vandaag’, ‘volgende week’, ‘5 januari’, ‘met kerst’ etc.

4: De requesting- en offering guidance geeft net als bij de onderwerpen en eigennamen aan wat de begeleiding is van de actie. Deze begeleiding over de actie kan gevraagd zijn of aangeboden worden. Een tekst waarin veel begeleidende acties staan kan bijvoorbeeld een instructie zijn over een bepaald onderwerp.

3.3.3.2.3 De Styles Analysis

De Styles Analysis geeft een aantal kenmerken over de gehele tekst. Een aantal van die kenmerken worden ook al gebruikt bij de Topics- en Actions Analysis. Deze kenmerken die al eerder zijn uitgelegd in de voorgaande hoofdstukken zijn: Polariteit, Offering Guidance, Requesting guidance, Decisiveness en een tijdsaanduiding. Bij de Styles Analysis geeft bijvoorbeeld de polariteit de mate van positiviteit / negativiteit van de gehele tekst i.p.v. over één specifiek onderwerp. Een aantal kenmerken die niet in de Topics- en de Actions analysis voorkomen maar wel in de Styles Analysis worden hieronder uitgewerkt.

1: De Flamboyance van de tekst geeft aan wat de schrijfstijl is van de tekst. Deze schrijfstijl kan erg uitgebreid of eenvoudig zijn. De Flamboyance geeft in een schaal van 1 t/m 4 de schrijfstijl van de tekst aan.

2: Slang geeft een percentage van de tekst aan dat is geschreven in slecht taalgebruik. Dit kenmerk kan nuttig zijn voor het identificeren van de auteur en doelgroep van de tekst.

3: Het Contrast geeft aan in welke mate de geanalyseerde tekst tegenstrijdig is. In één tekst kan de schrijver heel tegenstrijdige meningen hebben over hetzelfde onderwerp. Open Amplify maakt onderscheid in de volgende drie mate van tegenstrijdigheid: geen, een beetje of veel.

3.3.3.2.4 De Demographics Analysis

De Demographics Analysis geeft d.m.v. taalkundige technieken een schatting van de demografische signalen van de tekst. De signalen die de Demographics Analysis van iedere tekst geeft zijn 'Age', 'Gender', 'Education' en 'Language'.

1: Age geeft een schatting van leeftijd van de doelgroep voor de tekst. Open Amplify maakt hierbij onderscheid tussen 'Young', 'Adult' en 'Senior'.

2: Gender geeft de vermoedelijke sekse van de auteur / doelgroep van de tekst. Verschillende onderwerpen kunnen worden gekoppeld aan het interessegebied van een geslacht. Wanneer een tekst bijvoorbeeld over dure en luxe auto's gaat zou Open Amplify aangeven dat de tekst waarschijnlijk door een man, en voor mannen is geschreven. Uiteraard is dit niet 100% nauwkeurig.

3: Education geeft een vermoedelijke opleidingsniveau van de beoogde doelgroep van de tekst. Open Amplify maakt hiervoor onderscheid in de volgende waardes: undecided, pre-secondary, secondary, college of post graduate.

4: Bij **Language** wordt de taal aangegeven waarin de tekst is geschreven. Open Amplify kan de talen wel herkennen maar kan alleen overweg met teksten die in de Engelse taal zijn geschreven.

3.3.3.2.5 De Topic Intentions Analysis

De 'Topic Intentions Analysis' combineert alle signalen van de 'Topics Analysis' en de 'Actions Analysis'. Er zit vaak een direct verband tussen de onderwerpen en de acties en die worden met behulp van deze analyse duidelijk gemaakt. Open Amplify is al een erg krachtige methode voor het analyseren van teksten, door deze analyse is het nog veel krachtiger.

Iedere actie die is geïdentificeerd door Open Amplify wordt gekoppeld aan een onderwerp. Bij iedere actie wordt vervolgens de waarschijnlijkheid weergegeven van dat die actie wordt uitgevoerd. Tevens wordt bij iedere actie een tijdsaanduiding gegeven. Op deze manier kan bijvoorbeeld het sentiment van de iPhone worden vastgesteld met de waarschijnlijkheid dat dit toestel in de toekomst gekocht gaat worden.

3.3.3.3 Flowchart Open Amplify

Zoals eerder vermeld is Open Amplify als het ware een Black Box. Gebruikers weten wat de input en de output is van Open Amplify, maar hoe Open Amplify precies tot die output komt is voor de gebruikers onbekend. Uiteraard is het voor de ontwikkelaars van Open Amplify wel bekend wat er in de Black Box gebeurt.

3.3.3.4 Verwerkingstijd

Open Amplify is redelijk snel in het verwerken van artikelen en Twitterberichten maar niet zo snel als OpinionFinder. In één uur tijd is Open Amplify in staat om ongeveer vijfhonderd artikelen te verwerken of ongeveer duizend Twitterberichten.

3.3.3.5 Voordelen van de Open Amplify API

- Open Amplify heeft van de drie onderzochte methoden de meeste functionaliteiten;
- Open Amplify bepaald polariteit over onderwerpen en over de gehele tekst. Deze polariteit varieert van -1.00 tot 1.00. Hierdoor maakt Open Amplify niet alleen onderscheid in positief en negatief maar ook in welke mate het onderwerp of de tekst positief of negatief is. Het bepalen van sentiment is bij Open Amplify specifieker dan bij de andere geteste methoden;
- Open Amplify staat het toe dat de output voor alle commerciële doeleinden mag worden gebruikt. Dit geld ook voor de gratis variant van Open Amplify waarmee 1000 artikelen per dag verwerkt kunnen worden;
- Open Amplify blijft door ontwikkeld worden waardoor de functionaliteiten steeds uitgebreider worden en de resultaten nauwkeuriger. Zodra er een nieuwe versie van Open Amplify wordt gelanceerd kan er eenvoudig verbinding gemaakt worden met de nieuwere versie van Open Amplify. Momenteel heeft Open Amplify twee versies live.

3.3.3.6 Nadelen van de Open Amplify API

- Zodra een tekst groter is dan 5 kb kan Open Amplify de tekst niet in zijn geheel verwerken. De tekst moet dan opgesplitst worden in stukken van 5 kb;
- Tijdens de geteste periode is de server van Open Amplify regelmatig offline geweest. In sommige gevallen was dit zelfs langer dan één uur;
- De gebruikers van Open Amplify zijn volledig van hun afhankelijk. Als Open Amplify besluit om hun dienst niet meer beschikbaar te stellen kunnen huidige producten die zijn gebaseerd op hun software niet meer geleverd worden. Om dit te voorkomen zou het contractueel vastgelegd kunnen worden dat Open Amplify haar services zal blijven leveren;
- Open Amplify is een black box, hoe het systeem precies werkt zou niet bekent gemaakt worden.

3.3.4 Vergelijking van de drie methoden

In dit hoofdstuk worden de drie eerder uitgewerkte methoden voor het bepalen van sentiment met elkaar vergeleken. Het doel van deze vergelijking is vaststellen welke van deze tools het meest geschikt is voor het bepalen van sentiment. Er worden vier verschillende vergelijkingen gemaakt:

1. Een vergelijking van de **mogelijkheden** van de drie verschillende methoden;
2. Een vergelijking van geïdentificeerde **onderwerpen** en **acties**;
3. Een vergelijking van **sentiment** over een periode van tien dagen;
4. Een vergelijking van de **nauwkeurigheid** van de polariteit.

Voor de vergelijking van de onderwerpen, acties en het sentiment zijn er door elke onderzochte methode in totaal 100.000 Twitterberichten verwerkt die betrekking hebben op de onderwerpen ‘apple’ en ‘google’. Deze 100.000 berichten is een willekeurige set van 10.000 Twitterberichten per dag over een periode van 10 aaneensluitende dagen.

1: De mogelijkheden van de verschillende methoden worden hieronder met elkaar vergeleken:

	Methoden		
Identificeert	**Bayesian**	**OpinionFinder**	**Open Amplify**
1: Onderwerpen		✔	✔
Polariteit van de onderwerpen			✔
Begeleiding van de onderwerpen			✔
2: Eigennamen			✔
Polariteit van de eigennamen			✔
Begeleiding van de eigennamen			✔
3: Acties		✔	✔
Waarschijnlijkheid dat de actie is of wordt uitgevoerd			✔
Tijdsaanduiding van de actie			✔
Begeleiding van de actie			✔
4: Locaties			✔
5: Domeinen van de tekst			✔
6: Polariteit van de gehele tekst	✔		✔
7: Begeleiding van de gehele tekst			✔
8: Tijdsaanduiding van de gehele tekst			✔
9: Tegenstrijdigheid van de gehele tekst			✔
10: Subjectiviteit / objectiviteit		✔	

** OpinionFinder bepaalt gedeeltelijk de polariteit van de acties. Niet bij iedere actie geeft OpinionFinder een polariteit.*

In deze tabel is in één oogopslag duidelijk dat Open Amplify veruit de meeste mogelijkheden voor het analyseren van de teksten in huis heeft. Deze mogelijkheden kunnen van pas komen bij het specifiek bepalen van sentiment

over een bepaald onderwerp. Daarnaast heeft Open Amplify ook als groot voordeel dat het de mate van polariteit bepaalt van -1.00 tot 1.0. Op deze manier wordt er niet alleen aangegeven of bijvoorbeeld een onderwerp positief of negatief is, maar ook de mate van positiviteit en negativiteit wordt door Open Amplify aangegeven.

2: De geïdentificeerde onderwerpen & acties kunnen alleen worden vergeleken tussen OpinionFinder en Open Amplify. De Bayesian methode identificeert geen onderwerpen en acties en wordt voor deze test achterwege gelaten. Het goed identificeren van de onderwerpen is belangrijk voor het bepalen van sentiment over het bepaalde onderwerp.

Van OpinionFinder en Open Amplify wordt hieronder de top 5 meest geïdentificeerde onderwerpen en acties weergegeven van de Twitterberichten over de onderwerpen Apple en Google.

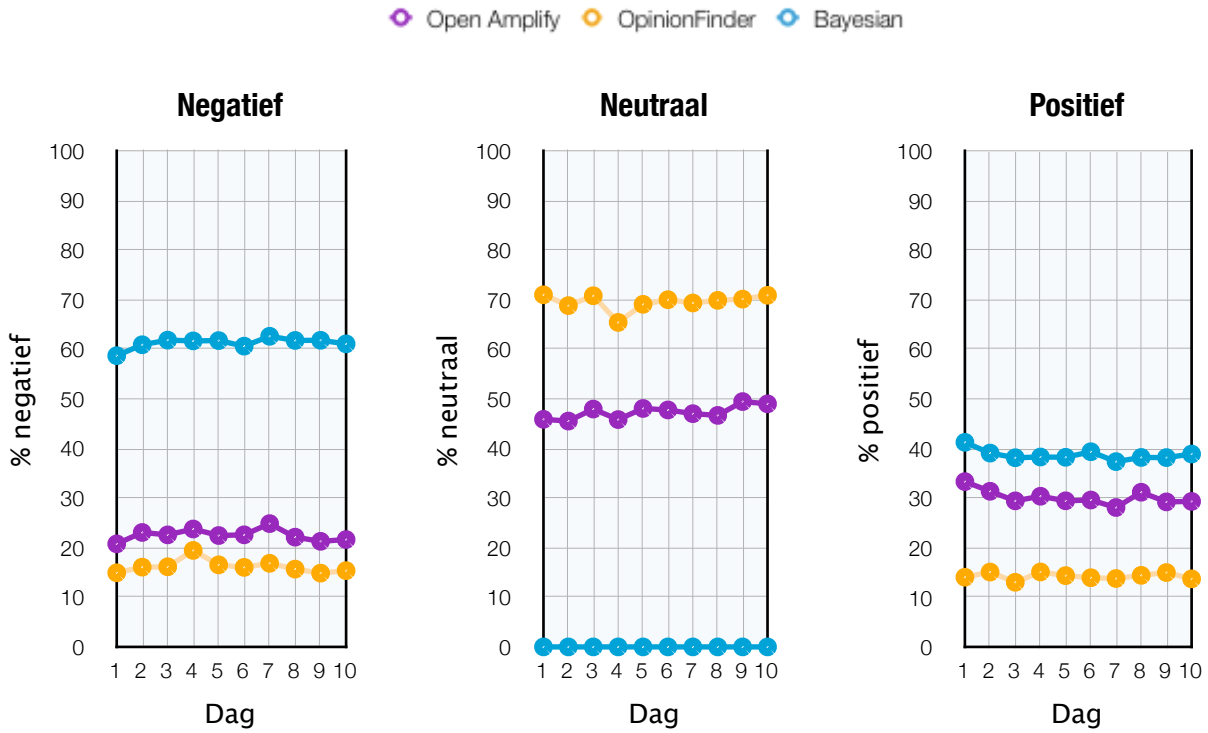
	Top 5 onderwerpen			
Nr	OpinionFinder		Open Amplify	
1	I	*5601x*	iPhone	*15470x*
2	You	*1483x*	Google	*13698x*
3	It	*859x*	iPad	*10101x*
4	He	*517x*	iPod	*9461x*
5	We	*368x*	Apple	*9378x*

	Top 5 acties			
Nr	OpinionFinder		Open Amplify	
1	Want	*3597x*	Say	*13087x*
2	Love	*2251x*	Move	*10892x*
3	Hate	*2005x*	Want	*9151x*
4	Think	*1958x*	Buy	*6932x*
5	Say	*1617x*	Communicate	*5904x*

In de bovenstaande tabellen is te zien dat OpinionFinder hele andere onderwerpen identificeert dan Open Amplify. Het lijkt erop dat OpinionFinder gebruik maakt van oude databestanden die de relatief nieuwe onderwerpen zoals iPhone en iPad niet kan identificeren, deze onderwerpen zijn namelijk helemaal niet geïdentificeerd door OpinionFinder. Daarnaast heeft OpinionFinder bijvoorbeeld het onderwerp ‘Apple’ 218x geïdentificeerd en komt het niet eens in de top 5 voor terwijl hij door Open Amplify 9378x is geïdentificeerd. Open Amplify identificeert veel vaker een onderwerp geeft voor ieder geïdentificeerd onderwerp extra informatie over het onderwerp zoals polariteit en begeleiding.

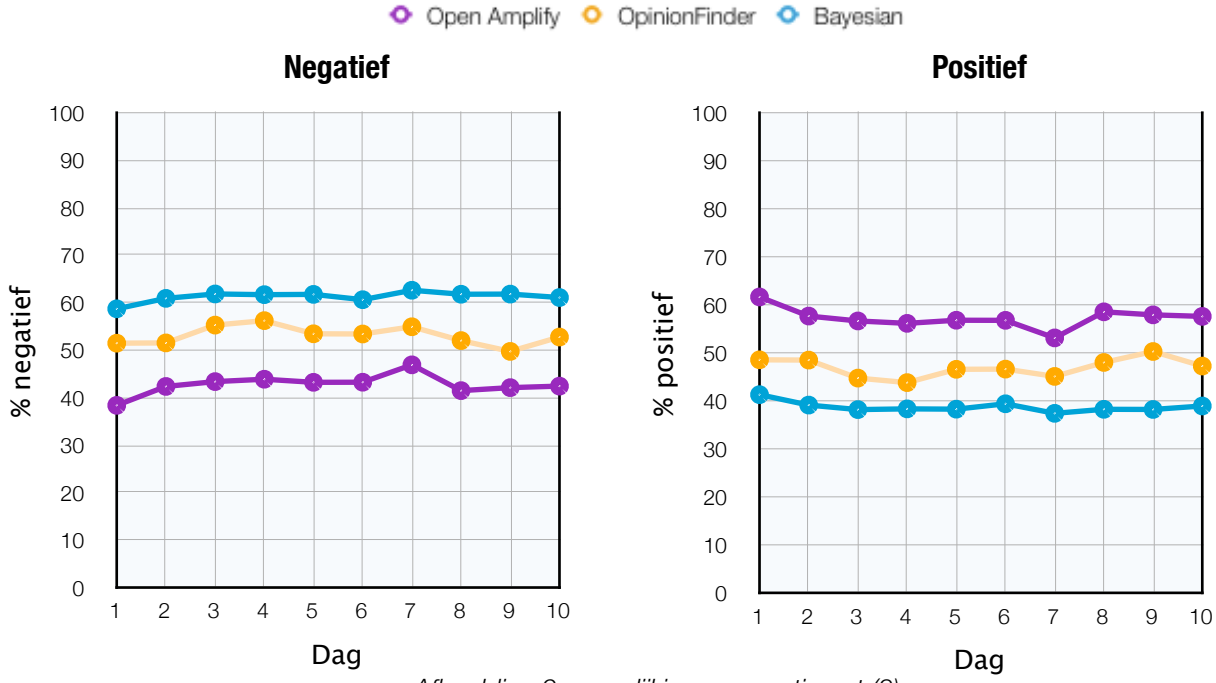
De acties die door OpinionFinder en Open Amplify zijn geïdentificeerd komen wel redelijk overeen, OpenAmplify heeft een actie zoals ‘want’ echter ongeveer 2,5x zo vaak geïdentificeerd als OpinionFinder. Buiten het feit dat OpinionFinder veel acties niet herkent is Open Amplify ook bruikbaarder dan OpinionFinder doordat Open Amplify meer informatie geeft bij iedere actie zoals tijdsaanduiding, begeleiding en de waarschijnlijkheid dat de actie is of wordt uitgevoerd.

3: Voor de vergelijking van sentiment van de drie methoden over een periode van 10 dagen wordt er gebruik gemaakt van verschillende afbeeldingen (zie afbeelding 8 en 9). In deze afbeeldingen wordt weergegeven wat het percentage van negatieve, neutrale en positieve Twitterberichten is volgens de drie onderzochte methoden.



Afbeelding 8: vergelijking van sentiment (1)

Zoals in de bovenstaande afbeelding is te zien maakt de Bayesian methode geen onderscheid in neutrale tekst. Hierdoor ligt het percentage van positieve en negatieve berichten van de bayesian methode veel hoger dan van de andere methoden. Om de vergelijking beter te kunnen maken is er voor gekozen om van de andere methoden de neutraal geïdentificeerde tekst niet mee te nemen in de vergelijking. De nieuwe afbeelding wordt hieronder weergegeven.



Afbeelding 9: vergelijking van sentiment (2)

In afbeelding 9 is te zien dat de verandering van het positieve en negatieve sentiment dat is bepaald d.m.v. de drie verschillende methoden redelijk dezelfde richting op gaat. Zoals bijvoorbeeld op dag 7 is te zien dat bij alle methoden het sentiment negatiever is dan de dag ervoor en de dag erna. Alle methoden zijn in staat om de

verandering van sentiment waar te nemen. Open Amplify heeft bij de drie geteste methoden de grootste volatiliteit[27] en bij de Bayesian methode is het zo goed als stabiel.

