Automatische hoogtepuntextractie en -annotatie van voetbalbeelden op basis van Twitter.

Fréderic Godin

Promotoren: prof. dr. ir. Rik Van de Walle, dr. ir. Sofie Van Hoecke
Begeleiders: dr. Steven Verstockt, Pieterjan De Potter

Masterproef ingediend tot het behalen van de academische graad van
Master in de ingenieurswetenschappen: computerwetenschappen

Vakgroep Elektronica en Informatiesystemen
Voorzitter: prof. dr. ir. Jan Van Campenhout
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2011-2012

UNIVERSITEIT
GENT

Automatische hoogtepuntextractie en -annotatie van voetbalbeelden op basis van Twitter.

Fréderic Godin

Promotoren: prof. dr. ir. Rik Van de Walle, dr. ir. Sofie Van Hoecke
Begeleiders: dr. Steven Verstockt, Pieterjan De Potter

Masterproef ingediend tot het behalen van de academische graad van
Master in de ingenieurswetenschappen: computerwetenschappen

Vakgroep Elektronica en Informatiesystemen
Voorzitter: prof. dr. ir. Jan Van Campenhout
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2011-2012

UNIVERSITEIT
GENT

Voorwoord

Het onderwerp van deze masterproef groeide vanuit mijn persoonlijke fascinatie voor Twitter en sportwedstrijden. De concrete invulling ervan ontstond echter tijdens een gesprek met mijn thesisbegeleiders waarbij we onszelf de vraag stelden of we Twitter niet konden gebruiken om belangrijke momenten te selecteren uit videobeelden van sportwedstrijden. Het idee om hoogtepunten te extraheren en annoteren door gebruik te maken van Twitter was geboren.

Graag had ik in dit voorwoord ook enkele mensen willen bedanken. In de eerste plaats wil ik mijn promotoren prof. dr. ir. Rik Van de Walle en dr. ir. Sofie Van Hoecke bedanken om mij de kans te geven mij te verdiepen in de analyse van Twitter en videosequenties. Ook wil ik mijn thesisbegeleiders, dr. Steven Verstockt en Pieterjan De Potter, alsook dr. Davy Van Deursen bedanken voor hun constructieve ondersteuning, tips en feedback bij het tot stand komen van deze masterproef.

Daarnaast wil ik ook mijn familie en vrienden bedanken die mij gesteund hebben tijdens deze intensieve periode. In het bijzonder wil ik mijn vriendin Kim bedanken voor het helpen verzamelen van de Twitterberichten en videosequenties.

Ten slotte, wil ik mijn ouders bedanken om mij de mogelijkheid te geven deze studie aan te vatten en mij gedurende deze 5 jaar altijd te steunen.

Fréderic Godin, juni 2012

Toelating tot bruikleen

Fréderic Godin, juni 2012

Automatische hoogtepuntextractie en -annotatie van voetbalbeelden op basis van Twitter.

door
Fréderic Godin

Afstudeerwerk ingediend tot het behalen van de graad van
Master in de ingenieurswetenschappen: computerwetenschappen

Academiejaar 2011-2012

Universiteit Gent
Faculteit Ingenieurswetenschappen en Architectuur
Vakgroep Elektronica en informatiesystemen
Voorzitter: prof. dr. ir. J. Van Campenhout

Promotoren: prof. dr. ir. R. Van de Walle, dr. ir. S. Van Hoecke
Thesisbegeleiders: dr. S. Verstockt, P. De Potter

Samenvatting

In deze masterproef wordt een systeem voorgesteld dat hoogtepunten kan extraheren uit videosequenties van voetbalwedstrijden in realtime door gebruik te maken van een algoritme dat gebeurtenissen detecteert in Twitter. In het eerste deel van de masterproef wordt een algoritme voorgesteld dat 6 verschillende gebeurtenissen kan detecteren en identificeren in realtime in Twitter: een doelpunt, een strafschop, een fout, een wissel, het einde van de eerste helft en het einde van de tweede helft. Hiervoor wordt gebruik gemaakt van een verzameling regels en een Support Vector Machine (SVM). In het tweede deel van de masterproef wordt een systeem voorgesteld om op basis van logodetecties in een videosequentie en gebeurtenisdetecties in de corresponderende Twitterberichtenstroom, hoogtepunten te extraheren uit de videosequentie en deze te annoteren met relevante informatie. Als toepassing werd er gekozen om de doelpunten te extraheren uit videosequenties van de Engelse Premier League 2011-2012.

Trefwoorden: Twitter, gebeurtenisdetectie, videoanalyse, hoogtepuntextractie

Automated highlight extraction and annotation of soccer video sequences using Twitter

Fréderic Godin

Supervisor(s): P. De Potter, dr. S. Verstockt, dr. ir. S. Van Hoecke, prof. dr. ir. R. Van de Walle

Abstract—**In this paper, a system is presented that can extract and annotate highlights in soccer video sequences in realtime based on the detection of events in Twitter. The first part of this system consists of an algorithm that is able to detect and identify 6 types of events in Twitter: a goal, a penalty, a foul, a substitution, the end of the first half and the end of the second half. To detect those events we make use of a set of rules and a Support Vector Machine to classify tweets. The second part of the system consist out of an algorithm that can extract highlights out of video sequences making use of logo detections in the video sequence and event detections in Twitter. The focus is on goal highlight extraction. Very promising results were obtained for the detection of 6 types of events and the extraction of goals out of video sequences.**

Keywords—**Twitter, sports analysis, event detection, highlight extraction**

I. Introduction

THE last few years, a massive growth of social media took place. All kinds of events around the globe are reported first-hand using social networks like Twitter. This is especially true for sports. In 2011, 6 out of the 16 events with the highest tweet-rate per second on Twitter were sport-related[1].

In this paper, a system is presented that can automatically detect and identify events in soccer games using tweets posted on Twitter. The goal is to automatically generate a game summary in real-time as current national news websites do. On top of that, an algorithm is proposed to extract highlights out of the corresponding video sequence in real-time based on the event detection in Twitter to improve the user experience. The output of the complete system is a video fragment annotated with information extracted from Twitter that can be used for digital archiving. The system is evaluated on soccer games of the English Premier League 2011-2012. Tweets were gathered using hashtags of those teams.