4: De nauwkeurigheid van de drie geteste methoden is ook van groot belang. Zoals eerder is beschreven zijn alle drie de methoden in staat om de verandering van sentiment waar te nemen maar hoe vaak hebben de methoden een tekst goed positief, neutraal of negatief geclassificeerd.

Voor het bepalen van de nauwkeurigheid is er handmatig van 100 willekeurige Twitterberichten bepaald wat de polariteit is. Hierbij is onderscheid gemaakt tussen positief, neutraal en negatief. Vervolgens is er gekeken hoe vaak de verschillende methoden dezelfde polariteit bepalen als de handmatige methode. De resultaten hiervan staan in de onderstaande tabel.

	Bayesian		OpinionFinder		Open Amplify	
	Goed	% goed	Goed	% goed	Goed	% goed
Positief:	*13 van de 26*	50%	*12 van de 26*	46,2%	*16 van de 26*	61,5%
Neutraal:	*0 van de 46*	0%	*37 van de 46*	80,4%	*36 van de 46*	78,3%
Negatief:	*21 van de 28*	75%	*12 van de 28*	42,9%	*22 van de 28*	78,6%
Totaal:	*34 van de 100*	**34%**	*61 van de 100*	**61%**	*74 van de 100*	**74%**

Deze test is niet 100% waterdicht maar geeft toch een goede indruk van de nauwkeurigheid. Zoals in bovenstaande tabel te zien is heeft Open Amplify met 74% de hoogste nauwkeurigheid. Zowel bij de positieve, neutrale en negatieve Twitterberichten scoort Open Amplify het beste.

Conclusie:

Open Amplify scoort van de drie geteste methoden op alle vergelijkingen het beste, het heeft veruit de meeste mogelijkheden, identificeert als beste de onderwerpen en acties, kan door de hoge volatiliteit het beste veranderingen in sentiment waarnemen en is met de hoogste nauwkeurigheid als beste in staat om sentiment te bepalen.

[27] Volatiliteit is een maatstaf voor de dagelijkse, wekelijkse, maandelijkse of jaarlijkse afwijking naar boven of naar beneden van een reeks ten opzichte van het gemiddelde.

3.4 Brondata verzamelen

In hoofdstuk 3.3 zijn er een drietal methodes beschreven en vergeleken voor het bepalen van sentiment d.m.v. tekst. Om een goed en betrouwbaar beeld te krijgen van het sentiment over een bepaald onderwerp zijn er grote hoeveelheden tekst nodig, er kunnen geen conclusies getrokken worden op basis van het sentiment van één willekeurige tekst. In dit hoofdstuk worden er drie methoden voor het verzamelen van brondata beschreven. Twee van deze methoden zijn gebruikt voor dit onderzoek.

3.4.1 Twitter

In de vorige hoofdstukken is de naam 'Twitter' al vaak naar voren gekomen. Twitter kwam in maart 2006 voor het eerst beschikbaar en sindsdien heeft het een gigantische groei ondergaan. Via het sociale netwerk versturen mensen wereldwijd dagelijks miljoenen berichten van maximaal 140 karakters lang. Naast dat gebruikers zelf berichten kunnen plaatsen kan ook automatisch de berichten van anderen gevolgd worden. Zodra men zichzelf aanmeldt als 'follower' van iemand anders, dan verschijnen zijn berichten automatisch op het scherm.

Omdat de doelgroep van Twitter erg breed is en er erg veel gebruik van wordt gemaakt is Twitter uitermate geschikt voor het bepalen van sentiment van een hele brede doelgroep.

3.4.1.1 Twitter API

Doordat Twitter een uitgebreide API beschikbaar stelt kunnen ontwikkelaars eenvoudig gebruik maken van de mogelijkheden van Twitter. Zo kunnen ontwikkelaars eenvoudig applicaties ontwikkelen waar alle functionaliteiten van Twitter in opgenomen kunnen worden. Er zijn al honderden Twitter clients ontwikkeld met behulp van de Twitter API.

Om automatisch sentiment te bepalen over verstuurde Twitterberichten moet er gebruik gemaakt worden van de Twitter API. In de meeste gevallen wordt sentiment bepaald over een van te voren gedefinieerd onderwerp, het moet dus mogelijk zijn om op een bepaald onderwerp te zoeken.

Helaas is het niet mogelijk om ruime historische data te verkrijgen via de Twitter API. Ook niet als er via de search API op een keyword wordt gezocht. Als er via de Twitter API op een keyword wordt gezocht worden alleen de resultaten van maximaal een paar dagen oud weergegeven. Dit zou betekenen dat er dagelijks (of meerdere malen per dag) de search API aangeroepen zou moeten worden. Wanneer de API te laat wordt aangeroepen zou dit resulteren in incomplete data. Dit is het grootste nadeel van de search API van Twitter.

Twitter heeft ook een streaming API beschikbaar gesteld. Met de streaming API is het mogelijk om gratis 1% van alle Twitterberichten realtime binnen te halen. Deze 1% van de Twitterberichten wordt door Twitter willekeurig gekozen van alle publieke Twitterberichten die verstuurd worden.

Ook is het via de streaming API mogelijk om gratis tot maximaal 400 keywords op te geven waarvan vervolgens 100% van de tweets realtime binnen komt die overeenkomsten heeft met 1 of meerdere opgegeven keywords. Tevens kan er gefilterd worden op userid en geografische locatie.

Als deze limieten van 1% en 400 keywords niet voldoende zijn kan er tegen betaling toegang verkregen worden tot firehose. Met firehose is het mogelijk om toegang te krijgen tot meer Twitterberichten. Firehose is een handmatige toelating en de kosten zijn afhankelijk van het business model. Gehele toegang tot firehose is niet altijd noodzakelijk. Er kan ook gekozen worden voor toegang tot subsets van de Twitterberichten.

De Twitter streaming API is uitermate geschikt voor een geautomatiseerd systeem dat sentiment bepaald over Twitterberichten. Voor dit onderzoek is er geen toegang nodig tot firehose. 1% van alle Twitterberichten en 400 keywords geeft voldoende inzicht in het algemene sentiment van Twitter.

Een groot nadeel is dat sentiment vaak alleen over het in het Engels geschreven tekst bepaald kan worden. Via Twitter wordt er in tientallen verschillende talen gecommuniceerd. Zowel met de streaming API als de search API van Twitter zullen ook deze (vreemde) talen binnen gehaald worden. Het is niet mogelijk om bij Twitter op taal te filteren. In de volgende paragraaf wordt uitgelegd hoe de Twitterberichten die zijn geschreven in een vreemde taal gefilterd kunnen worden.

3.4.1.2 Language detection

Zoals in de vorige paragraaf staat vermeld wordt er via Twitter veel gecommuniceerd in talen anders dan Engels. Buiten deze vreemde talen wordt er ook veel gecommuniceerd in het Engels waar de diverse methodes om sentiment te bepalen niet goed mee om kunnen gaan. Dit komt omdat het vaak onduidelijke of onvolledige zinnen zijn. Bijvoorbeeld: “RT @jaymohits: RT @L_Tido: Jay-z and kanye "Watch The Throne" Breaks iTunes Sales Record â—€ That's d Koko #WTT”.

Om ervoor te zorgen dat alle onduidelijke en niet Engelse Twitterberichten er uit worden gefilterd kan er gebruik gemaakt worden van ‘Text_LanguageDetect’. Dit is een gratis tool dat eenvoudig geïnstalleerd kan worden op een Linux machine. Na installatie kan er vanuit PHP een klasse aangeroepen worden waarmee de taal van een ingevoerde tekst in een fractie van een seconde herkend kan worden. ‘Text_LanguageDetect’ herkent 52 talen en geeft van iedere ingevoerde tekst per taal een mate van zekerheid dat een tekst in die taal geschreven is. Onderstaand voorbeeld geeft aan dat de ingevoerde tekst waarschijnlijk in het Duits is geschreven. Hoe hoger de waarde is dat erachter staat, hoe zekerder het script is.

```
Array(
    [german] => 0.407037037037
    [dutch] => 0.288065843621
    [english] => 0.283333333333
    [danish] => 0.234526748971
)
```

Hieronder wordt een voorbeeld gegeven van een test van ‘Text_LanguageDetect’. Er zijn 200.000 Twitterberichten over de onderwerpen apple en google door het language detect script gelopen. Met een grens van English ≥ 0,35 blijven er 10.140 Twitterberichten over (5,07%). Van deze overgebleven Twitterberichten is vrijwel zeker dat het duidelijke Engelse zinnen betreft.

Voor de verbeelding van deze test staan hieronder voorbeelden van willekeurige Twitterberichten die door deze test heenkomen, en willekeurige Twitterberichten die afvallen.

Voorbeeld van Twitterberichten die door de test heenkomen:

- @htc happens on other android devices too this is not a sense feature.
- This apple good as hell or im hungry one or the other
- What's this Google Plus thing :/
- Playing with the apple magic pad, so far, really like it .. perfect for #lion #apple
- Google is pretty disgusting. What a bunch of patent-babies and two-faced bastards.
- I cant believe i found an android yesterday :3
- after iphone 4 here comes iphone 5. .is there iphone 69 too?
- My last google search: "history of the couch". Sums up my weekend. #beantownscouch to the next one to the next one...
- Watch The Throne sold 290K first week on ITunes alone?
- Song went on itunes show some love and buy it :) love my fans couldnt do this without my family :)

Voorbeeld van Twitterberichten die niet door de test zijn gekomen:

- RT @9to5mac: Starbucks: Have a latte and download a free iPhone app http://t.co/k4NAguk
- Une biographie de Steve Jobs, le patron d'Apple, sera en vente en novembre http://t.co/sqEGfmK
- @tima1979 Ð°ÐºÑ†Ð¸Ñ oÑ, Apple: iPhone 3G Ð·Ð° 2000Ñ€ÑƒÐ±Ð»ÐµÐ¹! http://t.co/LozFMqa
- GTA Liberty City en Google map: Â¿Quien no ha jugado alguna vez a GTA? si me responden que no se los dejo como ta... http://t.co/PpR5yf3
- El efecto Panda se hace notar en las webs que copian contenido... http://t.co/obYeuGL
- Hy ziet er echt niet meer uit, en moet er nog 2 jaar mee doen. Ik wil een iPhone
- @palesamokemane lol olata hle mme ka nako e! Le lesedi fm eo fitile..google lols
- iPaard.. op iPad! #inderdaad #hhiihhiii
- ipad 3g: images Apple iPad 3G http://t.co/k4OIljZ
- RT @jaymohits: RT @L_Tido: Jay-z and kanye "Watch The Throne" Breaks iTunes Sales Record â—€ That's d Koko #WTT

In één oogopslag is al duidelijk dat de Twitterberichten die door de test heenkomen veel Engelser zijn dan de berichten die niet door de test heen komen.

De waarde ≥ 0,35 kan uiteraard verhoogd of verlaagd worden, dit is direct terug te zien in de resultaten. De kwaliteit van de Twitterberichten die worden doorgelaten door ‘Text_LanguageDetect’ wordt direct verhoogd of verlaagd. Door een handmatige steekproef is de waarde van ≥ 0,35 vastgesteld, er blijven genoeg Twitterberichten over en de kwaliteit van de Twitterberichten is voldoende voor het bepalen van sentiment.

3.4.2 Artikelen uit kranten en tijdschriften

Journalisten van kranten en tijdschriften schrijven vele artikelen die door een groot publiek worden gelezen. Omdat men wordt beïnvloed door het lezen van deze artikelen bepalen ze voor een groot gedeelte het sentiment. Wanneer journalisten een bepaald onderwerp erg negatief in beeld brengen zou dit directe invloed hebben op het sentiment van dit onderwerp.

OpinionFinder en voornamelijk Open Amplify zijn goed in staat om het sentiment van deze artikelen vast te stellen. Echter voor dat dit op grote wijze kan gebeuren moeten eerst al die artikelen verzameld worden. Er zijn verschillende grote databanken die al die nieuwsartikelen verzamelen, LexisNexis is er daar één van. Bij LexisNexis is het mogelijk om op keywords te zoeken en alle nieuwsartikelen waar dit keyword in voor komt te exporteren naar tekstbestanden. Het is zelfs mogelijk om artikelen op te zoeken die ruim 10 jaar terug geschreven zijn. D.m.v. LexisNexis is het redelijk eenvoudig om grote hoeveelheden tekstbestanden te verzamelen over een bepaald onderwerp.

Zodra alle artikelen zijn geëxporteerd naar tekstbestanden moeten ze nog goed worden geïmporteerd in de database die wordt gebruikt voor het opslaan en analyseren van alle artikelen. Omdat er vaak verscheidene artikelen in één tekstbestand staan is het erg moeilijk om de titel, de tekst, de datum en de bron van iedere tekst apart in de database op te slaan. Om dit toch voor elkaar te krijgen is er een script in PHP geprogrammeerd die de tekstbestanden uitleest en alle artikelen opslaat.

Een groot nadeel van LexisNexis is dat het proces niet volledig geautomatiseerd kan worden. Telkens moeten de artikelen handmatig geëxporteerd worden. Het is niet mogelijk om dit proces via LexisNexis te automatiseren. Om dit proces toch volledig te automatiseren zouden er verschillende contracten moeten worden afgesloten met de uitgevers van de kranten en tijdschriften.

3.4.3 Webcrawler

Een andere methode voor het verzamelen van grote hoeveelheden tekst is het ontwikkelen van een webcrawler. Een webcrawler is een geautomatiseerd computersysteem dat het wereldwijde web doorbladert. Van iedere webpagina die de webcrawler tegenkomt maakt het systeem een lokale kopie. Van deze lokale kopie kan vervolgens alle tekst ontleed en verwerkt worden met bijvoorbeeld Open Amplify. Helaas zullen er verschillende webpagina's niet publiekelijk toegankelijk zijn die wel interessant zijn voor het bepalen van sentiment. Denk hierbij bijvoorbeeld aan specifieke fora die zijn afgeschermd met een gebruikersnaam en wachtwoord.

Het voordeel van een webcrawler is dat echt het gehele internet afgespeurd kan worden. Hierdoor zou men niet meer afhankelijk zijn van Twitter en LexisNexis.

Vanwege de complexiteit en de omvang valt het ontwikkelen van een webcrawler buiten de scope van dit project.

4. Producten

In dit hoofdstuk wordt er een tweetal producten op basis van sentiment onderzocht en uitgewerkt. Deze producten zou Oxin aan kunnen bieden om de diversiteit van de producten en services te vergroten. Bij product één wordt geprobeerd om op basis van sentiment de beurs te voorspellen. Bij product twee wordt sentiment gebruikt als marketing tool voor bedrijven.

4.1 Sentiment als voorspellende kracht

Op de beurs wordt er veel gehandeld op basis van sentiment. Wanneer een bedrijf een positief sentiment heeft gaat men er vanuit dat het aandeel in de toekomst zal gaan stijgen, er zal een grotere vraag naar dit aandeel ontstaan waardoor de waarde van het aandeel zal gaan stijgen. Met dit gegeven kunnen er wellicht positieve rendementen[28] behaald worden door te handelen op basis van sentiment dat bepaald is met behulp van Open Amplify. Johan Bollen heeft op basis van sentiment de Dow Jones weten te voorspellen[29], reden genoeg om dit ook te onderzoeken.

Voor dit onderzoek zijn er door Open Amplify 295.000 artikelen verwerkt waar het woord Dow Jones minimaal één keer in voor komt. Deze artikelen komen bij LexisNexis vandaan en zijn van januari 2001 tot en met december 2010.

Na alle artikelen te hebben verwerkt is de database ruim 3,3 GB groot en bestaat het uit ruim 56.000.000 rijen. Met de omvang van deze database kan er een lange tijd onderzoek gedaan worden. Al snel kwam naar voren dat op basis van deze verwerkte artikelen er een kans bestond om de beurs te kunnen voorspellen, er kwamen verschillende modellen naar voren waarmee positieve rendementen behaald kunnen worden maar veel van deze modellen zijn niet altijd even significant.

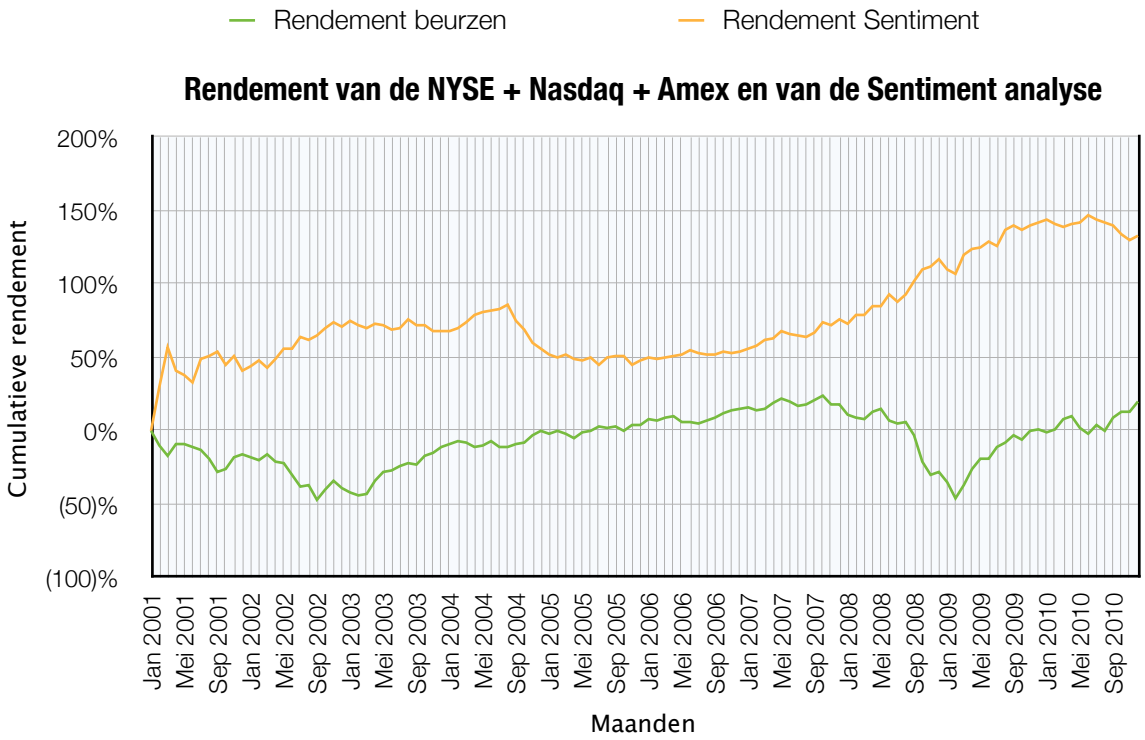
Van de onderzochte methoden om de beurs te kunnen voorspellen is er één die als beste uit de test komt, deze resultaten zijn boven alle verwachtingen. In onderstaande grafiek ziijn de cumulatieve[30] rendement van de NYSE, Nasdaq en Amex[31], en het cumulatieve rendementen van de methode die handelt op basis van sentiment te zien.