The remainder of this paper is organised as follows. In section II, the related work in the domain of Twitter and sports is discussed. In Section III an algorithm to detect and identify 6 types of events in Twitter is proposed. Next an algorithm to extract the corresponding highlight out of soccer video sequences is proposed in Section IV. In Section V both algorithms are evaluated. This paper ends with a conclusion in Section VI.

II. Related work

Because of the tremendous growth of Twitter in terms of users and tweets, and the potential of Twitter as breaking news broker, Twitter has become a hot research topic but is still premature. In this section the 2 most important articles are briefly discussed.

In [1], the authors present a two-stage algorithm to detect and identify 4 types of events in American Football games in real-time. They make use of an adaptive sliding window to roughly filter out the most important parts and a lexicon of words to identify the events.

Choudhury and Breslin [2] classify sports tweets in Twitter in 2 different ways: a tweet mentions a player or doesn't and a tweet is related to an event or isn't. They test their approach on cricket games. The disadvantages are that they don't discriminate between different types of events and that their approach is not real-time. They make use of a Support Vector Machine (SVM) to classify tweets.

III. Event detection in Twitter

The proposed event detection system consists of 6 steps. First the tweets are filtered. Next peak detection is applied. Then the remaining tweets are classified. The classified tweets are used to detect and identify the events in the following 2 steps. Finally, useful information related to each event is extracted. The system is able to detect 6 important soccer events: a goal, a penalty, a foul, a substitution, the end of the first half and the end of the second half.

A. Filtering tweets

To detect events accurately, it's important to only consider tweets that report events that just have happened. Therefore retweets (RT) and replies (@) are removed because they tend to be used when the event is already over. On top of that, they sometimes cause fake events. For the same reasons, tweets that contain a URL are also removed.

B. Peak detection

When an important event happens during a soccer game, we notice a peak in the tweet volume. To detect this peak, an adaptive sliding search window is used [1]. The search window will start with a length of 10 seconds. When no peek can be detected, the window will be iteratively increased to 20, 30 and 60 seconds. A peak is detected when:

$$\frac{\#tweets\ in\ second\ half\ of\ window}{\#tweets\ first\ half\ of\ window} > 1.85$$

under the condition that at least 10 tweets are included in the second half of the search window.

C. Classification of tweets

To detect and identify the events properly, the remaining tweets are classified into 1 of the 7 classes: no event, a goal, a penalty, a foul, a substitution, the end of the first half and the end of the second half. With every tweet, a tweet vector is associated. A tweet vector is a representation of a tweet containing

[1] http://yearinreview.twitter.com/en/tps.html

features that characterize the tweet. The features were selected based on the characteristics of the 6 events. Sentiment features like the use of capital letters and repetition and lexical features like keywords and player names are used. For the classification of the tweets, a Support Vector Machine (SVM) is used. The SVM was trained with 1324 annotated tweets.

D. Event detection

An event is detected when the fraction of tweets in the second half of the search window that belong to the class no event, is lower than 0.55.

E. Event identification

When an event is detected, the event is identified by measuring the fraction of tweets with class y_i in the second half of the search window:

$$\arg\max_{y_i} \frac{\#tweets\ with\ label\ y_i}{\#tweets}, \quad i = 1...7$$

This fraction should be larger than 0.4. Otherwise the event is identified as no event.

F. Extraction of relevant information

The last step is to extract the relevant information related to one of the soccer events. For each type of event, an example is given in Table I. Because the detection of the event penalty was sometimes cumbersome, this information is only used to enhance the goal event.

TABLE I
ILLUSTRATION OF THE OUTPUT OF THE EVENT DETECTION ALGORITHM.

Event	Example
Goal	31': R.van Persie scores! (1-1)
Penalty	7': W.Rooney scores! Penalty. (1-0)
Foul	20': Foul: Yellow Card.
Substitution	61': Substitution : A.Young, L.Nani
End of 1^{st} half	46': End of first half.(2-0)
End of 2^{nd} half	93': End of second half.(2-1)

IV. Highlight extraction in video sequences

For the highlight extraction algorithm, we focus on goals because they are themost important soccer event. We make use of the characteristics of modern soccer broadcast videos that use logo transitions to start and end the replay of a goal.

A. Logo detection

To detect logos in video sequences, standard techniques are used to detect gradual transitions and cuts between 2 shots. The Golden Section composition rule is used to divide every frame in 9 regions and calculate RGB histograms from the 4 most important regions [3]. For every 2 consecutive frames i and $i+1$, the difference $D(i, i+1)$ between those frames is calculated as the sum of the differences between the 4 local histograms. Within the video sequences of the English Premier League games, a small logo enters the video sequence in the middle of the frame and will enlarge until it covers the whole frame. Then it suddenly disappears. This is very similar to a gradual transition that ends with a cut. Therefore, a logo transition can be detected as a series of small but noticable differences $D_{gradual}$ that ends with a larger difference D_{cut}.

B. Combining Twitter with video sequences

When a goal is detected in Twitter, the system will start to analyse the broadcast video sequence during the game looking for the replay fragment. In practice, 2 cases arise: the event is detected before the start of the replay or the event is detected during the replay. To detect the start of the replay, the algorithm will first look for a logo transition in the past with a maximum of 40 seconds. If no logo was found, the algorithm starts looking for a logo in the future with a maximum of 20 seconds. If the first logo couldn't be detected, the algorithm stops. Otherwise the algorithm starts looking for end of the replay with a maximum of 70 seconds.

V. Results

The event detection system is evaluated on 15 games of the English Premier League. The results are shown in Table II. The precision is very high for all events. The recall of the foul and the substition is very low because fans tend to tweet less about less important events or even don't tweet about them at all. The highlight extraction algorithm is evaluated on a subset of 12 games. A high precision and recall of both 93.75% is obtained for the extraction of goal replays.

TABLE II
PRECISION, RECALL AND DELAY OF 6 EVENTS DETECTED IN TWITTER.

Event	Precision	Recall	Med. delay (s)
Goal	95.83%	97.87%	18
Penalty	100.00%	100.00%	10
Foul	100.00%	2.94%	27
Substitution	100.00%	2.50%	13
End 1^{st} half	75.00%	60.00%	24
End 2^{nd} half	100.00%	46.67%	14

VI. Conclusion and future work

We presented a system that is able to detect 6 types of events in Twitter with high precision. The recall of the foul and substition events can still be improved. In the second part of this paper, we presented a system to extract goal replays based on a goal event detection in Twitter. We obtained a high precision and recall.