[28] Rendement is de opbrengst van een belegging of investering uitgedrukt in een percentage van het daarmee samenhangende geïnvesteerde bedrag.

[29] Johan Bollen, Huina Mao en Xiao-Jun Zeng (2010). Twitter mood predicts the stock market.

[30] Het cumulatief rendement is het steeds bij elkaar opgetelde rendement.

[31] De NYSE, Nasdaq en Amex zijn de drie grootste Amerikaanse aandelen beurzen.



Afbeelding 10: rendement van de beurs & rendement van het model op basis van sentiment

Zoals in bovenstaande afbeelding is af te lezen hebben de NYSE, Nasdaq en Amex gezamenlijk in tien jaar tijd ongeveer 20% rendement behaald, de theorie die is gebaseerd op basis van sentiment heeft in dezelfde periode 135% rendement behaald.

Om de bovenstaande rendementen te behalen is het volgende model gebruikt.

- Alle nieuwsartikelen worden verzameld en verwerkt met Open Amplify. Voor deze methode zijn geen Twitterberichten gebruikt;
- Met een query[32] wordt per dag het gemiddelde sentiment bepaald van alle onderwerpen waar het woord economy in voor komt;
- Alle aandelen van NYSE, Nasdaq en Amex van de afgelopen tien jaar worden ingeladen. Het gaat hierbij om de naam van het aandeel en de dagelijkse prijs. Voor iedere maand zijn het gemiddeld 4750 aandelen;
- Met een regressie[33] wordt per maand de gevoeligheid[34] van de aandelen t.o.v. het dagelijkse sentiment bepaald;
- Vervolgens worden er per maand 6 portfolio's[35] gemaakt van alle ingeladen aandelen;

[32] Met een query wordt een opdracht aan de database gegeven om bepaalde gegevens op te halen.

[33] Een regressie is een statistische functie waarmee de afhankelijkheid tussen twee getallen reeksen kan worden bepaald.

[34] Hoe gevoeliger een aandeel is, hoe meer het aandeel afhankelijk is van het bepaalde sentiment.

[35] Een portfolio of portefeuille is de benaming voor een geheel van samengestelde aandelen. Beleggers houden vaak een portfolio van verschillende aandelen aan, om zo het risico op koersverlies te beperken.

    - Hiervoor worden eerst alle aandelen gesorteerd op gevoeligheid. Er wordt per maand een selectie gemaakt van de 10% meest gevoelige aandelen en van de 10% minst gevoelige aandelen in die maand;
    - Vervolgens worden de aandelen gesorteerd op grootte. Er wordt een selectie gemaakt van kleine en grote aandelen. Eerst wordt hiervoor de gemiddelde grootte van een aandeel bepaald. Vervolgens is de selectie van kleine aandelen kleiner dan het gemiddelde en de selectie van grote aandelen is groter dan het gemiddelde;
- Nu zijn er iedere maand 6 verschillende portfolio's;
    - Portfolio 1: Kleine bedrijven die niet gevoelig zijn voor het sentiment;
    - Portfolio 2: Kleine bedrijven die tussen gevoelig en niet gevoelig in zit;
    - Portfolio 3: Kleine bedrijven die gevoelig zijn voor het sentiment;
    - Portfolio 4: Grote bedrijven die niet gevoelig zijn voor het sentiment;
    - Portfolio 5: Grote bedrijven die tussen gevoelig en niet gevoelig in zit;
    - Portfolio 6: Grote bedrijven die gevoelig zijn voor het sentiment;
- Van deze zes verschillende portfolio's wordt er iedere maand portfolio 3 en 6 gekocht. Op de portfolio's 1 en 4 gaan we short[36].
- De geopende posities worden na zes maanden gesloten en de rendementen worden genoteerd;
- Het maken van de portfolio's en openen en sluiten van de posities gebeurt iedere maand opnieuw;

Het gebruikte model is een "zero cost portfolio", met het geld dat is ontvangen door short te gaan op de portfolio's 1 en 4 kunnen in theorie de portfolio's 3 en 6 gekocht worden. In een perfecte wereld zou er met dit model zonder eigen geld gehandeld kunnen worden. Tevens zijn de portfolio's "value weighted", in de portfolio's krijgen de aandelen met een kleine marktwaarde een kleiner gewicht dan aandelen met een grote marktwaarde. De resultaten van dit model zijn zeer significant[37].[38]

Het product 'sentiment als voorspellende kracht' kan als een abonnementsvorm worden aangeboden via een website. De resultaten van deze methode zullen voor veel mensen erg interessant zijn. Op de website worden de in het verleden behaalde rendementen weergegeven en tegen betaling krijgt men iedere maand te zien welke portfolio's er gekocht moeten worden en op welke portfolio's men short moet gaan om dezelfde rendementen te behalen. Klanten betalen voor informatie die uit het model voort vloeien.

[36] Short gaan is het verkopen van aandelen die men niet in bezit heeft, om zo te kunnen profiteren van een daling van de beurskoers.

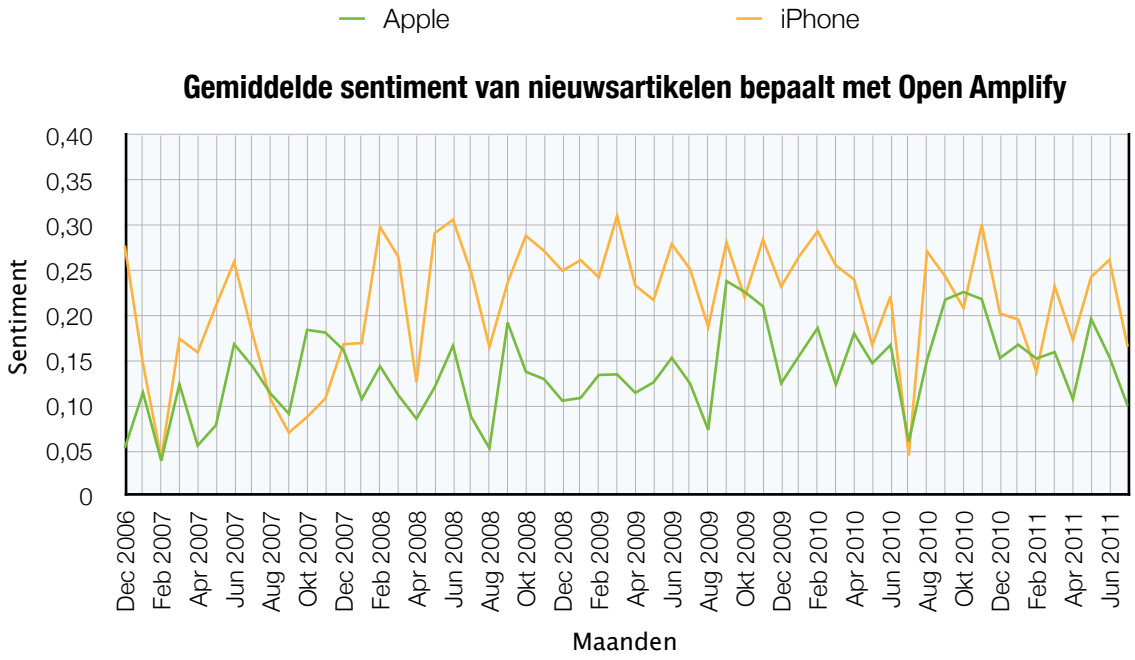
[37] Significant is een statistisch begrip dat betekent dat de kans dat een bepaald verschijnsel voorkomt groter is dan het toeval normaal gesproken wil.

[38] Meer informatie over dit model is te vinden in de paper: Dhr. S.C.E. Dekker, BSc (2011). Is headline risk priced in?

4.2 Sentiment als marketing tool

Het tweede mogelijke product is ‘Sentiment als marketing tool’. Veel bedrijven hebben diverse marketing campagnes lopen. Een doel van een marketingcampagne kan bijvoorbeeld zijn om een positievere beeldvorming te creëren bij de doelgroep. De verkoopcijfers tonen nog niet altijd direct het succes van de campagne aan. Om dit te kunnen meten willen marketeers weten wat het imago van hun merk, product of dienst is. In de meeste gevallen wordt het imago van een merk, product of dienst gemeten d.m.v. enquêtes. Dit imago kan nu ook perfect worden gemeten d.m.v. sentiment. Het gemiddelde sentiment van nieuwsartikelen en Twitterberichten geeft een duidelijk beeld van hoe men over een merk, product of dienst schrijft. Is dit sentiment positief dan zou het imago ook positief zijn.

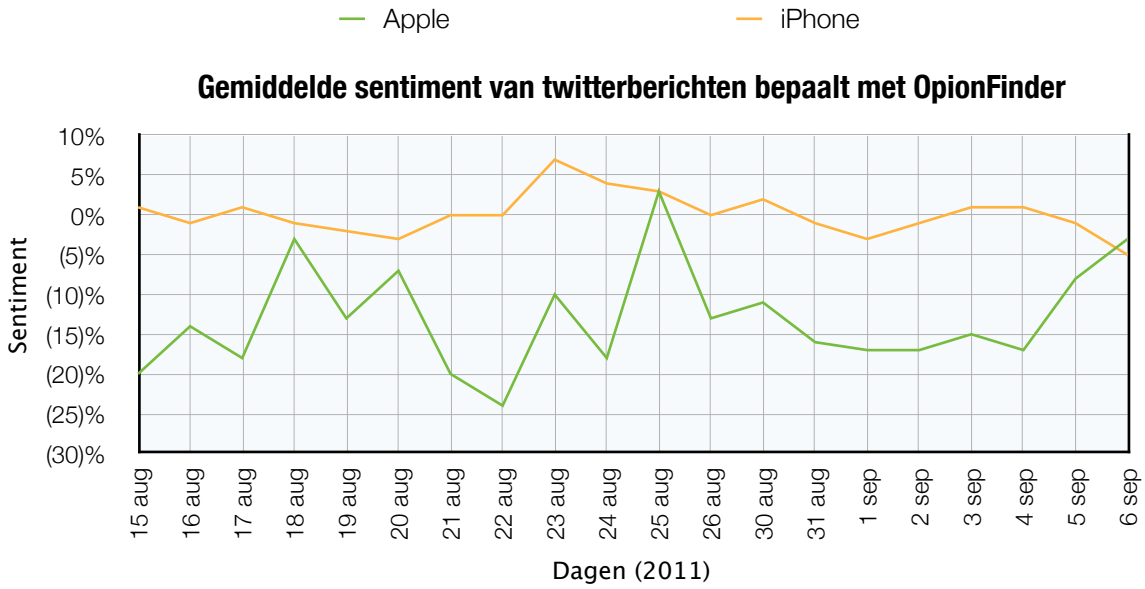
Om het product ‘sentiment als marketing tool’ te verduidelijken zijn er meer dan 75.000 nieuwsartikelen door Open Amplify verwerkt over de onderwerpen ‘Apple’ en ‘iPhone’. In onderstaande grafiek wordt een voorbeeld gegeven van het sentiment van deze onderwerpen over de periode december 2006 tot en met juli 2011.



Afbeelding 11: gemiddelde sentiment van Apple en iPhone (Open Amplify)

In afbeelding 11 is het opvallend dat het sentiment van ‘Apple’ en van de ‘iPhone’ altijd positief is geweest. Tevens zijn er nooit echt pieken in het sentiment maar voornamelijk alleen maar dalen. Met een correlatiecoëfficiënt van 0,39 zijn ‘Apple’ en ‘iPhone’ redelijk gecorreleerd aan elkaar. Zodra het sentiment van de ‘iPhone’ stijgt is er ook een grote kans dat het sentiment van ‘Apple’ stijgt en vice versa. Het verklaren van het verschil in sentiment over een gegeven periode is erg moeilijk. Neem bijvoorbeeld het lage sentiment van ‘Apple’ en ‘iPhone’ op juli 2010, er bestaat een grote kans dat dit komt door de Death Grip van de iPhone 4 maar dit kan niet worden afgelezen aan het verschil in sentiment, hiervoor zou er in deze periode meer onderzoek gedaan moeten worden naar de geschreven nieuwsartikelen. Wel is te zien dat in deze periode het sentiment van ‘Apple’ en ‘iPhone’ beide flink is gedaald. Dit zou kunnen betekenen dat het probleem met de ‘iPhone’ een groot probleem voor ‘Apple’ is geweest.

Naast het bepalen van sentiment over nieuwsartikelen kan er ook sentiment worden bepaald over Twitterberichten. Helaas vanwege de capaciteit van Open Amplify is het voor dit voorbeeld niet mogelijk geweest om de 500.000 Twitterberichten over 'Apple' en 'iPhone' door Open Amplify te verwerken. Daarom is er voor gekozen om het sentiment van de Twitterberichten te bepalen met OpinionFinder. Hieronder wordt het voorbeeld hiervan weergegeven.



Afbeelding 12: gemiddelde sentiment van Apple en iPhone (OpinionFinder)

Net als bij de nieuwsartikelen is het ook erg moeilijk om het verschil in sentiment te verklaren bij de Twitterberichten. In bovenstaande grafiek is op donderdag 25 augustus het sentiment van apple veel positiever, op deze dag heeft Steve Jobs bekent gemaakt om zijn functie als CEO op te geven, of dit ook daadwerkelijk positief is voor 'Apple' is nog maar de vraag.

Voor Apple is het sentiment en het verschil in sentiment van hun merk of product erg interessant. Omdat ze veel meer gegevens hebben over lopende campagnes, productlanceringen, persverklaringen, problemen met producten etcetera zijn ze beter in staat om het verschil van sentiment te verklaren.

Voor 'Apple' is het niet alleen interessant om te weten wat het imago is van hun eigen product of merk, ook het imago van concurrenten en gehele sectoren kan interessant zijn om te weten.

Het product 'sentiment als marketing tool' zou een website met webportal worden waar bedrijven tegen betaling een abonnement voor af kunnen sluiten. Met een abonnement heeft het bedrijf de mogelijkheid om één of meerdere woorden (bijvoorbeeld: 'Apple' en 'iPhone') op te geven waarover ze het sentiment willen monitoren. Op de webportal zullen vervolgens grafieken worden weergegeven die het verschil in sentiment over een bepaalde periode zullen weergeven. Dit sentiment zal bepaald worden met Open Amplify en heeft als input de Twitterberichten en nieuwsartikelen. Hieronder staat een overzicht van de belangrijkste functies die de webportal zou moeten hebben.

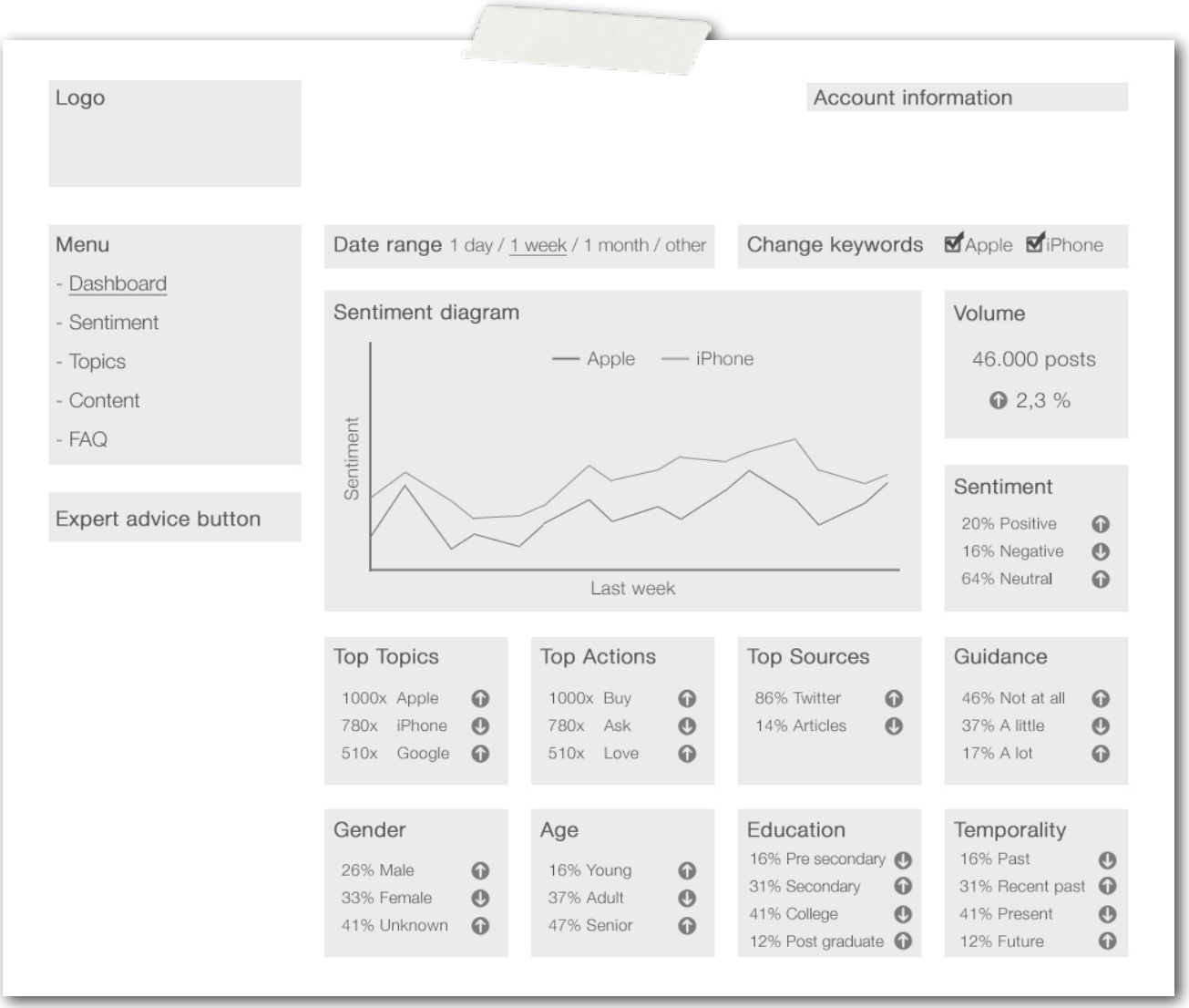
- Voordat een bedrijf toegang heeft tot de webportal moet er een abonnement worden afgesloten;
- Abonnementen moeten via de website kunnen worden afgesloten;

- Er moeten verschillende woorden opgeven kunnen worden waarover sentiment bepaald moet worden;
- Het sentiment over een bepaalde periode moet in grafieken worden weergeven, zoals bijvoorbeeld de grafieken hierboven;
- Het sentiment over een bepaalde periode moet in grafieken worden weergeven, waarvan het sentiment de tijdsaanduiding van de toekomst, heden of verleden aangeeft;
- Meerdere opgegeven onderwerpen moeten met elkaar vergelijken kunnen worden in een grafiek. Zo zou het bijvoorbeeld mogelijk moeten zijn om in één grafiek het sentiment van Apple en van Google weer te geven;
- De aangegeven periode moet gewijzigd kunnen worden: een datum range moet opgegeven kunnen worden en er moet aangegeven kunnen worden of de grafiek de resultaten per dag, per week of per maand moet weergeven;
- Er moet een overzicht worden weergegeven van de belangrijkste acties die gekoppeld zijn aan de opgegeven onderwerpen, met de waarschijnlijkheid dat de actie wordt uitgevoerd;
- Bij de grafieken moet duidelijk worden aangegeven wat de bron van het sentiment is: Twitter of nieuwsartikelen;
- Alle gescande berichten waar de opgegeven woorden in voor komen moeten via de webportal toegankelijk zijn;
- In de webportal moeten er statistieken worden getoond met het aantal berichten per bron, brontype en/of onderwerp van een opgegeven periode;
- Gegevens moeten geëxporteerd kunnen worden naar CSV bestanden;
- Via de webportal moet er een advies kunnen worden aangevraagd. Bedrijven kunnen bijvoorbeeld een specifieke vraag hebben waarom het sentiment in periode X veel lager is dan in periode Y. Door de verwerkte artikelen verder te onderzoeken valt dit verschil misschien wel te verklaren.

Naar aanleiding van de beschreven functies die de webportal zou moeten hebben zijn er drie wireframes uitgewerkt. Deze wireframes kunnen worden gezien als de bouwtekeningen van de webportal, ze geven een overzicht van de verschillende onderdelen die aanwezig zullen zijn. Het logo, het menu en de extra account informatie zullen op iedere pagina aanwezig zijn. Tevens staat er op iedere pagina onder het menu een knop waarmee gebruikers van de webportal extra advies kunnen vragen.

Wireframe 1 (het dashboard):
Zodra er is ingelogd door de gebruikers zou het dashboard worden weergegeven. Op het dashboard wordt alle algemene informatie beknopt weergegeven.

Van de opgegeven keywords wordt op het dashboard in een grafiek het gemiddelde sentiment weergegeven. Naast de grafiek staan er tien blokken die extra informatie geven over de verwerkte teksten waar de opgegeven keywords in voor komen. Denk hierbij onder andere aan het volume, de top topics, de top actions en de top sources. In ieder blok wordt met een pijl aangegeven of het betreffende item is gestegen of gedaald t.o.v. vorige periode. De periode die standaard wordt weergegeven is ‘afgelopen week’. Deze periode kan gewijzigd worden en alle informatie op het dashboard zal zich aanpassen aan de ingestelde periode.



Afbeelding 13: Wireframe 1 (het dashboard)

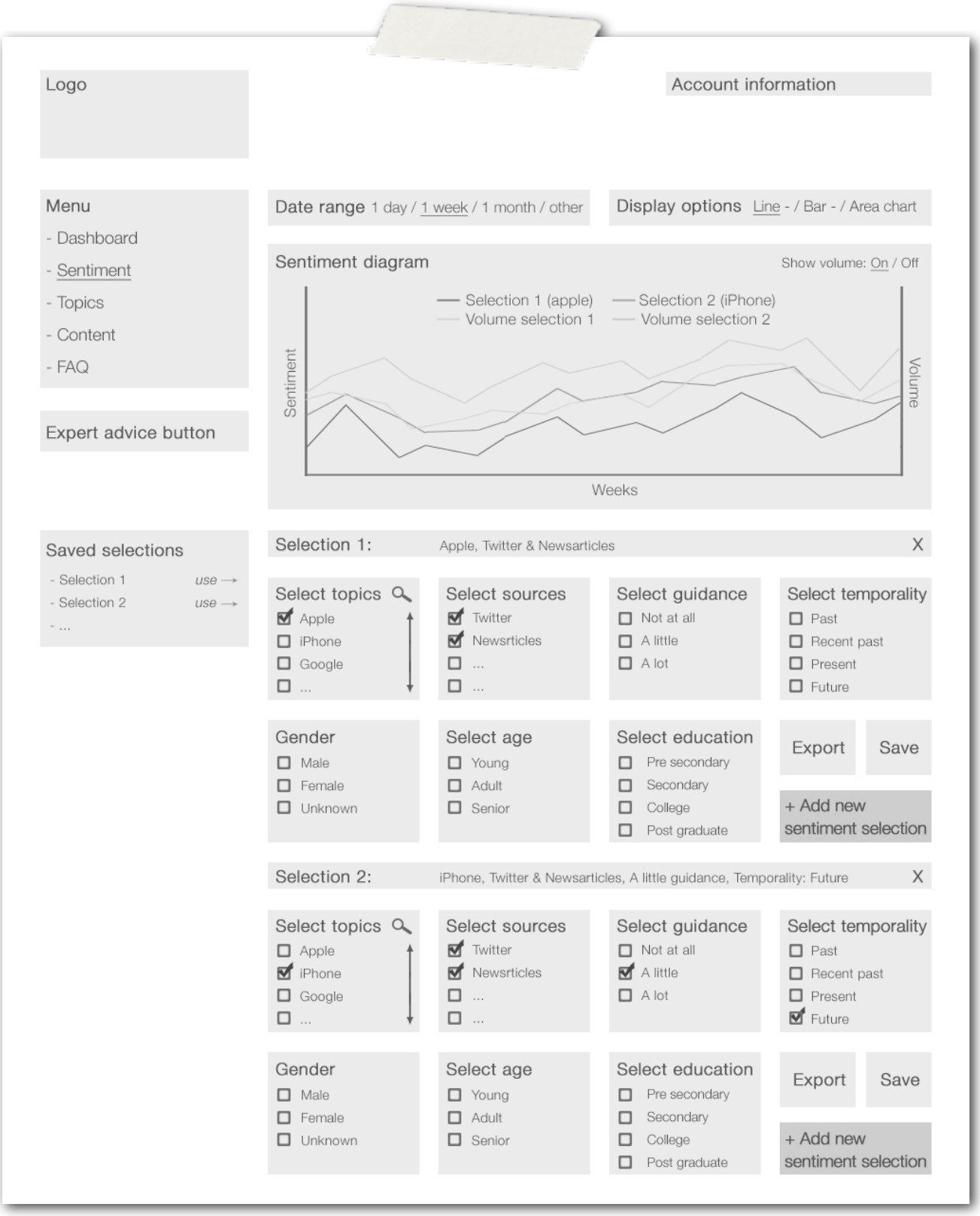
Wireframe 2 (sentiment):

In wireframe 2 draait het volledig om een nadere analyse van het sentiment zoals het op het dashboard wordt getoond. Dit sentiment dat wordt weergegeven in een grafiek kan op verschillende manieren met elkaar vergeleken worden door het aanmaken van meerdere ‘selections’. Er kunnen tot maximaal tien ‘selections’ in één grafiek worden vergeleken en iedere ‘selection’ kan volledig gespecificeerd worden. Een ‘selection’ kan worden opgeslagen en vervolgens op een later tijdstip weer ingeladen worden. De weergave van een grafiek kan worden gewijzigd en de gegevens van een selection kunnen worden geëxporteerd.

Een korte omschrijving van de selection wordt in de titelbalk van de selection weergegeven. Zodra op deze titelbalk wordt geklikt zou de selectie in- of uitklappen.

Net als op het dashboard kan ook op deze pagina de periode worden gewijzigd. Standaard wordt afgelopen week weergegeven.

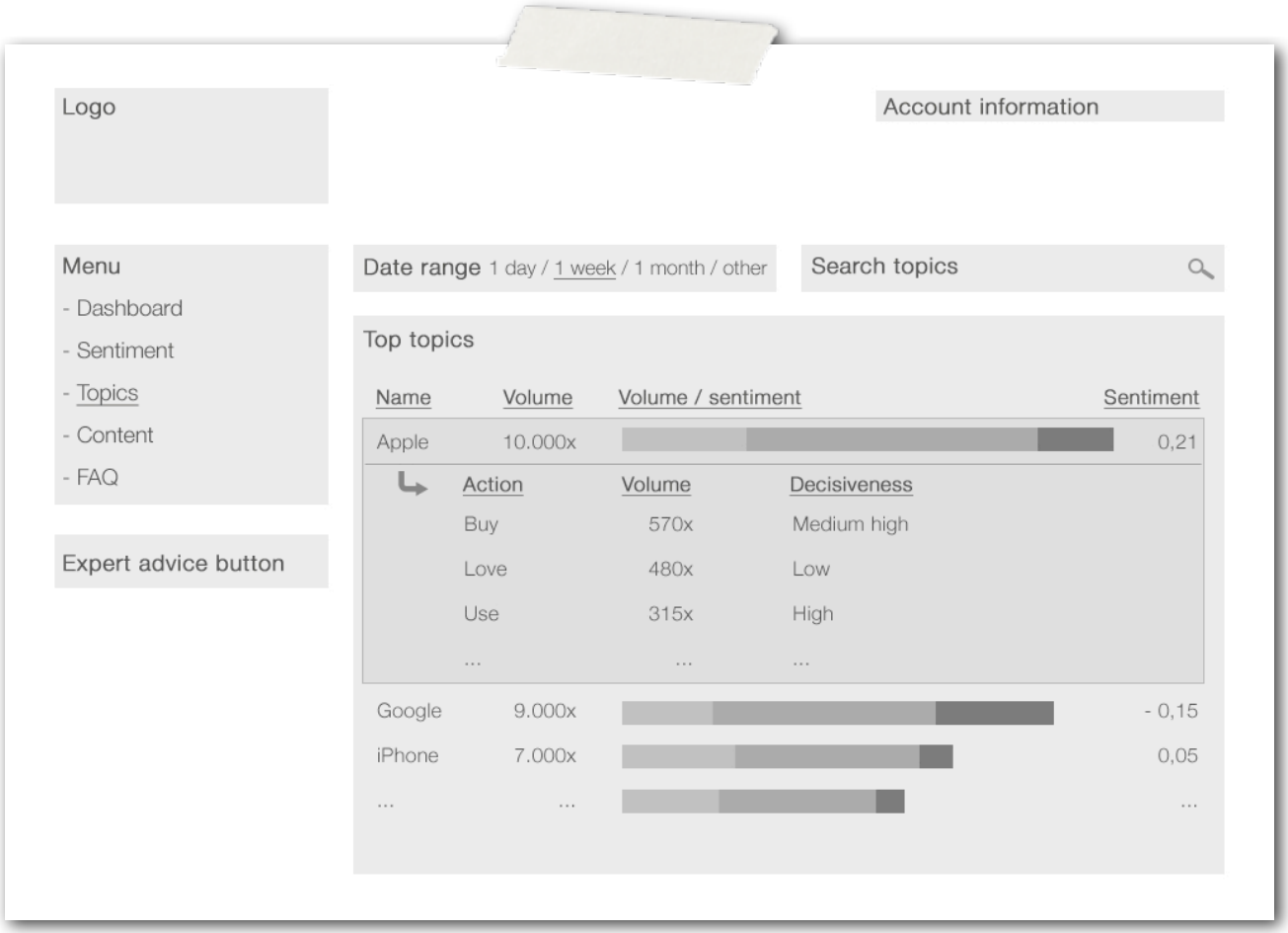
Zodra op de deze pagina op de knop ‘Expert advice’ wordt gedrukt zullen de huidige geselecteerde selecties worden meegestuurd naar Oxin.



Afbeelding 14: Wireframe 2 (sentiment)

Wireframe 3 (topics):

In wireframe 3 worden alle geïdentificeerde onderwerpen van de verwerkte teksten getoond. Van ieder onderwerp wordt vervolgens het volume en het gemiddelde sentiment weergegeven. Aan een onderwerp kunnen verschillende acties zijn gekoppeld, zodra er op een onderwerp wordt geklikt worden deze acties weergegeven. Van iedere actie wordt vervolgens weergegeven hoe vaak de actie voor komt en wat de gemiddelde waarschijnlijkheid is dat de actie is of wordt uitgevoerd. Boven in het scherm kan de periode worden gewijzigd en kan er worden gezocht op bepaalde onderwerpen. In onderstaand wireframe worden de acties die zijn gekoppeld aan het onderwerp 'Apple' weergegeven.



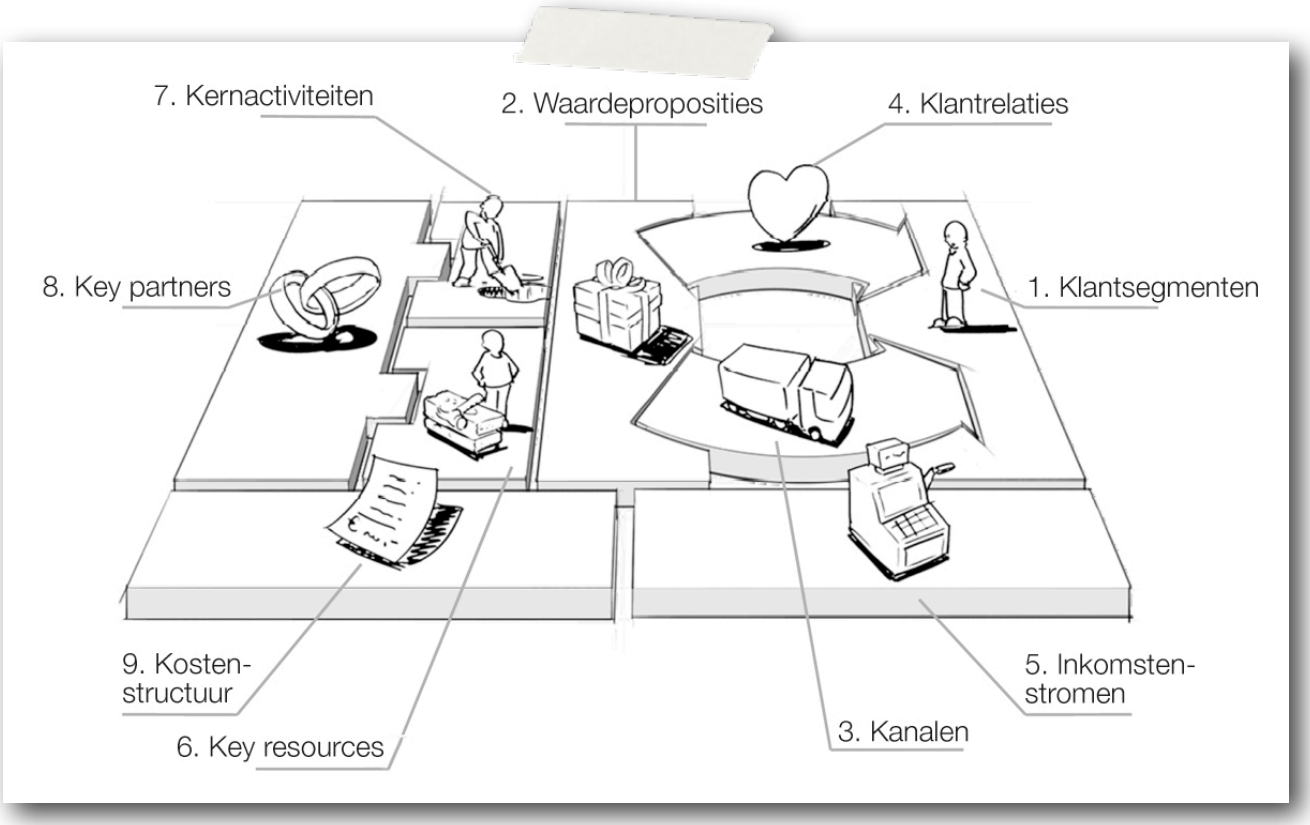
Afbeelding 15: Wireframe 3 (topics)

Er zijn twee pagina's van de webportal die niet in wireframes zijn uitgewerkt: de pagina 'content' en de pagina 'FAQ'. Op de pagina 'content' worden alle verwerkte teksten volledig weergegeven. Van iedere verwerkte tekst kunnen alle taalkundige gegevens worden weergegeven zoals Open Amplify dit heeft geïdentificeerd. Op de pagina 'FAQ' zullen alle veel gestelde vragen worden weergegeven en beantwoord.

5. Business model

Voor dit onderzoek zijn er drie methoden getest en vergeleken voor het bepalen van sentiment. Open Amplify komt het beste uit de vergelijking en is gebruikt voor het onderzoeken van twee producten die Oxin kan aanbieden op basis van sentiment. Er is voor gekozen om het tweede product: 'sentiment als marketing tool' uit te werken in dit business model. Dit product past veel beter bij de huidige producten en diensten die Oxin momenteel aanbiedt ten opzichte van het eerste product. Oxin is tenslotte geen financiële instelling maar een full service internet bureau. Dit betekent niet dat het eerste product 'sentiment als voorspelende kracht' geen goed product zou zijn om aan te bieden.

Dit business model beschrijft de grondgedachten van hoe Oxin door middel van het product 'sentiment als marketing tool' waarde creëert, levert en behoudt. In de negen hoofdstukken van dit business model komen vier hoofdgebieden van het product 'sentiment als marketing tool' aan bod: de klanten, het product, de infrastructuur en de financiële levensvatbaarheid. Voor dit business model is er gebruik gemaakt van het business model canvas (zie de afbeelding hieronder) dat is ontwikkeld door middel van co-creatie van 470 experts uit 45 landen.[39]



Afbeelding 16: Het business model canvas

[39] Alexander Osterwalder & Yves Pigneur (2009). Business Model Generatie.

Dit business model is geschreven vanuit het perspectief van de klant, en niet vanuit het perspectief van Oxin. Welke taken moeten onze klanten gedaan krijgen en hoe kunnen wij hen daarmee van dienst zijn? En dus niet: wat kan Oxin aan haar klanten verkopen? In dit business model wordt er onderzoek gedaan naar de haalbaarheid van het product 'Sentiment als marketing tool'.

5.1 Klantsegmenten

Deze paragraaf definieert een tweetal klantsegmenten die Oxin wil bereiken en bedienen met het product: 'Sentiment als marketing tool'. Zonder deze klanten zou Oxin niet winstgevend kunnen worden, de klanten vormen daardoor het hart van dit business model.

De twee klantsegmenten waar Oxin voor het product onderscheid in maakt zijn bedrijven waar veel in de engelse taal over wordt geschreven op sociale media en in kranten en tijdschriften, en bedrijven die afhankelijk zijn van het sentiment van een ander bedrijf of markt. Beide klant segmenten hebben specifieke klantbehoeften en verschillen in de distributiekanalen, klantrelaties, winstgevendheid en zijn bereid te betalen voor verschillende aspecten van het aanbod. Deze klantsegmenten maken een klein onderscheid tussen de marktsegmenten met kleine verschillen in de gewenste problemen en bestaande behoeften.

5.2 Waardeproposities

Oxin heeft de ambities om de problemen van haar klanten op te lossen en in de behoeften van haar klanten te voorzien met waardeproposities[40]. Deze waardeproposities zullen in deze paragraaf aan bod komen.