References

[1] Siqi Zhao, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan, "Human as real-time sensors of social and physical events: A case study of twitter and sports games," *CoRR*, vol. abs/1106.4300, 2011.

[2] Smitashree Choudhury and John Breslin, "Extracting semantic entities and events from sports tweets," in *Making Sense of Microposts (#MSM2011)*, 2011, pp. 22–32.

[3] Li Li, Xiaoqing Zhang, Weiming Hu, Wanqing Li, and Pengfei Zhu, "Soccer video shot classification based on color characterization using dominant sets clustering," in *Proceedings of the 10th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, Berlin, Heidelberg, 2009, PCM '09, pp. 923–929, Springer-Verlag.

Inhoudsopgave

1 Inleiding **1**

1.1 Probleemstelling 1

1.2 Overzicht van de masterproef 2

2 Twitter **3**

2.1 De Tweet 3

2.2 Ondernemingsmodel 4

2.3 Sportevenementen 4

2.4 Karakteristieken van Twittergebruikers 6

2.5 Besluit 7

3 Gerelateerd werk **8**

3.1 Twitter 8

3.1.1 Algemeen 9

3.1.2 Sport 13

3.2 Hoogtepuntextractie uit sportvideo's op basis van audio-, video- en webtekstanalyse 16

3.2.1 Hoogtepuntextractie op basis van videoanalyse 17

3.2.2 Alternatieve manieren van hoogtepuntextractie 19

3.2.3 Alternatieve toepassingen van hoogtepuntextractie 19

3.2.4 Webtekstanalyse 19

3.3 Machinaal leren 20

3.3.1 Naive-Bayes 21

3.3.2 MaxEnt 22

3.3.3 Support Vector Machines 22

3.3.4 Gradient Boosted Decision Trees 24

3.3.5 Vergelijking van de verschillende technnieken 25

3.4 Besluit 26

4 Detectie van gebeurtenissen in Twitter tijdens voetbalwedstrijden 27
4.1 Analyse van het tweetgedrag tijdens voetbalwedstrijden 28
4.1.1 Pieken en dalen 28
4.1.2 Pieken versus dalen 30
4.1.3 Schaduwgebeurtenissen 32
4.1.4 Piekmomenten in detail 34
4.1.5 Besluit 37
4.2 Een algoritme voor gebeurtenisdetectie en -identificatie 37
4.2.1 Filteren van de tweets 39
4.2.2 Piekdetectie 39
4.2.3 Classificatie van de tweets 41
4.2.4 Gebeurtenisdetectie 47
4.2.5 Gebeurtenisidentificatie 47
4.2.6 Extractie van relevante informatie 48
4.2.7 Besluit 52
4.3 Geavanceerde gebeurtenisdetectie en -identificatie 52
4.3.1 Geavanceerde piekdetectie 52
4.3.2 Geavanceerde gebeurtenisdetectie en -identificatie 54
4.3.3 Geavanceerde extractie van relevante informatie 54
4.4 Besluit 56

5 Extractie van hoogtepunten uit voetbalvideosequenties in realtime 57
5.1 Analyse van live-voetbaluitzendingen met betrekking tot doelpunten 57
5.1.1 Cinematografische karakteristieken van live-voetbaluitzendingen . . 57
5.1.2 Cinematografische technieken voor en na een doelpunt 59
5.2 Shotgrens- en logodetectie 60
5.2.1 Lokale histogrammen 60
5.2.2 Shotgrensdetectie 62
5.2.3 Logodetectie 63
5.3 Het combineren van Twitter en videosequenties 64
5.3.1 Vertragingen in Twitter en live-uitzendingen 65
5.3.2 Een algoritme om Twitter en videosequenties te combineren 67
5.4 Besluit 70

6 Evaluatie **71**

6.1 Evaluatie van de individuele stappen van het Twittergebeurtenisdetectiesysteem . . . 72

6.1.1 Filteren van tweets . . . 72

6.1.2 Classificatie van tweets . . . 72

6.1.3 Gebeurtenisdetectie . . . 74

6.1.4 Gebeurtenisidentificatie . . . 76

6.2 Globale evaluatie van het Twittergebeurtenisdetectiesysteem . . . 77

6.2.1 Het doelpunt . . . 77

6.2.2 De strafschop . . . 82

6.2.3 De fout . . . 84

6.2.4 De wissel . . . 86

6.2.5 Het einde van de eerste helft . . . 86

6.2.6 Het einde van de tweede helft . . . 87

6.2.7 Besluit . . . 88

6.3 Evaluatie van de individuele componenten van het hoogtepuntextractiealgoritme . . . 89

6.3.1 Shotgrensdetectie . . . 89

6.3.2 Logodetectie . . . 91

6.4 Globale evaluatie van het hoogtepuntextractiealgoritme . . . 92

7 Besluit en toekomstig werk **95**

7.1 Gebeurtenisdetectie en -identificatie in Twitter . . . 95

7.2 Hoogtepuntextractie uit videosequenties op basis van Twitter . . . 96

A Uitgebreid overzicht van de gebruikte wedstrijden **98**

Bibliografie **106**

Lijst van figuren

2.1 Een schermafdruk van het realtime scoreverloop tijdens een wedstrijd op Twitter. 6

3.1 Schermafdruk van Statler. 12
3.2 Vergelijking tussen samenvattingen op basis van Twitter en op basis van audiovisuele karakteristieken. 16
3.3 Illustratie van de guldensnedecompositieregel. 18
3.4 Illustratie van de 4 belangrijkste shottypes. 18
3.5 Illustratie van lineaire SVM in 2D. 22
3.6 Illustratie van de invloed van de kostparameter op de zachte marges bij een SVM met een lineaire kern. 24
3.7 Illustratie van de invloed van γ op de zachte marges bij een SVM met een RBF kern. 25