Nieuwheid:
Het automatisch bepalen van sentiment over geschreven tekst is pas sinds kort vanwege de technologische vooruitgang mogelijk en vanwege de globalisering is er steeds meer nieuws en overig geschreven tekst digitaal beschikbaar. Het product 'sentiment als marketing tool' is niet volledig nieuw. Er zijn al verschillende bedrijven die een vergelijkbaar product aanbieden in de vorm van een online knipselkrant. Bij deze knipselkranten worden alle online geschreven tekst over een bepaald onderwerp verzameld en inzichtelijk gemaakt. Oxin zou vernieuwend zijn door het product totaal op het sentiment en de verandering van sentiment te richten. Het sentiment zou op meerdere wijzen gespecificeerd, gepersonaliseerd en vergeleken kunnen worden.

Performance:
Van de drie geteste methoden voor het automatisch laten bepalen van sentiment over geschreven tekst is Open Amplify veruit de beste methode vanwege de mogelijkheden en nauwkeurigheid. De techniek die het mogelijk maakt om sentiment te bepalen over geschreven tekst is één kant van het verhaal. Het verzamelen en opslaan van de geschreven tekst is de andere kant. Wederom vanwege de technologische vooruitgang en de globalisering kan de performance voor het verzamelen en opslaan van deze teksten erg omhoog. Systemen kunnen zo ontwikkeld worden dat ze de geschreven tekst automatisch verzamelen en verwerken met Open Amplify. De verzamelde en verwerkte teksten zullen realtime via de webportal worden aangeboden aan de klanten.

[40] Met een waardepropositie wordt vastgesteld wat de toegevoegde waarde van een organisatie is.

Zodra alle teksten zijn verzameld, opgeslagen en verwerkt moet het nog aangeboden worden aan de klanten van Oxin, hiervoor zou de webportal uitermate geschikt zijn. Om dit goed en vloeiend te laten verlopen zouden de gebruikte servers goed ingericht moeten zijn, voldoende performance moeten hebben en zou de webportal zo geoptimaliseerd moeten zijn om zo snel- en met zo min mogelijk serverkracht de klant te kunnen bedienen.

De performance van het product ‘sentiment als marketing tool’ is voor beide klantsegmenten even belangrijk, voor zowel de bedrijven waar veel over wordt geschreven als voor de bedrijven waar minder veel over wordt geschreven is de performance van de webportal van groot belang.

Customization:
Het op maat afstemmen van het product ‘sentiment als marketing tool’ op de klanten van Oxin creëert extra waarde. Samen met Oxin kan de klant een set met keywords samenstellen die geanalyseerd zullen worden door het systeem. Alle artikelen en Twitterberichten waar één van de opgegeven keywords in voorkomen zullen worden opgeslagen en worden verwerkt met Open Amplify. Via de webportal zal vervolgens een gepersonaliseerde weergave van sentiment worden getoond op basis van de opgegeven keywords. De functionaliteiten van de webportal zal voor alle klanten hetzelfde zijn, de inhoud zou erg verschillen.

Zodra de klant extra informatie wil over een bepaalde verandering in sentiment, of als ze een andere specifieke vraag hebben kunnen klanten via de webportal een verzoek tot extra advies en informatie indienen. Oxin zou vervolgens het probleem handmatig gaan onderzoeken om zo aan de wensen van de klant te kunnen voorzien.

Ontwerp:
Het product ‘sentiment als marketing tool’ zal veel door marketeers van grote bedrijven worden gebruikt. Om deze marketeers eerder over de streep te halen om het product te gaan gebruiken zou de website en webportal een uitblinkend en innovatief ontwerp krijgen. De verwacht is dat de marketeers hier erg gevoelig voor zullen zijn. Tevens kunnen de marketeers de rapportages direct aan hun directieleden overhandigen.

Merk en status:
Oxin zou, zeker voorlopig, nog totaal geen waarde kunnen halen uit zijn of haar eigen merk en status. Wel is het belangrijk om snel een klantenbestand met een aantal grote namen op te bouwen. Deze referenties kunnen helpen om andere bedrijven over te halen om ook gebruik te gaan maken van het product.

Het product zal niet worden gelanceerd onder de naam Oxin, voor de naam van het product zal er eerst onderzoek gedaan moeten worden naar de beschikbaarheid van de domeinen en zoekvolumes op bepaalde keywords.

Prijs:
De diverse huidige aanbieders vragen behoorlijke bedragen voor hun product. Door hetzelfde en zelfs een beter product aan te bieden voor een lagere prijs zou Oxin erg onderscheidend zijn in de markt. Of de lagere prijs haalbaar is zal in het verloop van dit business model duidelijk worden.

Door het product gratis te mogen proberen kan de klant kennis maken met het product. Iets gratis aanbieden is altijd al een aantrekkelijke waardepropositie geweest. Maar het doel van om het product tijdelijk gratis aan te

bieden is kennis laten maken met het product, Oxin zou de klanten overtuigen van het product wat hopelijk vervolgens een betaald abonnement door zet.

Kostenbeperking:
Een van de belangrijkste waardeproposities is misschien wel kostenbesparing. Door gebruik te maken van het product 'sentiment als marketing tool' kunnen bedrijven de dure en tijdrovende enquêtes voor het bepalen van sentiment de deur uit doen.

Gemak en bruikbaarheid:
Door de gebruiksvriendelijke interface van de website en webportal zou het product extra waarde creëren. De mogelijkheden van de webportal zijn erg uitgebreid, door middel van illustraties en video's kunnen niet alleen de mogelijkheden getoond worden, ook zou direct duidelijk worden hoe met de webportal kan worden gewerkt. Op iedere positie waar extra vragen mogelijk zullen zijn zal door middel van een pop-up extra informatie gegeven kunnen worden.

5.3 Kanalen

Het product wordt aan klanten geleverd via communicatie en verkoopkanalen. In deze paragraaf komt naar voren via welke kanalen we interesse bij de klant creëren, verkoop afhandelen, producten leveren, support geven en via welke kanalen klanten de mogelijkheid hebben om het product te evalueren.

Interesse creëren:
Tijdens de lancering van het product zal nog geen enkele onderneming vanuit de doelgroep gehoord hebben van het product 'sentiment als marketing tool' dat Oxin wil gaan aanbieden. Hoe kan Oxin toch interesse creëren voor het product?

De website zou volledig in staat moeten zijn om het product te promoten, interesse te creëren, vragen te beantwoorden en tot slot het afsluiten van abonnementen. Bezoekers zullen echter niet automatisch op de website komen. Om de bezoekers toch te krijgen zou de website volledig geoptimaliseerd worden voor de zoekmachines. Om in de organische zoekresultaten van de zoekmachine hoog te scoren zal de website minstens één jaar online moeten zijn. Om dit gat op te vangen zullen er gerichte campagnes worden gestart met Google AdWords[41].

Er zullen nog steeds tal van bedrijven zijn die alleen door middel van de website nooit met het product in aanraking zullen komen. Omdat de doelgroep van het product redelijk klein is zal er via de telefoon direct verkoop plaatsvinden waarmee bedrijven geheel gratis en vrijblijvend op het aanbod in kunnen gaan om gebruik te maken van de proefperiode. Het gaat hierbij in eerste instantie alleen over de grote bedrijven waar veel in de engelse taal over wordt geschreven.

[41] AdWords is een belangrijk onderdeel van Google. Het zijn advertenties gebaseerd op zoekwoorden die zijn gedefinieerd door de adverteerder. Als er op één van deze zoekwoorden wordt gezocht, wordt de advertentie naast of boven de zoekresultaten weergegeven.

Aankoop:
Het aankoop proces vindt voor zowel de gratis als betaalde variant plaats via de website. Gebruikers kunnen de keuze maken uit verschillende abonnementsvormen, indien er vragen zijn of begeleiding nodig is tijdens het aankoopproces kan er uiteraard altijd contact op worden genomen.

Aflevering:
Het sentiment dat is bepaald door middel van geanalyseerde teksten wordt aangeboden via een beveiligde verbinding op de webportal. Dit geld voor zowel de betaalde als voor de niet betalende klanten die gebruik maken van de gratis proef periode. Betalende klanten kunnen ook via de webportal aangeven of ze geïnteresseerd zijn om dagelijkse rapporten via de e-mail te ontvangen.

After sales:
Na verkoop van het product 'sentiment als marketing tool' krijgen abonnementhouders toegang tot een servicenummer waarmee ze eenvoudig en snel contact op kunnen nemen. Dit servicenummer is op werkdagen van 10.00 tot 18.00 uur (Nederlandse tijd) beschikbaar.

Evaluatie:
Omdat het product nog verder door ontwikkeld kan worden staan we open voor kritiek Oxin bereid zijn om aanpassingen aan het systeem te maken als dit ten goede komt voor het product. Alle abonnementhouders gaan gevraagd worden om een evaluatie van het product te geven, deze referenties zullen op de website worden weergegeven.

5.4 Klantrelaties

In deze paragraaf worden de verschillende klantrelaties beschreven die Oxin met haar klanten zou hebben en onderhouden.

Klant acquisitie en het stimuleren van verkoop zal plaatsvinden door middel van telefonisch verkeer, voornamelijk de grote bedrijven zullen via de telefoon benaderd worden om gratis met het product kennis te maken. Indien er vragen zullen zijn is het in uitzonderlijke gevallen mogelijk om persoonlijke hulp op locatie te krijgen. Dit is alleen mogelijk voor bedrijven waarvan verwacht wordt dat ze het product mede dankzij de persoonlijke hulp op locatie voor langere tijd zullen aanschaffen. Voor beide klantsegmenten zal verkoop ook automatisch (d.m.v. selfservice) via de website kunnen verlopen.

Door het product naar alle tevredenheid en volgens alle afspraken te leveren wordt verwacht dat de bestaande klantrelaties zullen worden behouden, hiervoor zou het product wel door ontwikkeld moeten worden zodat ze niet overstappen naar een concurrent.

5.5 Inkomstenstromen

De enige inkomsten zullen voortvloeien uit betalende klanten. Deze klanten betalen voor een abonnementsvorm plus eventueel bijkomende opties zoals extra keywords en extra berichten die verwerkt zullen worden. In de volgende tabel zijn de drie aangeboden abonnementsvormen opgenomen. Men kan kiezen uit een free, basic of advanced pakket. De gratis variant zal na één maand automatisch worden omgezet naar het basis pakket.

	Free	Basic	Advanced
Aantal keywords:	*1*	*1*	*10*
Aantal berichten:	*10.000*	*10.000*	*100.000*
Rapportage per email:	*Wekelijks*	*Dagelijks*	*Dagelijks*
Tijdsduur:	*Max 1 maand, automatisch verlengd naar Basic*	*Per maand automatisch verlengd*	*Per maand automatisch verlengd*
Kosten per maand:	*0 euro*	*450 euro*	*2500 euro*
Extra opties:			
1 extra keyword:	-	*150 euro*	*150 euro*
100.000 extra berichten:	-	*1250 euro*	*1250 euro*
1 uur maatwerk:	*80 euro*	80 euro	*80 euro*

Zodra gebruikers buiten de limieten van het pakket treden zullen automatisch de extra kosten in rekening worden gebracht.

5.6 Key resources

De key resources die in deze paragraaf worden beschreven maken het mogelijk om het product ‘sentiment als marketing tool’ te creëren, aan te kunnen bieden, te verkopen en inkomsten mee te verdienen. Het gaat hierbij voornamelijk om fysieke- en human resources die voor dit product benodigd zijn.

Fysieke resources:

Om het product te kunnen leveren zijn we afhankelijk van een aantal fysieke resources. Één van de belangrijkste fysieke resource is de webportal (moet nog ontwikkeld worden), zonder de webportal kunnen klanten nooit van het product gebruik maken. Alle ontwikkelde systemen (waaronder de webportal) zullen komen te draaien op een Virtual Private Server (VPS). Naast de server en webportal zijn we ook afhankelijk van onze bronnen waar we de berichten van binnen halen, denk hierbij aan Twitter en LexisNexis. In de toekomst zullen hier ook andere sociale media aan toegevoegd worden. Tot slot is de Open Amplify API een belangrijke fysieke resource, hiermee zullen alle opgeslagen berichten verwerkt worden.

Als één van de bovenstaande fysieke resources niet meer beschikbaar zou zijn kan het product niet meer aangeboden worden.

Human resources:

Na de voorbereidingen die tijdens dit project zijn getroffen zou het project door één programmeur in drie maanden ontwikkeld kunnen worden. Deze programmeur zou niet alleen verantwoordelijk zijn voor de ontwikkeling van de webportal, ook voor de ontwikkeling van de website zou de programmeur verantwoordelijk zijn. Na de lancering van het product zou het ook nog met enige regelmaat door de programmeur verder door ontwikkeld moeten worden. Hiervoor wordt in de eerste periode 10 uur per week begroot.

Naast de ontwikkeling van het systeem is er ook één persoon nodig die fulltime bezig is met het binnenhalen van de nieuwe klanten en het bieden van support aan bestaande klanten. Zodra er meer klanten en dus ook meer financiële mogelijkheden zijn moet hier meer personeel voor aangenomen worden. Vanwege het tijdverschil met Amerika zou deze persoon elke werkdag van 10.00 tot 18.00 uur (Nederlandse tijd) telefonisch beschikbaar zijn.

5.7 Kernactiviteiten

In deze paragraaf worden de kernactiviteiten besproken die nodig zijn om het product aan te kunnen bieden. Deze kernactiviteiten zullen de belangrijkste acties zijn die Oxin moet ondernemen om met succes te opereren.

De website en webportal zou volledig ontwikkeld en getest moeten worden voordat het product gelanceerd kan worden. Na de lancering van het product zouden klanten kennis moeten maken met het product, dit zal in het begin voornamelijk gaan via telefonische verkoop. Zodra een klant geïnteresseerd is en een abonnement wil afsluiten zullen we in overeenstemming de gewenste keywords invoeren in het systeem. Het systeem verzamelt vervolgens automatisch alle berichten waar deze keywords in voorkomen. Opgeslagen berichten worden verwerkt met Open Amplify en de resultaten zullen direct beschikbaar worden gemaakt via de webportal. Dagelijks zullen er automatisch rapporten worden gegenereerd met de statistieken van afgelopen dag, deze rapporten zullen vervolgens automatisch per e-mail worden verstuurd naar de gebruikers.

5.8 Key partners

De key partners die in dit hoofdstok worden beschreven zijn de verschillende partijen die nodig zijn voor het product ‘sentiment als marketing tool’. Zonder deze key partners zou het product niet kunnen worden geleverd. De belangrijkste key partner is Open Amplify. Open Amplify voert één van de belangrijkste kernactiviteiten uit: het dagelijks verwerken van duizenden tekstberichten voor het bepalen van sentiment.

Naast Open Amplify zijn er nog een aantal key partners voor het verzamelen van de teksten. Denk hierbij aan Twitter en LexisNexis. Zodra het product is gelanceerd zou er zo snel mogelijk onderzoek worden gedaan naar andere mogelijkheden, Facebook en een webcrawler staan hoog op de prioriteitenlijst.

Ook zou Oxin opzoek gaan naar een ‘launching customer’. Deze ‘launching customer’ moet een groot bedrijf zijn dat vanaf de lancering van het product bereid is om als actieve referentie te dienen.

5.9 Kostenstructuur

Deze kostenstructuur beschrijft alle kosten die gemaakt moeten worden om het product aan te kunnen bieden aan de klanten. Voor de producten wordt er onderscheid gemaakt in opstart kosten, maandelijkse kosten en kosten die afhankelijk zijn van het aantal gebruikers.

Opstart kosten:

Het systeem dat ontwikkeld moet worden kan voornamelijk door Oxin worden uitgevoerd. De eenmalige kosten die de ontwikkeling met zich mee brengen zijn geschat op 10.000 euro. Deze kosten zullen terugverdiend moeten worden met de verkoop van het product.

Maandelijkse kosten:

De maandelijkse kosten zijn de kosten die iedere maand terug zullen komen. Deze kosten worden in de volgende lijst opgesomd.

- Kosten voor de server: 350 euro;
- Kosten voor de telefoon: 200 euro;
- Kosten voor internet: 50 euro;
- Huur kantoor ruimte: 1000 euro;
- Personeelskosten: 3000 euro;
- Onvoorziene kosten: 500 euro.

Maandelijkse kosten afhankelijk van het aantal klanten:

Afhankelijk van de hoeveelheid klanten zullen er verschillende berichten worden verwerkt door Open Amplify. Er is een globale schatting gemaakt dat een gemiddelde klant ongeveer 40.000 berichten per maand zou verwerken. Omgerekend brengt Open Amplify ongeveer 1 eurocent per verwerkt artikel in rekening. Gemiddeld zou een klant Oxin ongeveer 400 euro kosten om de verzamelde en opgeslagen artikelen te laten verwerken door Open Amplify.

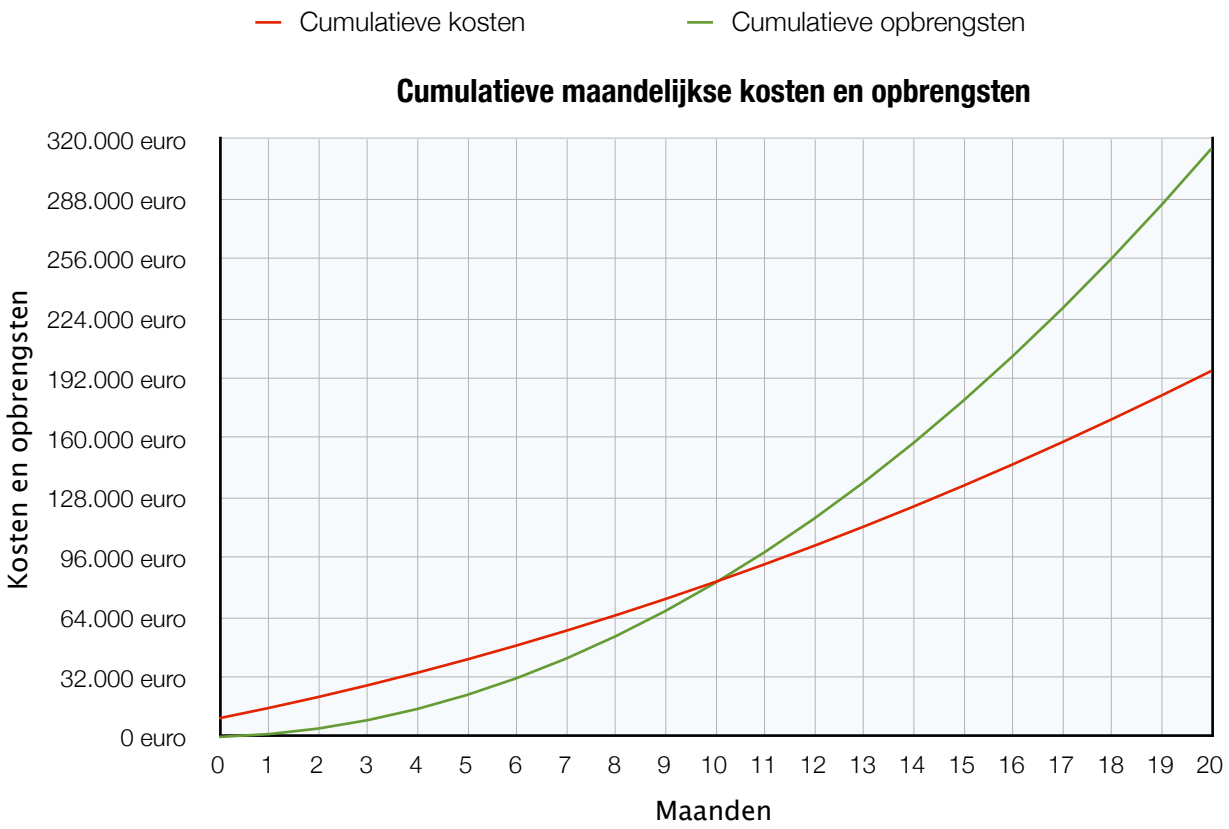
Om een eventuele kosten explosie voor het verwerken van alle artikelen van de gratis accounts te voorkomen zou tijdens de registratie worden gevraagd naar een bedrijfsnaam en bewijs van inschrijving (bij een KvK of soort gelijken). Dit zou een kleine drempel vormen waardoor niet iedereen zo maar in staat is om zich te registreren. Tevens kunnen geregistreerde gegevens gebruikt worden voor verdere sales activiteiten.