4.1 Fragment van de live-verslaggeving op de website van Soccerway. 27
4.2 Grafiek van het aantal berichtjes dat per minuut verzonden werd tijdens Chelsea - Arsenal. 29
4.3 Illustratie van de samenstelling van een dalmoment. 30
4.4 Illustratie van de samenstelling van een piekmoment. 31
4.5 Histogram van de samenstelling van de minuten 6 en 25 tijdens Chelsea-Wolverhampton. 33
4.6 Grafiek van het aantal berichtjes dat per minuut werd verzonden tijdens de eerste helft Chelsea-Wolverhampton. 33
4.7 Schematische voorstelling van het algoritme om gebeurtenissen te detecteren in Twitter 39
4.8 Illustratie van de tweetratio voor een zoekvenster van 20 seconden 53

5.1 Illustratie van de 4 belangrijkste shottypes. 58
5.2 Illustratie van de 2 belangrijkste types overgang tussen 2 shots. 59
5.3 Illustratie van een logotransitie. 59
5.4 Opeenvolging van de types shots voor en na het vallen van een doelpunt. . 60
5.5 Illustratie van een afbeelding en het bijhorend histogram. 61

5.6 Illustratie van de guldensnedecompositieregel. 62
5.7 Illustratie van de 6 laatste beelden van 2 verschillende logotransities. . . . 64
5.8 Kaart van de regio Manchester met de locaties van mobiele gebruikers. . . 66
5.10 Illustratie van het algoritme om videosequenties en Twitter te combineren (geval 1). 68
5.11 Illustratie van het algoritme om videosequenties en Twitter te combineren (geval 2). 69

6.1 Illustratie van de 6 laatste beelden van het eerste type logo. 89
6.2 Illustratie van de 6 laatste beelden van het tweede type logo. 89
6.3 Illustratie van de oorzaak van een falende logodetectie. 92

Gebruikte afkortingen

URL	Uniform Resource Locator
GBDT	Gradient Boosted Decision Tree
SVM	Support Vector Machine
PoS	Part-of-Speech
TF-IDF	Term Frequency - Inverse Document Frequency
HMM	Hidden Markov Model
DCT	Discrete Cosine Transform
FSM	Finite State Machine
MAP	Maximum A Posteriori
RBF	Radial Base Function
CART	Classification And Regression Tree
MPEG	Moving Picture Experts Group
RGB	Red Green Blue
HSV	Hue Saturation Value
HSI	Hue Saturation Intensity

Hoofdstuk 1

Inleiding

1.1 Probleemstelling

Sociale media zijn op dit moment het belangrijkste medium om sportwedstrijden te volgen [1]. Maar liefst 41% van de supporters beschouwt Twitter en Facebook als de media bij uitstek om op de hoogte te blijven van de vorderingen van hun favoriete ploeg of speler tijdens een wedstrijd. Sportwebsites vormen de tweede beste bron met een voorkeur van 40% van de supporters.

Twitter is hét sociale netwerk om snel op de hoogte te zijn van gebeurtenissen. In tegenstelling tot vele andere sociale netwerken, is Twitter publiek toegankelijk en laat het gebruikers toe automatisch berichten van anderen te ontvangen die dezelfde interesses delen. Bovendien mag een bericht op Twitter slechts 140 karakters bevatten. Dit bevordert de snelheid waarmee supporters gebeurtenissen rapporteren.

Wanneer supporters een wedstrijd willen volgen via Twitter maken ze gebruik van een hashtag, die eigen is aan een ploeg of een speler, om berichten te groeperen. Voor de voetbalploeg Chelsea is dit bijvoorbeeld #CFC. Het probleem hierbij is echter dat gebruikers niet enkel berichten vinden over de wedstrijd die ze willen volgen, maar ook clubnieuws en berichten die helemaal niet met Chelsea gerelateerd zijn. De afkorting CFC is immers niet exclusief toegekend aan berichten over de voetbalploeg Chelsea. Een systeem dat in staat is om automatisch belangrijke gebeurtenissen tijdens een voetbalwedstrijd uit deze berichten te filteren, zou hier echter een oplossing kunnen bieden. Bovendien kan deze informatie gebruikt worden om in realtime aan sportverslaggeving te doen en wedstrijdsamenvattingen te generen.

Parallel met Twitter zijn er ook een live-uitzendingen op de televisie. Het laatste decennium is er reeds veel onderzoek verricht naar het extraheren van belangrijke gebeurtenissen uit sportvideosequenties. De gebruikte technieken steunen veelal op de audiovisuele

karakteristieken van de sportwedstrijd. Bijgevolg gaat de analyse gepaard met rekenintensieve operaties en kunnen een aantal belangrijke semantische concepten niet altijd afgeleid worden. Denk hierbij bijvoorbeeld aan wie het doelpunt gescoord heeft. Een systeem dat de juiste videofragmenten selecteert op basis van de gebeurtenissen gedetecteerd in Twitter kan beide problemen omzeilen. De analyse van tekst is minder rekenintensief dan de analyse van video. De tekst laat bovendien toe de videofragmenten veel rijker semantisch te annoteren.

In deze masterproef hebben we er voor gekozen om een systeem te implementeren dat in realtime gebeurtenissen in Twitter detecteert. Dit sluit nauw aan bij zowel het doel van Twitter, namelijk de snelle verslaggeving, als de manier waarop supporters sport beleven. Als toepassing van het systeem kiezen we ervoor om de gedetecteerde gebeurtenis te verrijken met een bijhorend videofragment. Omgekeerd kan dit ook beschouwd worden als het realtime detecteren en annoteren van hoogtepunten in een sportvideo op basis van Twitter. Als sport hebben we gekozen voor voetbal omdat dit de meest beoefende sport ter wereld is. Als competitie hebben we gekozen voor de Engelse Premier League omdat dit onder andere de meest bekeken competitie ter wereld is, de Twitterberichten Engelstalig zijn en omdat er voldoende wedstrijden in België worden uitgezonden.

1.2 Overzicht van de masterproef

Na deze korte algemene inleiding, overlopen we in Hoofdstuk 2 de basisconcepten van Twitter en de karakteristieken van een typische Twittergebruiker. In Hoofdstuk 3 verdiepen we ons in reeds bestaande literatuur en worden enkel concepten geïntroduceerd die gebruikt zullen worden in de ontwikkeling van de verschillende algoritmes. In Hoofdstuk 4 introduceren we een algoritme dat in staat is verschillende gebeurtenissen in Twitter te detecteren en te identificeren. In Hoofdstuk 5 maken we gebruik van dit algoritme om hoogtepunten uit videosequenties te extraheren en te annoteren. Daarna vindt in Hoofdstuk 6 een evaluatie plaats van het finale algoritme en de verschillende deelcomponenten. We beëindingen deze masterproef met een besluit in Hoofdstuk 7 en geven een overzicht van mogelijke verbeteringen en andere onderzoekspaden.

Hoofdstuk 2

Twitter

Twitter definieert zichzelf als een realtime informatienetwerk dat de gebruiker verbindt met de laatste verhalen, ideeën, opinies en nieuwsfeiten die de gebruiker interesseren [2].

> "De snelste en makkelijkste manier om dichtbij hetgeen te blijven waar jij om geeft." - Twitter

Op het moment van schrijven (april 2012) heeft Twitter 140 miljoen actieve gebruikers. Zij zijn goed voor een gemiddelde van 340 miljoen tweets per dag [2].

2.1 De Tweet

Aan de basis van Twitter ligt het concept *tweet*. Een tweet is een bericht van maximaal 140 karakters dat gebruikers de wereld insturen of verdelen binnen hun eigen sociaal netwerk. Oorspronkelijk beantwoordde een tweet de vraag "Wat ben je aan het doen?". Tegenwoordig is Twitter medium nummer 1 om snel informatie te verkrijgen en te verspreiden. Bovendien is de functionaliteit van een tweet uitgebreid met foto- en videomogelijkheden. Een tweet bevat heel wat tekens om de communicatie te vereenvoudigen, we beschrijven hier de belangrijkste die nodig zijn om wat volgt te begrijpen.

Het hekje (#) Het hekje (Engels: hash) wordt gebruikt in een tweet om het onderwerp of een trefwoord aan te geven. Een *hashtag* bestaat uit de opeenvolging van een hekje en een sleutelwoord. Twittergebruikers gebruiken het om hun berichten te kunnen onderverdelen in categorieën. De Twitterzoekmachine maakt hiervan gebruik om tweets op onderwerp te indexeren en makkelijk vindbaar te maken [3].

Het apenstaartje (@) Daar waar het hekje gevolgd wordt door een trefwoord of onderwerp, wordt een apenstaartje gevolgd door de naam van een andere gebruiker. Elke gebruiker kan door middel van een apenstaartje gerefereerd worden. Op deze manier wordt een tweet gericht aan een bepaalde persoon. Er zijn 2 verschillende manieren van gebruik. Indien een tweet begint met een apenstaartje gevolgd door een gebruikersnaam, dan is dit een antwoord op een reeds bestaande tweet. Indien een apenstaartje wordt gevolgd door een gebruikersnaam midden in een tweet, dan spreekt men over een vermelding [3].

Retweet (RT) Naast het beantwoorden van een tweet kan men een tweet van een gebruiker ook opnieuw verzenden naar andere gebruikers. Dit concept noemt men retweeten. Een retweet zal voorafgegaan worden door de letters RT [3].

2.2 Ondernemingsmodel

Het inkomstenmodel van Twitter is gebaseerd op advertenties. Twitter biedt bedrijven 3 mogelijkheden aan om in beeld te komen bij potentiële klanten, namelijk gepromote tweets, gepromote gebruikers en gepromote trends.

Een systeem dat automatisch detecteert wanneer de meest interessante momenten in sportwedstrijden zich voordoen, kan helpen om gerichter te adverteren. Gepromote tweets kunnen verstuurd worden wanneer het meeste aantal gebruikers actief is op Twitter.

2.3 Sportevenementen

Supporters maken vaak gebruik van Twitter tijdens sportevenementen. Tabel 2.1 geeft een overzicht van de 16 evenementen in 2011 waarbij het hoogst aantal tweets per seconde geregistreerd werden. De winnaar is de "MTV Music and Video Awards" waar er op een bepaald moment een piek van 8868 tweets per seconde werd gemeten. Opmerkelijk is dat maar liefst 6 van de 16 evementen sportevenementen zijn [4].

Het marketingbureau GMR [1] deed een onderzoek bij 350 sportfanaten over hoe zij sport beleven anno 2012:

- **Wat is de beste bron om brekend sportnieuws te vinden?** 41% gebruikt Twitter of Facebook, 40% gebruikt sportwebsites, 13% gebruikt de televisie en slechts 3% de radio. De overige 3% gebruikt andere bronnen.

Tabel 2.1: Overzicht van de 16 evenementen met het hoogste aantal tweets/seconde in 2011[4]. Evenementen met een (*) zijn sportwedstrijden.

Datum	**Evenement**	**Tweets/sec**
1 jan	Nieuwjaar	6939
6 feb	Superbowl*	4064
11 maa	Japanese aardbeving en tsunami	5530
29 apr	Trouw Prins William en Kate	3966
2 mei	Raid op Osama bin Laden	5106
28 mei	UEFA Champions League*	6303
13 jun	NBA finale*	5531
27 jun	BET Awards	6436
11 jul	Home Run Derby*	4995
17 jul	Brazilië uitgeschakeld in Copa America*	7166
17 jul	Einde van FIFA wereldbeker voor vrouwen*	7196
23 aug	Aardbeving Amerikaanse oostkust	5449
25 aug	Steve Jobs neemt ontslag	7064
28 aug	MTV Video Music Awards	8868
20 sep	Troy Davis geëxecuteerd	7671
6 okt	Steve Jobs is dood	6049

- **Waar heeft u al sociale media gebruikt om sportinformatie te verkrijgen?** 86% heeft al tijdens de werkuren social media gebruikt, 74% op een uitgaansgelegenheid, 69% tijdens het eten, 58% in de badkamer of het toilet, 33% tijdens een vergadering op het werk, 22% tijdens een concert en zelfs 9% tijdens een misviering in de kerk.

- **Heeft u al sociale media gebruikt voor het vinden van sportinformatie tijdens één van volgende activiteiten?** 83% heeft al sociale media gebruikt tijdens het bekijken van de wedstrijd op televisie. Maar liefst 63% heeft al sociale media gebruikt terwijl ze naar de wedstrijd kijken op de speellocatie. 62% heeft al sociale media gebruikt tijdens wedstrijdanalyses op televisie voor en na de wedstrijd. 61% heeft al sociale media gebruikt tijdens het surfen op sportwebsites en 30% heeft al sociale media gebruikt terwijl ze de wedstrijd via de radio aan het volgen waren.