Break-even point:

Om het break-even point[42] te kunnen bepalen is er van uit gegaan dat iedere maand gemiddeld één nieuwe klant een betaald abonnementsvorm zal afsluiten. Samen met de veronderstelling dat een gemiddelde klant ongeveer 1500 euro per maand zal opleveren is afbeelding 17 opgebouwd. In deze afbeelding worden de cumulatieve kosten tegenover de cumulatieve opbrengsten weergegeven. Iedere maand zal er gemiddeld één nieuwe betalende klant bijkomen.

Met de bovenstaande aannames zal tot en met maand vier het verlies oplopen tot bijna 20.000 euro. Vanaf maand 5 zal het verlies steeds verder afnemen. Het break-even point zal ongeveer op tien maanden liggen, na deze periode zullen alle kosten afbetaald zijn en kan er winst worden gemaakt.

[42] Het break-even point is het punt waarbij de totale opbrengsten gelijk zijn aan de totale kosten.



Afbeelding 17: break-even point

6. Toekomstperspectief

In dit hoofdstuk wordt een vooruitblik gegeven op de mogelijkheden voor het bepalen van sentiment en de mogelijkheden om dit als product of dienst aan te bieden. Dit toekomstperspectief is een persoonlijke inschatting.

Mogelijkheden met betrekking tot het bepalen van sentiment

In de toekomst zullen steeds meer systemen sentiment bepalen over geschreven tekst. Zo zou bijvoorbeeld van elk geschreven nieuwsartikel automatisch het sentiment kunnen worden weergegeven. Vanwege de globalisering en technologische vooruitgang zal er steeds meer tekst digitaal beschikbaar zijn. Systemen zullen steeds meer met elkaar in verbinding komen te staan en sentiment zal hierbij een steeds grotere rol krijgen. De systemen voor het bepalen van sentiment zouden steeds nauwkeuriger en uitgebreider worden.

Sentiment zal ook een steeds belangrijkere rol krijgen in de profielvorming van personen. Zo zou Google bijvoorbeeld niet alleen advertenties kunnen tonen op basis van interesses maar ook op basis van het sentiment van het gevormde profiel van een bepaald persoon.

Enquêtes over producten of diensten zullen in de toekomst wellicht overbodig worden. Via websites als Twitter en Facebook kan perfect worden gemeten hoe het product wordt ontvangen bij de doelgroep. Twitter komt wellicht in de toekomst zelf met een betaalde variant voor ‘sentiment als marketing tool’, Twitter is nog op zoek naar een verdien model.

Maar ook het onderzoek naar het voorspellen van de beurs zal steeds verder gaan. Meerdere partijen hebben het succesverhaal van Johan Bollen gelezen en zijn zelf nu onderzoek aan het doen naar de mogelijkheden om op basis van sentiment over geschreven tekst de beurs te kunnen voorspellen. Dankzij deze modellen zal de markt steeds perfecter worden, hierdoor zullen de rendementen die nu behaald kunnen worden op basis van sentiment in de toekomst wellicht verdwijnen.

In de toekomst zou het wellicht ook mogelijk zijn om sentiment te bepalen over afbeeldingen en video’s. Van een nieuwszender zoals CNN kan realtime het sentiment worden bepaald. Deze sentiment analyses kunnen bijvoorbeeld weer worden gebruikt voor de producten ‘sentiment als marketing tool’ en ‘sentiment als voorspellende kracht’.

Mogelijkheden voor Oxin

Oxin gaat zeker proberen om het product ‘sentiment als marketing tool’ zo snel mogelijk te ontwikkelen en te lanceren. De ontwikkeling kan grotendeels door Oxin zelf worden uitgevoerd waardoor de kosten voor de lancering van het product relatief laag zullen blijven. Momenteel is het product nog redelijk nieuw en innovatief en hier kan Oxin mogelijk goed van profiteren. Ook zou Oxin zo snel mogelijk onderzoek doen naar de mogelijkheden om van andere bronnen de teksten op te slaan en het sentiment te bepalen. Facebook en een webcrawler staan hoog op de lijst.

7. Conclusies

Sentiment is een gevoel (emotie) dat betrekking heeft op iets of iemand anders. Met andere woorden: sentiment is de algemene gemoedstoestand over een bepaald onderwerp. Voor vele personen, bedrijven en instanties is sentiment een belangrijke indicator. Op basis van het sentiment worden grote beslissingen genomen. Wanneer er bijvoorbeeld een negatief sentiment heerst over een bepaald product kan er voor worden gekozen om nieuwe campagnes te starten om dit negatieve imago te verbeteren. Sentiment wordt ook gebruikt om bijvoorbeeld de peilingen van politieke partijen te meten, om boekverkopen of filmopbrengsten te voorspellen maar ook zijn er onderzoeken uitgevoerd om aan de hand van sentiment de beurskoers te voorspellen.

Sentiment kan op verscheidene manieren bepaald worden. De meest gangbare manier om sentiment te bepalen is aan de hand van enquêtes. Deze enquêtes zijn duur en tijdrovend. Vanwege de globalisering en de technologische vooruitgang is het tegenwoordig ook mogelijk om op een geautomatiseerde wijze sentiment te bepalen met behulp van computersystemen. Deze computersystemen zijn dusdanig ingericht dat ze teksten taalkundig kunnen ontleden. Ze zijn in staat om de onderwerpen en werkwoorden te identificeren. Maar het belangrijkste is dat ze ook in staat zijn om een tekst (of een gedeelte van een tekst) positief of negatief te classificeren. Door honderden berichten over een bepaald onderwerp door deze computersystemen te laten analyseren kan het sentiment over dit onderwerp vastgesteld worden.

Voor dit onderzoek zijn er drie methoden voor het bepalen van sentiment getest en vergeleken. Elke methode heeft zijn eigen benadering, werkwijze en mogelijkheden. De drie onderzochte methoden zijn: de Bayesian methode, OpinionFinder en Open Amplify. Open Amplify scoort van de drie geteste methoden op alle vergelijkingen het beste. Open Amplify heeft veruit de meeste mogelijkheden, het identificeert als beste de onderwerpen en acties, kan door de hoge volatiliteit het beste veranderingen in sentiment waarnemen en is met de hoogste nauwkeurigheid als beste in staat om sentiment te bepalen.

Op basis van sentiment, dat is bepaald door Open Amplify zijn er twee mogelijke producten onderzocht die Oxin zou kunnen aanbieden om een breder aanbod te krijgen van producten en services.

Voor het eerste onderzochte product is op basis van sentiment de beurs voorspeld. In de periode van januari 2001 tot en met december 2010 heeft de markt gemiddeld 20 procent rendement weten te behalen. De theorie die is gebaseerd op basis van sentiment heeft in dezelfde periode 135% rendement behaald. Deze theorie zou aan een geïnteresseerde partij verkocht kunnen worden of de maandelijkse resultaten zullen in een abonnementsvorm aangeboden kunnen worden.

Voor het tweede product is er gekeken of sentiment kan worden aangeboden als marketing tool. Oxin verzamelt alle nieuws- en Twitterberichten over een bepaald onderwerp. Alle verzamelde berichten worden verwerkt met Open Amplify en de resultaten worden via een webportal toegankelijk gemaakt. De webportal zal voornamelijk gericht zijn op het sentiment en het verschil van sentiment over een bepaalde periode. Er kan eenvoudig een vergelijking worden gemaakt tussen diverse vormen van sentiment. De resultaten kunnen opgeslagen of

geëxporteerd worden. Om toegang te krijgen tot de webportal kunnen bedrijven kiezen uit drie verschillende abonnementsvormen.

Hoewel de resultaten van het eerste product verrassend goed zijn is er toch voor gekozen om het tweede product ‘sentiment als marketing tool’ uit te werken in een business model. Oxin is tenslotte geen financiële instelling maar een full service internetbureau dat veel meer affiniteit heeft met marketing dan met finance. Uit het business model is naar voren gekomen dat dit product zeker door Oxin aangeboden zou kunnen worden. Het product is realistisch, heeft een financiële levensvatbaarheid en de investering van rond de 20.000 euro zal na ongeveer tien maanden zijn terug verdiend.

8. Bronnen

- Alexander Osterwalder & Yves Pigneur (2009). Business Model Generatie.

- Bo Pang & Lillian Lee (2005). Movie review data. Verkregen op 20 juni 2011, via http://www.cs.cornell.edu/people/pabo/movie-review-data/

- George Lowery (2011). Tweets study: People across the globe report similar, ever-changing moods. Verkregen op 30 september 2011, via Cornell University.

- Google translate API v2. (2011). Verkregen op 29 juli 2011, via http://code.google.com/intl/nl/apis/language/translate/overview.html

- Ian Barber (21 januari 2010). Bayesian opinion mining. Verkregen op 20 juni 2011, via http://phpir.com/bayesian-opinion-mining

- Johan Bollen, Huina Mao en Xiao-Jun Zeng (2010). Twitter mood predicts the stock market.

- Nick O'Neill (15 maart 2010). Twitter Charging more companies for access to Firehose. Verkregen op 26 juli 2011, via http://socialtimes.com/twitter-charging-firehose_b3905

- Open Amplify (2011), Open Amplify version 2.1 API documentation. Verkregen op 15 augustus 2011 via www.openamplify.com/node/46.

- Open Amplify (2011), Customers. Verkregen op 22 juli 2011 via www.openamplify.com/customers.

- OpinionFinder (2005), Documentation for OpinionFinder 1.5 (opinionfinderv1.5.readme).

- OpinionFinder (2005), SourceFinder.readme, dit bestand wordt meegeleverd met OpinionFinder.

- OpinionFinder (2005), Speech_DirSubj.readme, dit bestand wordt meegeleverd met OpinionFinder.

- OpinionFinder (2005), Subjectivity.readme, dit bestand wordt meegeleverd met OpinionFinder.

- OpinionFinder (2005), Subjectivity.readme, dit bestand wordt meegeleverd met OpinionFinder.

- OpinionFinder (2005), Polarity.readme, dit bestand wordt meegeleverd met OpinionFinder.

- Paul Hoffmann (12 december 2006). OpinionFinder: A Developer's Manual (PDF)

- Pricing and Terms of Service (z.d.). Verkregen op 5 augustus 2011, via http://code.google.com/apis/language/translate/v2/pricing.html

- S.C.E. Dekker, BSc (2011). Is headline risk priced in? (PDF)

- Twitter API documentation (2011). Verkregen op 27 juni 2011, via http://dev.twitter.com/doc

Bijlage 1: OpinionFinder

Deze bijlage is ter ondersteuning van hoofdstuk 3.3.2. In deze bijlage wordt een voorbeeld gegeven van een tekst die door OpinionFinder is verwerkt. Daarnaast worden er verschillende voorbeelden gegeven van de output van OpinionFinder. Voor deze voorbeelden zijn er in totaal 10.850.000 willekeurige Twitterberichten verwerkt met OpinionFinder.

1.1 Voorbeeld van een verwerkte tekst door OpinionFinder

Voorbeeld input OpinionFinder:

Om een voorbeeld van de output van OpinionFinder te kunnen geven is de onderstaande tekst gebruikt.[43]

> Nowhere is the human being more truly revealed than in his letters. Notin literary letters--prepared with care, and the thought of possible publication--but in those letters wrought out of the press of circumstances, and with no idea of print in mind. A collection of such documents, written by one whose life has become of interest to mankind at large, has a value quite aside from literature, in that it reflects in some degree at least the soul of the writer.
>
> The letters of Mark Twain are peculiarly of the revealing sort. He was a man of few restraints and of no affectations. In his correspondence, as in his talk, he spoke what was in his mind, untrammeled by literary conventions.
>
> Necessarily such a collection does not constitute a detailed life story, but is supplementary to it. An extended biography of Mark Twain has already been published. His letters are here gathered for those who wish to pursue the subject somewhat more exhaustively from the strictly personal side. Selections from this correspondence were used in the biography mentioned. Most of these are here reprinted in the belief that an owner of the "Letters" will wish the collection to be reasonably complete.

[43] Deze tekst is verkregen op 5 juli 2011, via: http://www.marktwainbooks.org/complete-letters-mark-twain.

Voorbeeld output OpinionFinder:

Zodra de bovenstaande tekst met OpinionFinder is verwerkt zal de output er als volgt uitzien. Deze tekst in SGML kan vervolgens worden ontleed.

<MPQASENT autoclass1="unknown" autoclass2="obj" diff="4.1">Nowhere is <MPQASRC>the human</MPQASRC> being more truly revealed than in his <MPQASD>letters</MPQASD>.</MPQASENT>

<MPQASENT autoclass1="unknown" autoclass2="subj" diff="2.9">Notin literary letters--prepared with care, and the thought of possible publication--but in those letters wrought out of the press of circumstances, and with no <MPQASD>idea</MPQASD> of print in mind.</MPQASENT> <MPQASENT autoclass1="unknown" autoclass2="subj" diff="6.7">A collection of such documents, written by one whose life has become of interest to mankind at large, has a value quite aside from literature, in that it reflects in some degree at least the soul of the writer.</MPQASENT>

<MPQASENT autoclass1="unknown" autoclass2="obj" diff="0.4">The letters of Mark Twain are <MPQAPOL autoclass="negative">peculiarly</MPQAPOL> of the revealing sort.</MPQASENT> <MPQASENT autoclass1="obj" autoclass2="obj" diff="23.2">He was a man of few restraints and of no affectations. <MPQASENT> <MPQASENT autoclass1="obj" autoclass2="obj" diff="23.0">In his correspondence, as in his <MPQASD>talk</MPQASD>, <MPQASRC>he</MPQASRC> <MPQASD>spoke</MPQASD> what <MPQASD>was in</MPQASD> his <MPQASD>mind</MPQASD>, untrammeled by literary conventions.</MPQASENT>

<MPQASENT autoclass1="obj" autoclass2="obj" diff="20.2">Necessarily such a collection does not constitute a detailed life story, but is supplementary to it.</MPQASENT> <MPQASENT autoclass1="obj" autoclass2="obj" diff="20.0">An extended biography of Mark Twain has already been published.</MPQASENT> <MPQASENT autoclass1="unknown" autoclass2="subj" diff="2.2">His <MPQASD>letters</MPQASD> are here gathered for those who <MPQASD><MPQAPOL autoclass="positive">wish</MPQAPOL></MPQASD> to pursue the subject somewhat more exhaustively from the strictly personal side.</MPQASENT> <MPQASENT autoclass1="unknown" autoclass2="obj" diff="4.4">Selections from this correspondence were used in the biography mentioned.</MPQASENT> <MPQASENT autoclass1="subj" autoclass2="subj" diff="16.4">Most of these are here reprinted in the belief that <MPQASRC>an owner of the "Letters</MPQASRC>" will <MPQASD><MPQAPOL autoclass="positive">wish</MPQAPOL></MPQASD> the collection to be reasonably complete.</MPQASENT>

1.2 Output voorbeelden

Onderwerpen

Met 10.850.000 willekeurig verwerkte Twitterberichten heeft 'SourceFinder' 1.473.840 onderwerpen geïdentificeerd (waarvan 89.739 verschillende). Dit is de top 30 van de meest voorkomende onderwerpen die door 'SourceFinder' zijn geïdentificeerd:

Meest voorkomend (1 t/m 10)		Meest voorkomend (11 t/m 20)		Meest voorkomend (21 t/m 30)	
Onderwerp	Aantal keer in tweet	Onderwerp	Aantal keer in tweet	Onderwerp	Aantal keer in tweet
1. I	*592825 x*	11. Anyone	*8318 x*	21. God	*1883 x*
2. You	*302544 x*	12. Everyone	*7168 x*	22. Everybody	*1780 x*
3. Me	*52596 x*	13. RT	*4914 x*	23. D	*1725 x*
4. They	*52082 x*	14. Us	*4306 x*	24. CNN	*1614 x*
5. He	*48667 x*	15. Them	*3795 x*	25. My mom	*1551 x*
6. We	*44073 x*	16. U	*3356 x*	26. I'm	*1499 x*
7. She	*36166 x*	17. Him	*2459 x*	27. A girl	*1402 x*
8. It	*30975 x*	18. Guys	*2441 x*	28. The world	*1387 x*
9. People	*14679 x*	19. Somebody	*2433 x*	29. The people	*1370 x*
10. Someone	*10349 x*	20. Justin	*2134 x*	30. Don't	*1303 x*

Zoals in de bovenstaande tabel is te zien zijn de top 30 van de meest voorkomende onderwerpen redelijk goed geïdentificeerd. Echter zijn er een aantal voorbeelden niet correct: RT, D, I'm en Don't. De 81,3% nauwkeurigheid die de onderzoekers claimen te behalen komt redelijk overeen met deze test.

DSESE'S

De meest voorkomende subjectieve uitdrukkingen & de gesproken & schriftelijke gebeurtenissen (DSESE) die OpinionFinder met behulp van de 'DSESE Classifier' identificeert staan hieronder in de tabel weergegeven. De 'DSESE Classifier' heeft in totaal 4.683.601 DSESE's geïdentificeerd, waarvan 41.330 verschillende.