Alhoewel het onderzoek bij een beperkte groep mensen is afgenomen geeft het een goede indicatie over de relevantie van Twitter als medium bij uitstek voor het vinden van sportnieuws en het delen van informatie. De meest frapante cijfers zijn toch dat 41% sociale media als beste bron beschouwen om sportnieuws heet van de naald te vinden en dat 83% van de fans die de wedstrijd via televisie volgen op hetzelfde moment ook al eens sociale media hebben gebruikt.

Ook Twitter is op de kar gesprongen om zelf verslaggeving te doen tijdens voetbalwedstrijden. Sinds februari 2012 loopt er een proefproject op de Duitstalige versie van Twitter waarbij een gebruiker die een zoekopdracht invoert naar een ploeg die op dat moment ook

Figuur 2.1: Een schermafdruk van het realtime scoreverloop tijdens de wedstrijd Schalke 04 - Borussia Dortmund van 14/4/2012 op Twitter.

speelt, een scorebord te zien krijgt met de stand op dat moment tijdens de wedstrijd. Verder promoot Twitter ook een eigen hashtag voor elke wedstrijd die gevormd wordt door de 3 letters die verwijzen naar de thuisploeg en 3 letters die verwijzen naar de uitploeg. In Figuur 2.1 zien we het scoreverloop van de wedstrijd Schalke 04 - Borussia Dortmund wanneer we zoeken op #S04 (Schalke 04), #BVB (Borussia Dortmund) of #S04BVB.

2.4 Karakteristieken van Twittergebruikers

Wat is nu de drijfveer van Twittergebruikers? In [5] proberen de auteurs te achterhalen waarom mensen Twitter gebruiken. De auteurs onderscheiden 4 intenties (anno 2007):

- De dagelijkse babelaar: Dit is de meest voorkomende Twittergebruiker. Hij praat over zijn dagelijkse routine of over wat hij op dit moment aan het doen is.
- Conversaties: Gebruikers reageren op elkaars tweets door gebruik te maken van het @-symbool. Ongeveer 21% van de gebruikers gebruikt deze techniek en 1 op 8 tweets bevat het @-symbool.
- Het delen van informatie en URL's: Ongeveer 13% van alle tweets bevat een URL, al dan niet spam.

- Verspreiden van nieuws: Veel gebruikers verspreiden het laatste nieuws of reageren op de laatste nieuwtjes. Sommige gebruikers zijn automatische verdelers van berichten.

Naast de 4 intenties die aangeven waarom mensen Twitter gebruiken onderscheiden de auteurs ook 3 types gebruikers:

- De informatiebron: Deze gebruiker heeft typisch een groot aantal volgers en zal op een al dan niet frequente manier informatie verspreiden. Hij heeft net een groot aantal volgers omdat de tweets die hij verzendt waardevolle informatie bevatten. Vaak gebeurt dit verzenden automatisch. Voorbeelden zijn onder andere nieuwswebsites of beroemdheden, maar het kunnen evengoed gewone gebruikers zijn die groot geworden zijn door Twitter.
- Vrienden: Sommige gebruikers gebruiken Twitter voornamelijk om sociale contacten te onderhouden met vrienden, familie of collega's.
- De informatiezoeker: Dit type gebruiker zal niet actief tweets verzenden maar is eerder opzoek naar informatie. Hij zal veel gebruikers volgen die regelmatig tweets verzenden met voor hem interessante informatie.

Een gebruiker is niet beperkt tot 1 type maar kan uiteraard meerdere rollen innemen.

2.5 Besluit

In dit hoofdstuk hebben we enkele basisconcepten van Twitter geïntroduceerd. Naast de karakteristieken van de doorsnee Twittergebruiker hebben we ook de drijfveer van een sportfanaat geanalyseerd. Sportfanaten beschouwen sociale media als de nummer 1 bron om informatie over een wedstrijd te verkrijgen. Vaak gebeurt dit op het werk of tijdens uitgaansgelegenheden. Bovendien wordt het gebruik van Twitter vaak gecombineerd met het bekijken van een sportwedstrijd of zelfs tijdens het bekijken van een wedstrijd in een stadion. Snel informatie verkrijgen op eender welk moment blijkt van cruciaal belang te zijn voor een sportfanaat.

Hoofdstuk 3

Gerelateerd werk

Om het systeem, beschreven in deze masterproef, te realiseren kunnen we de gerelateerde literatuur onderverdelen in 3 grote delen. Eerst bekijken we de bestaande literatuur rond de analyse en classificitatie van Twitterberichten waarbij we de focus leggen op algemene systemen voor gebeurtenis- en sentimentanalyse. We besteden hier ook uitgebreid aandacht aan de beperkte literatuur rond de analyse en toepassing van Twitter tijdens sportwedstrijden. Het tweede onderdeel bespreekt de bestaande literatuur die hoogtepunten uit sportvideosequenties probeert te extraheren door middel van video-, audio- en webtekstanalyse. In het laatste deel behandelen we nog enkele concepten uit de theorie van het machinaal leren die we in deze masterproef zullen toepassen.

3.1 Twitter

Twitter, en meer algemeen micro-blogging, is een relatief nieuw onderzoeksthema binnen de computerwetenschappen. Sinds de oprichting van Twitter is er al redelijk wat onderzoek verricht in het domein van de natuurlijke taalverwerking door computers. Dit onderzoek bestaat voornamelijk uit het toepassen van bestaande technieken, die vroeger werden toegepast op formele documenten, op informele spreektaal. Dat Twitter een schat aan informatie bevat blijkt ook uit de belangrijkste thema's die hierbij aan bod komen. We beginnen deze sectie met een algemeen overzicht (Sectie 3.1.1) van de meest interessante onderzoeken rond Twitter die waardevol kunnen zijn voor het vervolg van deze masterproef. Vervolgens geven we in Sectie 3.1.2 een uitgebreide bespreking van gerelateerd werk dat zich specifiek op Twittergebruik tijdens sportwedstrijden focust.