Meest voorkomend (1 t/m 10)		Meest voorkomend (11 t/m 20)		Meest voorkomend (21 t/m 30)	
DSESE's	Aantal keer in tweet	DSESE's	Aantal keer in tweet	DSESE's	Aantal keer in tweet
1. Love	*386792 x*	11. Hope	*70835 x*	21. Happy	*36896 x*
2. Know	*318977 x*	12. Thought	*66702 x*	22. Feeling	*36760 x*
3. Want	*274986 x*	13. Believe	*57163 x*	23. Wants	*35978 x*
4. Think	*253867 x*	14. Talking	*50210 x*	24. Wanted	*35687 x*
5. Say	*177117 x*	15. Told	*49912 x*	25. Talk	*34836 x*
6. Please	*121095 x*	16. Ask	*45993 x*	26. Tell	*34476 x*
7. Said	*119283 x*	17. Says	*45523 x*	27. Mind	*31866 x*
8. Feel	*98137 x*	18. Guess	*45336 x*	28. Smile	*31674 x*
9. Hate	*92535 x*	19. Saying	*42266 x*	29. Called	*27469 x*
10. Wish	*79610 x*	20. Thinking	*40485 x*	30. Knew	*22598 x*

Sentiment positief / negatief

De 'Polarity Classifier' heeft 1.643.063 woorden als positief geclassificeerd en 2.088.481 woorden als negatief. In totaal zijn het 2.185 verschillende positieve en 6.687 verschillende negatieve woorden. Opvallend is dat er bij deze test ruim 3 keer zoveel verschillende woorden negatief als positief zijn geclassificeerd. Een aantal mogelijke verklaringen hiervan zijn: er zijn gewoonweg meer verschillende negatieve dan verschillende positieve woorden, OpinionFinder herkent meer verschillende negatieve dan positieve woorden of er wordt gewoonweg over het algemeen negatiever geschreven op Twitter.

Hieronder worden de top 30 van de meest voorkomende positief geclassificeerde en de meest voorkomende negatief geclassificeerde woorden die de 'Polarity Classifier' heeft geclassificeerd weergegeven:

Nr.	Positief geclassificeerd woord	Aantal keer in Twitterbericht	Negatief geclassificeerd woord	Aantal keer in Twitterbericht
1	Want	255543 x	Feel	95913 x
2	Good	226364 x	Bad	86669 x
3	Hope	115732 x	Hate	82126 x
4	Better	94481 x	Too	52985 x
5	Happy	90734 x	Wrong	51554 x
6	Wish	80447 x	Hell	45045 x
7	Please	60947 x	Sorry	39747 x
8	Like	58742 x	Crazt	34641 x
9	Just	56173 x	Stupid	33505 x
10	Wanted	35378 x	Want	33350 x
11	Wants	34220 x	Sure	33152 x
12	Very	25926 x	Sad	32580 x
13	Proud	24262 x	Damn	31258 x
14	Beautiful	24199 x	Good	28511 x
15	Support	24069 x	So	21693 x
16	Love	21900 x	Weird	18513 x
17	Best	15615 x	Very	16777 x
18	Swear	14289 x	Wort	16717 x
19	Feel	13981 x	Really	15855 x
20	Glad	13578 x	Ugly	15551 x
21	Luck	11669 x	Awkward	15176 x
22	Hoping	11466 x	Worst	14051 x
23	Interested	10301 x	Cry	13975 x
24	Would	9689 x	Scared	13786 x
25	Most	9353 x	Pain	13612 x
26	Respect	8944 x	Forget	13083 x
27	Welcome	8826 x	Laughing	12144 x
28	Honest	6870 x	Boring	12139 x
29	Much	6767 x	Annoying	12004 x
30	Positive	6104 x	Tired	11993 x

Bijlage 2: Open Amplify

Deze bijlage is ter ondersteuning van hoofdstuk 3.3.3. In deze bijlage wordt allereerst een voorbeeldtekst weergegeven. Deze voorbeeldtekst wordt vervolgens gebruikt om enkele voorbeelden van de structuur van Open Amplify te geven. Tot slot worden er in deze bijlage verschillende voorbeelden gegeven van de output van Open Amplify.

2.1 Voorbeeldtekst

Onderstaande tekst[44] is gebruikt in de voorbeelden die worden gegeven in deze bijlage.

Teen accused in turkey toss now charged with throwing apples at car on Long

A teenager who was indicted last year in connection with a turkey tossing incident was arraigned Monday on charges that he and others threw apples at a car last October.

Steven Manzolina, 17, was charged with third-degree criminal mischief for allegedly throwing apples at a 1988 Lincoln Navigator on Oct. 17, causing $435.77 worth of damage to its side mirror, a criminal complaint said. Manzolina was accused of acting with others, but no one else was named or charged in the incident. No injuries were reported in the criminal complaint.

Manzolina pleaded not guilty at his arraignment Monday afternoon before Judge John Toomey, the Suffolk County district attorney's office said. Bail was set at $7,500, and Manzolina was scheduled to return to court Friday. A message left Monday for Manzolina's attorney, Patrick O'Connell, was not immediately returned.

Manzolina was accused of stealing a credit card that was used to buy a 20-pound turkey that was thrown from the window of a car into the windshield of an oncoming vehicle on Nov. 12, critically injuring a 44-year-old woman. Five other teenagers also were indicted on various charges relating to the turkey incident.

The 18-year-old who allegedly threw the turkey faces the most serious charges: assault, reckless endangerment and criminal mischief. He, Manzolina and a third teenager who also was charged with stealing the credit card are expected back in court Feb. 10 for a pretrial conference. The three other

[44] The Associated Press State & Local Wire (2005, 31 januari). Teen accused in turkey toss now charged with throwing apples at car on Long

2.2 Structuur output

Onderwerpen

Hieronder wordt de top vijf van onderwerpen die Open Amplify heeft geïdentificeerd tijdens het verwerken van de bovenstaande voorbeeld tekst weergegeven. Achter ieder onderwerp staat een cijfer. Hoe hoger dit cijfer hoe vaker dit onderwerp in de tekst voor komt.

- Steven Manzolina: 38.00
  - Polarity: *Negative ,-0.09*
  - Offering guidance: *To Some Extent ,2.00*
  - Requesting guidance: *Not At All ,1.00*
  - NER: *Person, male,*
- Teenager: 18.00
  - Polarity: *Negative ,-0.86*
  - Offering guidance: *Not At All ,1.00*
  - Requesting guidance: *Not At All ,1.00*
  - NER:
- Turkey: 8.00
  - Polarity: *Neutral ,0.00*
  - Offering guidance: *Not At All ,1.00*
  - Requesting guidance: *Not At All ,1.00*
  - NER:
- Charge: 8.00
  - Polarity: *Negative ,-0.50*
  - Offering guidance: *Not At All ,1.00*
  - Requesting guidance: *Not At All ,1.00*
  - NER:
- Office: 6.00
  - Polarity: *Neutral ,0.00*
  - Offering guidance: *Not At All ,1.00*
  - Requesting guidance: *Not At All ,1.00*
  - NER:

Van ieder geïdentificeerd onderwerp geeft Open Amplify aan wat de polariteit van het onderwerp is, of de onderwerpen gevraagde of aangeboden begeleiding hebben. En in sommige gevallen wordt er ook begeleidende informatie gegeven over het onderwerp.

Eigennamen

Hieronder wordt de top vijf van eigennamen die Open Amplify heeft geïdentificeerd door de voorbeeldtekst te verwerken weergegeven. Van iedere geïdentificeerde eigennaam geeft Open Amplify dezelfde waarden terug als bij de onderwerpen.

- Steven Manzolina: 38.00
  - Polarity: *Negative ,-0.09*
  - Offering guidance: *To Some Extent ,2.00*
  - Requesting guidance: *Not At All ,1.00*
  - NER: *Person, male,*
- Patrick O'Connell: 6.00
  - Polarity: *Neutral ,0.00*
  - Offering guidance: *To Some Extent ,2.00*
  - Requesting guidance: *Not At All ,1.00*
  - NER: *Person, male,*
- Suffolk County: 6.00
  - Polarity: *Neutral ,0.00*
  - Offering guidance: *Not At All ,1.00*
  - Requesting guidance: *Not At All ,1.00*
  - NER: *Location, city,*
- Judge John Toomey: 3.00
  - Polarity: *Positive ,0.74*
  - Offering guidance: *Not At All ,1.00*
  - Requesting guidance: *Not At All ,1.00*
  - NER: *Person, male,*
- Lincoln Navigator: 3.00
  - Polarity: *Neutral ,0.00*
  - Offering guidance: *Not At All ,1.00*
  - Requesting guidance: *Not At All ,1.00*
  - NER: *Person, male,*

Domeinen

Om ter verbeelding te spreken wordt hieronder een voorbeeld gegeven van domeinen die Open Amplify heeft geclassificeerd van de voorbeeldtekst uit deze bijlage:

- Food and drink: 8.00
  - Turkey: 8.00
- Crime: 5.00
  - Law enforcement: 4.00
    - Attorney: 2.00
    - Court: 2.00
  - Violence: 1.00
    - Assault: 1.00
- Sports: 4.00
  - Tennis: 2.00
    - Court: 2.00

- Court: 2.00

Zoals in dit voorbeeld is te zien wordt er voor ieder domein en sub-domein een waarde meegegeven. Hoe hoger deze waarde, hoe zekerder Open Amplify is dat de geanalyseerde tekst daadwerkelijk over dit domein gaat.

Acties

Hieronder staat een voorbeeld van de output van de 'Actions Analysis'. Om het overzichtelijk te houden wordt alleen de top 5 van acties weergegeven van de voorbeeldtekst uit deze bijlage.

- Plead: 5.00
  - Decisiveness: *Low ,1.00*
  - Action type: *Persuade, 1.00*
  - Temporality: *Past, 1.00*
  - Offering guidance: *Not At All ,1.00*
  - Requesting guidance: *Not At All ,1.00*
- Say: 5.00
  - Decisiveness: *Low ,1.00*
  - Action type: *Say, 1.00*
  - Temporality: *Past, 1.00*
  - Offering guidance: *Not At All ,1.00*
  - Requesting guidance: *Not At All ,1.00*
- Return to court: 3.00
  - Decisiveness: *Low ,1.00*
  - Action type: *Move, 1.00*
  - Temporality: *Past, 1.00*
  - Offering guidance: *Not At All ,1.00*
  - Requesting guidance: *Not At All ,1.00*
- Throw apples: 3.00
  - Decisiveness: *Low ,1.00*
  - Action type: *Other, 1.00*
  - Temporality: *Past, 1.00*
  - Offering guidance: *Not At All ,1.00*
  - Requesting guidance: *Not At All ,1.00*
- Arraign: 2.00
  - Decisiveness: *Low ,1.00*
  - Action type: *Other, 1.00*
  - Temporality: *Past, 1.00*
  - Offering guidance: *Not At All ,1.00*
  - Requesting guidance: *Not At All ,1.00*

2.3 Output voorbeelden

Hieronder worden enkele voorbeelden gegeven van de top 30 meest voorkomende onderwerpen, eigennamen, acties en locaties. Hiervoor zijn er 157.354 artikelen verwerkt waar het woord 'apple' of 'google' minimaal één keer in de titel of in de inleidende paragraaf voor komt. Deze artikelen komen uit de periode van januari 2005 t/m juli 2011.

Top 30 meest voorkomende onderwerpen:

Meest voorkomend (1 t/m 10)		Meest voorkomend (11 t/m 20)		Meest voorkomend (21 t/m 30)	
Onderwerp	Aantal keer	Onderwerp	Aantal keer	Onderwerp	Aantal keer
1. Google	*43531 x*	11. Site	*9548 x*	21. Million	*5919 x*
2. Apple	*32975 x*	12. Microsoft	*9485 x*	22. I'm	*5647 x*
3. Company	*30220 x*	13. Share	*9336 x*	23. Phone	*5369 x*
4. $	*16686 x*	14. U.S.	*8919 x*	24. New York	*5297 x*
5. It	*16178 x*	15. Service	*8577 x*	25. Application	*4957 x*
6. Internet	*14341 x*	16. Yahoo	*7758 x*	26. Don't	*4950 x*
7. Web	*13783 x*	17. Analist	*7042 x*	27. Video	*4850 x*
8. iPhone	*13018 x*	18. Stock	*6910 x*	28. Technology	*4724 x*
9. People	*12004 x*	19. Information	*6655 x*	29. Customer	*4656 x*
10. User	*10959 x*	20. iPod	*6333 x*	30. Business	*4597 x*

Google en Apple zijn de meest voorkomende onderwerpen die Open Amplify heeft geïdentificeerd. Dit kan uiteraard heel goed kloppen omdat de verzameling artikelen is samengesteld op basis van deze woorden.

Top 30 meest voorkomende eigennamen:

Meest voorkomend (1 t/m 10)		Meest voorkomend (11 t/m 20)		Meest voorkomend (21 t/m 30)	
Eigennaam	Aantal keer	Eigennaam	Aantal keer	Eigennaam	Aantal keer
1. Google	*38243 x*	11. New York	*4028 x*	21. AAPL	*2227 x*
2. Apple	*17505 x*	12. TV	*3212 x*	22. Mac	*2226 x*
3. Internet	*11567 x*	13. Nasdaq	*3155 x*	23. Apple Valley	*2186 x*
4. iPhone	*11036 x*	14. China	*3143 x*	24. Apple Inc	*2160 x*
5. Web	*10592 x*	15. Comtex SmarTrend	*2904 x*	25. Comtex News	*2072 x*
6. Microsoft	*7162 x*	16. iPad	*2691 x*	26. Steve Jobs	*2060 x*
7. Yahoo	*6467 x*	17. YouTube	*2686 x*	27. Facebook	*2028 x*
8. U.S.	*6457 x*	18. TM	*2684 x*	28. I've	*1955 x*
9. iPod	*5483 x*	19. AT&T	*2557 x*	29. CEO	*1912 x*
10. I'm	*4123 x*	20. Washington	*2251 x*	30. GOOG	*1907 x*

Net als bij de onderwerpen komt ook bij de eigennamen de woorden 'Google' en 'Apple' het meeste naar voren.

Top 30 meest voorkomende domeinen:

Meest voorkomend (1 t/m 10)		Meest voorkomend (11 t/m 20)		Meest voorkomend (21 t/m 30)	
Domein	Aantal	Domein	Aantal	Domein	Aantal
1. Internet	*68632 x*	11. Political organization	*17474 x*	21. Video games	*11256 x*
2. Sports	*47789 x*	12. Stockmarket	*17077 x*	22. Crime	*10340 x*
3. Computers	*47616 x*	13. Arts	*16149 x*	23. Consumer electronics	*10334 x*
4. Entertainment	*47487 x*	14. Food and drink	*15728 x*	24. Immediate family	*9453 x*
5. Politics	*46418 x*	15. Web	*13483 x*	25. Computer programming	*9240 x*
6. Business	*40679 x*	16. Television	*13333 x*	26. Fashion	*9214 x*
7. Education	*24740 x*	17. Computer software	*11992 x*	27. Fishing	*9140 x*
8. Stock	*23026 x*	18. School	*11773 x*	28. Microsoft	*8954 x*
9. Family	*20330 x*	19. Advertising	*11680 x*	29. Law enforcement	*8542 x*
10. Music	*18176 x*	20. Transportation	*11344 x*	30. Baseball	*7655 x*

Hierboven staat een lijst met de top 30 meest voorkomende domeinen die Open Amplify heeft geclassificeerd tijdens het verwerken van de artikelen.

Top 30 meest voorkomende acties:

Meest voorkomend (1 t/m 10)		Meest voorkomend (11 t/m 20)		Meest voorkomend (21 t/m 30)	
Actie	Aantal keer	Actie	Aantal keer	Actie	Aantal keer
1. Say	*640984 x*	11. Believe	*66406 x*	21. Advise	*26846 x*
2. Communicate	*286168 x*	12. Travel	*52087 x*	22. Impel	*26315 x*
3. Move	*228966 x*	13. Sell	*50776 x*	23. Repair	*26032 x*
4. Give	*112058 x*	14. Like	*44259 x*	24. Emote	*23429 x*
5. Create	*107973 x*	15. Request	*39017 x*	25. Persuade	*22548 x*
6. Use	*78012 x*	16. Attack	*35994 x*	26. Choose	*22048 x*
7. Attend	*73173 x*	17. Want	*35301 x*	27. Damage	*18764 x*
8. Help	*71701 x*	18. Assess	*34862 x*	28. Consume	*17680 x*
9. Buy	*69961 x*	19. Learn	*29155 x*	29. Have	*14085 x*
10. Compete	*67184 x*	20. Sport	*27680 x*	30. Be	*13698 x*

Hierboven staat een lijst met de top 30 meest voorkomende acties die Open Amplify heeft geïdentificeerd tijdens het verwerken van de artikelen.

Top 30 meest voorkomende locaties:

Meest voorkomend (1 t/m 10)		Meest voorkomend (11 t/m 20)		Meest voorkomend (21 t/m 30)	
Locatie	Aantal keer	Locatie	Aantal keer	Locatie	Aantal keer
1. U.S.	*68632 x*	11. Silicon Valley	*17474 x*	21. Mountain View, Calif.	*11256 x*
2. New York	*47789 x*	12. Apple Valley	*17077 x*	22. Germany	*10340 x*
3. United States	*47616 x*	13. Europe	*16149 x*	23. Texas	*10334 x*
4. San Francisco	*47487 x*	14. Chicago	*15728 x*	24. Boston	*9453 x*
5. US	*46418 x*	15. USA	*13483 x*	25. Hollywood	*9240 x*
6. Washington	*40679 x*	16. New York City	*13333 x*	26. Japan	*9214 x*
7. China	*24740 x*	17. York	*11992 x*	27. Milwaukee	*9140 x*
8. California	*23026 x*	18. Los Angeles	*11773 x*	28. Calif.	*8954 x*
9. IN	*20330 x*	19. London	*11680 x*	29. Minnesota	*8542 x*
10. America	*18176 x*	20. Mountain View	*11344 x*	30. Canada	*7655 x*

Hierboven staat een lijst met de top 30 meest voorkomende locaties die Open Amplify heeft geïdentificeerd tijdens het verwerken van de artikelen.

Bijlage 3. Technische details

In deze bijlage worden de technische details besproken van de serverinstallatie en -configuratie. Gebruikte programma's, instellingen en geprogrammeerde software dat is gebruikt en ontwikkeld voor dit onderzoek komt aan bod.

Hardware server:
Voor dit onderzoek is er gebruik gemaakt van de volgende server configuratie.

Merk:	Acer;
Type:	Veriton M265;
CPU:	Pentium(R) Dual-Core CPU @ 2.50 GHz;
Geheugen:	3GB intern;
Harde schijf:	160GB intern;
Netwerkadapter:	Realtek 1000M Ethernet.

Operating system:
Vanwege de stabiliteit, functionaliteit en beschikbare programma's is het noodzakelijk geweest om Linux te installeren. Na vele Linux varianten te hebben getest blijkt dat OpinionFinder alleen compatible is met oude Linux varianten. De gebruikte servernaam van de Linux server is: server01.

Gebruikt operating system: Ubuntu 6.06 LTS (the Dapper Drake) - released in juni 2006

Geïnstalleerde pakketten:
In deze paragraaf worden de belangrijkste geïnstalleerde en geconfigureerde pakketten weergegeven. Alle genoemde pakketen zijn noodzakelijk voor de werking van de server.