3.1.1 Algemeen

Het eerste onderwerp dat we belichten is de detectie en classificatie van gebeurtenissen in Twitter. Vervolgens gaan we verder met het detecteren van deze gebeurtenissen in realtime en de manier waarop het gedrag van gebruikers wijzigt wanneer er een interessante gebeurtenis plaatsvindt tijdens een evenement. Het volgende grote thema is de sentimentanalyse. Hierbij wordt geprobeerd de neutraliteit of polariteit van een tweet te bepalen om zo een opinie over producten, personen of evenementen te extraheren. Finaal bespreken we nog enkele artikels waarin het classificeren van gebruikers en tweets binnen een bepaald thema behandeld wordt (bijvoorbeeld: nieuws, opinie, persoonlijk...).

3.1.1.1 Evenement- en gebeurtenisanalyse

Controversiële gebeurtenissen tijdens een evenement lokken vaak een publieke discussie uit waarbij toeschouwers tegengestelde opinies hebben, verrast zijn of blijk geven van ongeloof. De auteurs in [6] proberen op zoek te gaan naar deze gebeurtenissen door gebruik te maken van momentopnames. Hierbij wordt elke momentopname gekenmerkt door een persoon (in dit geval een beroemdheid die onderwerp van onderzoek is), een tijdsspanne (1 dag) en een verzameling van tweets betreffende deze persoon en binnen deze tijdsspanne. De auteurs evalueren een aantal systemen waarbij het belangrijkste systeem eerst een gebeurtenisdetectie doet en vervolgens een controversiële score berekent. Voor de gebeurtenisclassificatie wordt gebruik gemaakt van machinaal leren, meer bepaald van Gradient Boosted Decision Trees (GBDT's), zie Sectie 3.3.4. De geselecteerde kenmerken zijn linguïstisch, structureel, vergelijkingen met vorige momentopnames en het aantal positieve, negatieve en neutrale tweets. Het meest uitgebreide model dat getest wordt, maakt bovendien gebruik van andere bronnen zoals nieuwswebsites. Het beste systeem behaalt een gemiddelde precisie van 61,8%.

Dezelfde auteurs [7] bouwen verder op dit systeem wanneer ze de betrokkenheid van gebruikers in het becommentariëren van gebeurtenissen onderzoeken. Hiervoor beschouwen de auteurs 4 soorten beroemdheden: acteurs, atleten, musici en politici. De auteurs concluderen dat Twittergebruikers het meest intens reageren op gebeurtenissen rond acteurs. Verder blijkt dat gebruikers voornamelijk positief reageren op een gebeurtenis. Dit maakt Twitter een positief getint medium. Positieve emoties komen 6 keer zo vaak voor als negatieve.

3.1.1.2 Analyse van het Twittergedrag tijdens evenementen

In [8] onderzoeken de auteurs hoe het Twittergedrag van gebruikers wijzigt tijdens massa-evenementen en noodsituaties. De auteurs analyseerden de democratische en republikeinse

congressen en het aan land komen van de orkanen Gustav en Ike. De belangrijkste conclusie is dat er minder @-gebruik is tijdens deze evenementen dan op andere momenten. Ook is er een opmerkelijke stijging in het aantal tweets dat een URL bevat. Dit bevestigt de algemene trend dat Twitter meer en meer als informatieverdeler wordt gebruikt.

In [9] wordt er dieper ingegaan op het @-gebruik tijdens evenementen. In dit artikel keken de auteurs hoe het gedrag van de Twittergebruikers wijzigde tijdens de inhuldiging van president Barack Obama. Hierbij stellen ze een sterke positieve correlatie vast tussen gemiddeld aantal karakters per tweet per minuut en het aantal keer dat het @-symbool wordt vermeld per minuut. Verder blijkt een abrupte daling van het @-gebruik te wijzen op het begin van een belangrijke gebeurtenis tijdens een evenement. Een stijging van het @-gebruik bleek echter niet altijd het einde van een gebeurtenis aan te geven.

3.1.1.3 Realtime gebeurtenisanalyse

In [10] wordt voor het eerst een praktische dienst beschreven die gebeurtenissen detecteert in Twitter. De auteurs hebben een dienst ontwikkeld dat in realtime aardbevingen en tyfonen detecteert door gebruik te maken van Twitter en vervolgens een e-mail stuurt naar alle ingeschrevene om hen te waarschuwen. Voor het detecteren van een aardbeving worden binnenkomende tweets eerst geclassificeerd door middel van een Support Vector Machine (SVM), zie Sectie 3.3.4. De SVM maakt gebruik van 3 soorten eigenschappen van een tweet: statistische eigenschappen, sleutelwoordeigenschappen en contexteigenschappen. Indien men op een bepaald moment een stel tweets heeft die een aardbeving signaleren, wordt er een probabiliteitsmodel toegepast om finaal te beslissen of er een aardbeving heeft plaatsgevonden. Het systeem detecteert aardbevingen met een probabiliteit van 96%.

3.1.1.4 Sentimentanalyse

Voor bedrijven is het heel belangrijk dat hun product op een positieve manier in de media verschijnt. Op Twitter worden heel wat meningen gedeeld over deze producten. Sentimentanalyse is de techniek die tweets probeert te classificeren als positief, neutraal of negatief. In [11] wordt een eerste poging ondernomen. De tweets worden verzameld op basis van emoticons die bovendien als label van de tweet dienden. De auteurs testen 3 technieken van machinaal leren, namelijk een SVM, Naive-Bayes en MaxEnt (zie Sectie 3.3). Als kenmerken worden unigrams[1] gebruikt, bigrams[2] en Part-of-Speech (PoS) gebruikt. De SVM presteert het best op unigrams. Naive-Bayes en MaxEnt presteren het best op een combinatie van unigrams en bigrams alhoewel het verschil beperkt is.

[1]De aanwezigheid van elk woord wordt als een apart kenmerk beschouwd. Zie Sectie 3.3.
[2]De aanwezigheid van 2 opeenvolgende woorden wordt als kenmerk beschouwd. Zie Sectie 3.3

PoS zorgt voor een daling in de accuraatheid van de SVM en Naive-Bayes en heeft een verwaarloosbaar effect op MaxEnt.