- Apache 2.0.55;
- PHP 5.1.2.1;
- MySQL 5.0.22;
- Cron 3.0pl1;
- Perl 5.8.7;
- Python 2.4.2;
- Java 1.5;
- SSH 1:4.2p1;
- VNC server 3.3.7;
- JSON 1.1.0-2;
- Curl 5.1.2;
- Gawk 1:3.1.5;

- GCC 1:3.3.6-10.

Opslaan Twitterberichten:

Voor dit onderzoek hebben zijn er een tweetal Twitter streams gebruikt. Om deze twee streams gelijktijdig te kunnen draaien moet er gebruik gemaakt worden van twee verschillende Twitteraccounts die vanaf twee verschillende ip-adressen verbinding maken met Twitter. Twitter staat het niet toe om twee verschillende streams tegelijkertijd te draaien vanaf het zelfde ip-adres of vanaf hetzelfde Twitter account. De twee streams die continu de Twitterberichten binnenhalen zijn:

1. sample.json: Via een PHP bestand dat lokaal op de server (server01) draait wordt er verbinding gemaakt met de Random Twitter stream (sample.json) die gemiddeld 1% van de willekeurige Twitterberichten binnen haalt. De Twitterberichten die via het JSON formaat worden aangeleverd worden via het LanguageDetect script gehaald. Vervolgens worden alleen de Twitterberichten die volgens het LanguageDetect script een engelse waarde van hoger dan 0,2 hebben opgeslagen in de database, dit zijn er ongeveer 100.000 per dag.

2. filter.json: Via een PHP bestand dat op een server in Amsterdam draait wordt er verbinding gemaakt met de filter stream van Twitter (filter.json). Via deze filter stream worden alle Twitterberichten op de lokale server (server01) opgeslagen in een database. Het gaat hierbij niet om de willekeurige Twitterberichten, via filter.json kunnen er keywords opgegeven worden waar op gezocht moet worden. Alle Twitterberichten die die een overeenkomst hebben met één van deze keywords worden binnen gehaald. Vervolgens haalt de lokale server Twaalf keer per door middel van een ander PHP bestand alle Twitterberichten uit de database op die nog niet zijn gecontroleerd op taal en controleert dit vervolgens met het LanguageDetect script. Alle Twitterberichten die niet voldoen aan de gestelde eisen (een engelse waarde van hoger dan 0.35) worden weer verwijderd. De opgegeven keywords voor filter.json zijn: apple, iphone, imac, macbook, ipad, ipod, itunes, osx, IOS, google, gmail, chrome, analytics, adwords, adsense, picassa en android.

Van ieder Twitterbericht dat wordt opgeslagen worden de volgende gegevens bewaard in de database:

- Het Twitterbericht;
- Datum en tijd van het Twitterbericht;
- Het id van de gebruiker;
- Het aantal volgers van de gebruiker;
- Het aantal aantal Twitter accounts die worden gevolgd door de gebruiker.

Opslaan tekstbestanden:

De tekstbestanden die zijn geëxporteerd bij LexisNexis moeten nog worden opgeslagen in een database. In ieder tekstbestand staan meerdere artikelen. Om deze artikelen te kunnen scheiden en om van ieder artikel de titel, de tekst, de bron en de datum op te slaan in een database is een script geschreven in PHP. Dit PHP bestand zal in dit document verder niet worden beschreven omdat het te veel uitzonderingen bevat. Dit komt omdat ieder artikel een ander export-formaat kan hebben; zo staat bijvoorbeeld de titel, de datum en de bron altijd in een ander formaat en op een andere locatie.

Installatie OpinionFinder:

OpinionFinder draait en linux en heeft een behoorlijk groot stappenplan die nauwkeurig moet worden gevolgd voordat OpinionFinder werkt. Voor dit onderzoek is er gebruik gemaakt van OpinionFinder 1.5. OpinionFinder 1.5 kan gedownload worden via: http://www.cs.pitt.edu/mpqa/opinionfinderrelease/.

Voor de de werking van OpinionFinder moet de volgende software geïnstalleerd zijn, en via de terminal toegankelijk zijn bij hun naam:

- Python (versie 2.3 of hoger);
- Java (versie 1.5);
- Perl.

Naast de bovenstaande software dienen ook onderstaande programma's geïnstalleerd te worden alvorens OpinionFinder geïnstalleerd kan worden.

- Sundance: dit programma wordt meegeleverd met OpinionFinder en is te vinden in de map: /opinionfinder/software. In deze map staat tevens ook een readme file die stap voor stap instructies geeft voor de installatie van dit programma;
- Scol 1k: in de map: /opinionfinder/software staat een ZIP van het programma. Na het uitpakken van dit bestand staat er wederom een readme file die stap voor stap instructies geeft voor de installatie van Scol;
- BoosTexter 2.1: dit programma wordt niet meegeleverd met OpinionFinder maar moet apart gedownload worden via de website: http://www.research.att.com/~gsf/download/ref/boostexter/boostexter.html. Insctructies voor de installatie van BoosTexter zijn ook te vinden op deze website;
- WordNet 1.6: dit programma wordt ook niet meegeleverd met OpinionFinder en zal ook apart gedownload moeten worden. Dit programma kan gedownload worden via http://wordnet.princeton.edu/obtain. LET OP: OpinionFinder is ontwikkeld met behulp WordNet 1.6 en kan niet werken met andere versies van WordNet.

Zodra alle voorgaande software en programma's correct zijn geïnstalleerd kan OpinionFinder worden geïnstalleerd. In de OpinionFinder map staat een configuratie bestand genaamd config.txt die correct ingevuld moet worden. Zodra dit bestand correct is ingevuld kan via de terminal het volgende commando aangeroepen worden: 'python install.py config.txt'.

De onderstaande programma's hoeven niet door de gebruiker geïnstalleerd te worden maar worden tijdens de installatie van OpinionFinder geïnstalleerd.

- OpenNLP 1.3.0;
- SourceFinder;
- PyWordNet 1.6.

De volgende foutmeldingen kunnen tijdens de installatie van OpinionFinder worden negeert: 'Note: Some input files use unchecked or unsafe operations' & 'Note: Recompile with -Xlint:unchecked for details'.

OpinionFinder aanroepen:
Zodra de installatie van OpinionFinder succesvol is verlopen kan OpinionFinder via de terminal worden aangeroepen. De eenvoudigste manier om dit te doen is met het volgende commando: 'python opinionfinder.py -f doclist.txt'. Het bestand doclist.txt verwijst naar alle tekstbestanden met teksten die verwerkt moeten worden door OpinionFinder.

Voor meer informatie over hoe OpinionFinder gebruikt kan worden: typ het volgende commando via de terminal: 'python opinionfinder.py'.

Open Amplify:
Opgeslagen Twitterberichten en nieuwsartikelen worden door middel van een PHP bestand verwerkt met Open Amplify. Dit PHP bestand zorgt ervoor dat de tekst wordt opgeknipt in stukken van 5kb en stuurt vervolgend via CURL de desbetreffende tekst naar de Open Amplify API. Binnen een aantal seconden geeft Open Amplify de verwerkte tekst terug in XML formaat. Deze XML wordt uitgelezen met XPATH en vervolgens opgeslagen in de MySQL database. Een gemiddelde tekst van 600 woorden geeft ongeveer 200 rijen terug die moeten worden opgeslagen in de database.

Zodra een verwerkte tekst succesvol is opgeslagen zou het script automatisch een nieuwe tekst selecteren zodat het proces herhaald kan worden.

Bayesian methode:
Om de Bayesian methode te kunnen testen is er een script ontwikkeld dat verbinding maakt met een database en vervolgens een tekst ophaalt. Met de Bayesian methode wordt er gekeken of de opgehaalde tekst positief of negatief is en het resultaat hiervan wordt opgeslagen in de database. Zodra dit proces is afgerond zou het script automatisch worden herstart zodat er een nieuwe tekst wordt ingeladen.

Cronjobs:
Een cronjob is een commando dat een programma of script op een ingesteld tijdstip uitvoert. Op de server (server01) zijn er meerdere cronjobs ingesteld. Deze cronjobs zorgen ervoor dat de Twitterberichten worden binnen gehaald, Twitterberichten worden gecontroleerd op taal, Twitterberichten worden opgeslagen en dat nieuwsartikelen & Twitterberichten worden verwerkt door OpinionFinder, Open Amplify en door de Bayesian methode.

Cronjob 1 *(elke 15 minuten):* deze cronjob zorgt ervoor dat elke 15 minuten een bestand wordt aangeroepen die de 1% willekeurige Twitterberichten binnenhaalt, controleert met het LanguageDetect script en de Twitterberichten opslaat. Elke 15 minuten wordt dit script opnieuw aangeroepen omdat het kan voorkomen dat het script wel eens de verbinding met Twitter verliest.

Cronjob 2 *(elke 15 minuten)*: deze cronjob draait op de externe server en zorgt ervoor dat dit filter stream elke vijftien minuten wordt aangeroepen en dat de binnekomende Twitterberichten worden opgeslagen op de lokale server.

Cronjob 3 *(elke 5 minuten)*: deze cronjob draait op de lokale server (server01) en zorgt ervoor dat de nieuwe opgeslagen Twitterberichten van cronjob 2 met het LanguageDetect script worden gecontroleerd op taal. Zodra

de opgeslagen Twitterberichten niet voldoen aan de gestelde eisen (een engelse waarde van hoger dan 0,35) worden ze verwijderd.

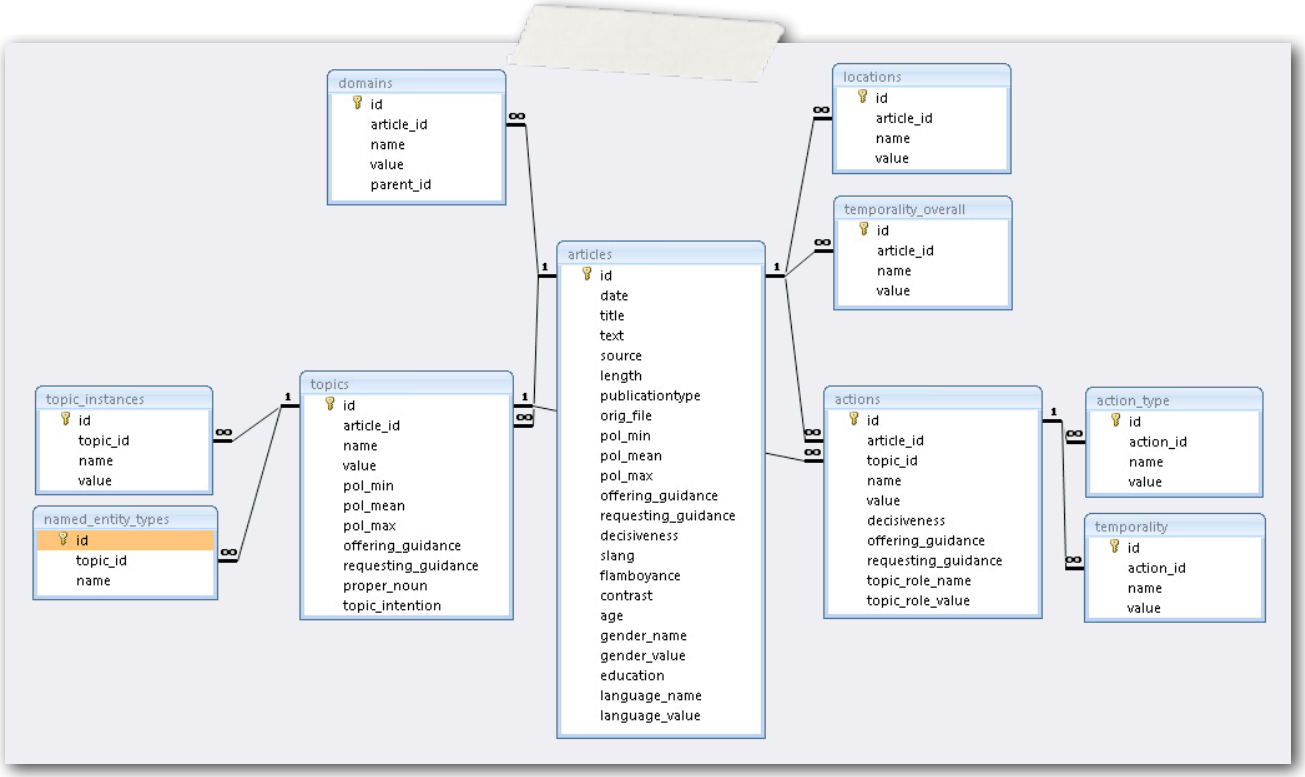
Cronjob 4 *(elke 15 minuten):* deze cronjob draait op de lokale server en wordt elke vijftien minuten aangeroepen. Het zorgt ervoor dat elke 15 minuten wordt gekeken naar nieuw opgeslagen Twitterberichten. Deze Twitterberichten worden op de juiste formaat geëxporteerd naar tekstbestanden zodat deze geschikt zijn voor OpinionFinder. OpionFinder wordt aangeroepen en zal alle geëxporteerde Twitterberichten verwerken. Zodra dit proces is afgerond zal de output van OpinionFinder (SGML markup) worden uitgelezen en opgeslagen in de database. Als laatste zorgt de cronjob ervoor dat alle tijdelijke bestanden van OpinionFinder weer worden verwijderd. Er zit een crontrole op het script zodat OpinionFinder twee maal tegelijkertijd aangeroepen kan worden.

Cronjob 5 *(elke 60 minuten):* deze cronjob zorgt ervoor dat elk uur het script wordt aangeroepen dat alle teksten verwerkt met Open Amplify.

Cronjob 6 *(elke 60 minuten):* deze cronjob zorgt ervoor dat elk uur het script wordt aangeroepen dat alle teksten verwerkt met de Bayesian methode.

Database ontwerp Open Amplify:

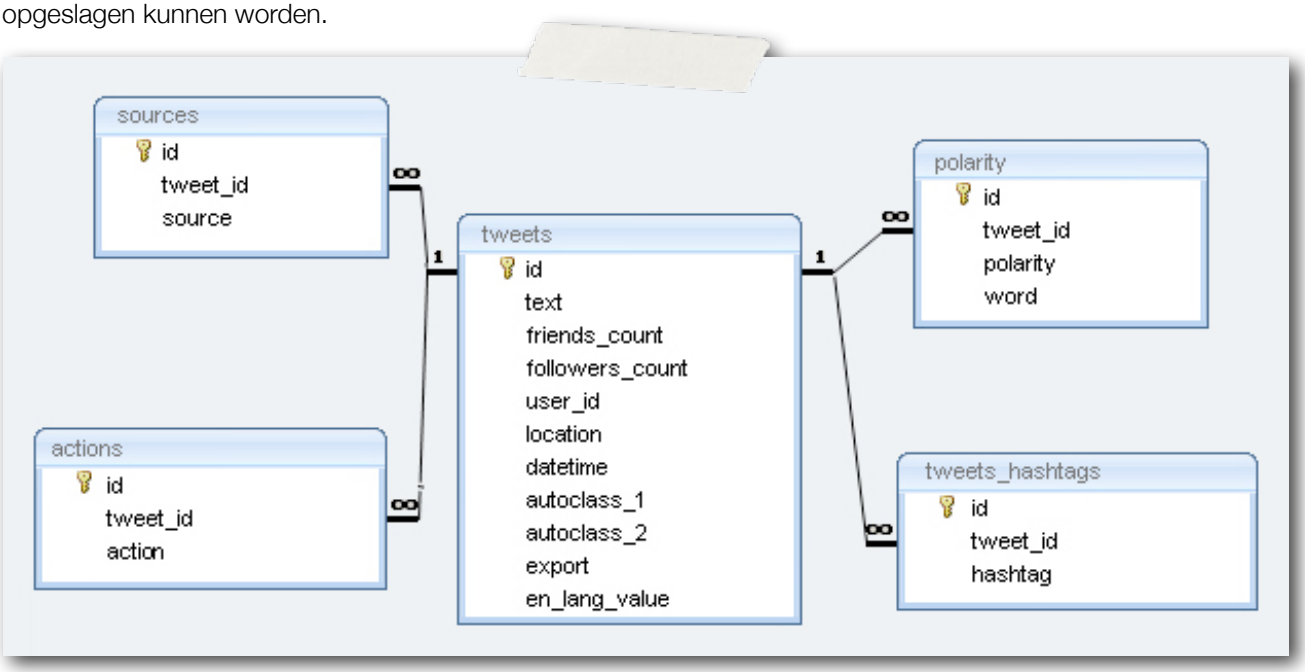
Hieronder staat het ontwerp van de database waar alle nieuwsartikelen en de resultaten van Open Amplify in opgeslagen kunnen worden.



Afbeelding 18: database ontwerp Open Amplify

Database ontwerp OpinionFinder:

Hieronder staat het ontwerp van de database waar alle twitterberichten en de resultaten van OpinionFinder in opgeslagen kunnen worden.



Afbeelding 19: database ontwerp OpinionFinder

Indices:

Voor de optimalisatie van de database zijn indices van groot belang. Een index van een database werkt eigenlijk hetzelfde als een index van een boek: een index wordt gebruikt om er snel achter te komen waar bepaalde gegevens staan.

Door gebruik te maken van verschillende indices kan een query in sommige gevallen ruim 1000 keer zo snel uitgevoerd worden. Onderstaande query kan met een database van meer dan 53 miljoen rijen zonder indices niet binnen een dag worden uitgevoerd, terwijl MySQL met indices de query binnen 10 seconden heeft uitgevoerd.

```
SELECT YEAR(a.date) AS 'year' ,MONTH(a.date) AS 'month',COUNT(t.id) AS 'count topics', COUNT(DISTINCT(a.id)) AS 'count articles',avg(t.pol_mean) AS 'AVG topics_pol_mean' ,avg(a.pol_mean) AS 'AVG articles_pol_mean', SUM(t.pol_mean) AS 'SUM topics_pol_mean', SUM(a.pol_mean) AS 'SUM articles_pol_mean'
FROM `articles` a
INNER JOIN `topics` t ON a.id = t.article_id
WHERE t.name LIKE '%apple%' AND proper_noun=0 AND topic_intention=0
GROUP BY YEAR(a.date),MONTH(a.date)
ORDER BY YEAR(a.date),MONTH(a.date)
```

Matlab:

Het software pakket Matlab is een technische omgeving dat wordt gebruikt voor wiskundige toepassingen zoals het berekenen van functies, bewerken van matrices, statistiek, tekenen van grafieken, schrijven en implementeren van algoritmen en het maken van grafische gebruikersinterfaces. Voor het onderzoek of sentiment een voorspellende kracht op de beurs heeft is er gebruikt gemaakt van Matlab. Vanuit Matlab is er verbinding gemaakt met de database die op de server draait.