In [12] wordt er verder gebouwd op deze resultaten. De data werden verzameld op basis van reeds bestaande sentimentanalysesystemen, zogehete mash-ups: Twendz, Twitter Sentiment en TweetFeel. De auteurs ontwikkelden een 2-delig systeem waarbij tweets eerst onderverdeeld worden in 2 categorieën: objectief en subjectief. Vervolgens wordt de polariteit van de subjectieve tweets als positief of negatief bepaald. Er worden 2 soorten eigenschappen gebruikt: meta-eigenschappen, die informatie geven over het type woord (PoS, positief woord,...) en syntactische eigenschappen, zoals het gebruik van hoofdletters, emoticons, links,... Om het onderscheid te maken tussen objectieve en subjectieve tweets blijkt de grootste informatiewinst te halen uit de volgende eigenschappen: het gebruik van positieve woorden, links, subjectieve woorden, hoofdlettergebruik en werkwoorden. Voor het bepalen van de polariteit zijn de meest interessante eigenschappen: positieve woorden, negatieve woorden, werkwoorden, positieve emoticons en hoofdlettergebruik. Verder blijkt experimenteel dat een SVM het beste resultaat geeft en dat de voorgestelde selectie van eigenschappen betere resultaten geeft dan het gebruik van unigrams. Voor de subjectiviteitsdetectie rapporteren de auteurs een foutenpercentage van 18,1% voor de voorgestelde aanpak tegenover 27,6% voor de aanpak op basis van unigrams en voor de polariteitsdetectie een foutenpercentage van 18,7% voor de voorgestelde aanpak tegenover 20,9% voor de aanpak op basis van unigrams.

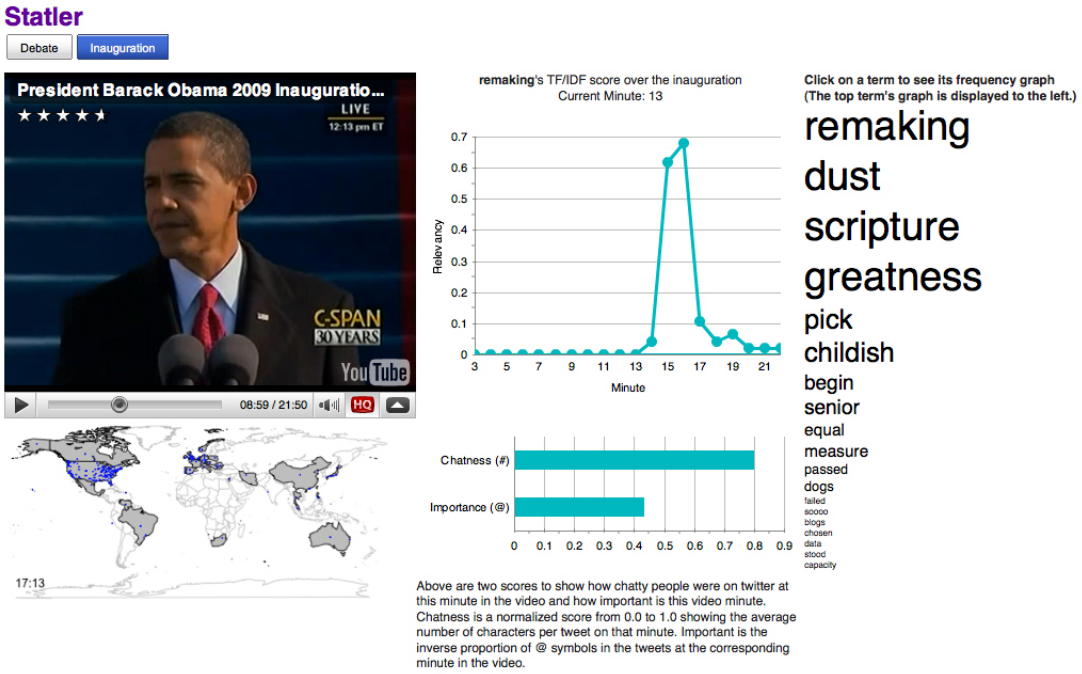
Dezelfde auteurs verbeteren dit resultaat door op te merken dat de Twittertaal voortdurend evolueert. Zo tonen ze aan dat de woordfrequentie in tweets gemiddeld lager ligt dan in formele teksten [13]. Als oplossing hiervoor introduceren ze onderwerpklassen. Sommige woorden of producten komen typisch in een postieve context voor, andere in een negatieve. Alle woorden kunnen via externe diensten zoals AlchemyAPI[3] herleid worden tot enkele onderwerpklassen. Na evaluatie met een Naive-Bayes classificeerder geeft dit een stijging van de accuraatheid van 81% naar 86,3% ten opzichte van unigrams.

Samengevat leren we dat micro-blog eigenschappen zoals het gebruik van hoofdletters, herhaling en emoticons een sterke informatiewinst opleveren [12]. Het gebruik van lexicon met typische sentimentwoorden [12] of onderwerpen [13] heeft ook een positieve invloed op de accuraatheid. PoS-technieken blijken slechts in beperkte mate een positieve invloed te hebben [12]. Sommige auteurs rapporteren zelfs een daling in accuraatheid [11].

3.1.1.5 Classificatie van tweets

Naast het classificeren van tweets op basis van sentiment kunnen we tweets ook in een aantal andere categorieën verdelen. Zo proberen de auteurs in [14] tweets te verdelen

[3] http://www.alchemyapi.com/



Figuur 3.1: Schermafdruk van Statler. De inauguratie van Barack Obama wordt afgespeeld. In het midden wordt een analyse weergegeven van de corresponderde tweets tijdens de 13de minuut [16].

in de categorieën nieuws, evenementen, opinies, koopjes en private berichten. Hierbij maken ze gebruik van een Naive-Bayes classificeerder en 7 binaire eigenschappen van tweets: (1)gebruik van Twitterafkortingen, (2)tijd-evenement zinnen, (3)opiniewoorden, (4)nadruk op woorden (herhaling), (5)munt- en percentagetekens, (6)het gebruik van een @-symbool in het begin van een tweet en (7)het gebruik van een @-symbool in de tweet. Bovendien werd rekening gehouden met welke gebruikers de tweet stuurden. De Naive-Bayes classificeerder werd aangevuld met een bag-of-words strategie. De auteurs behalen een accuraatheidspercentage tussen de 90% en 100% voor de 5 categorieën.

In [15] probeert de auteur tweets te classificeren in 3 categorieën: gebruiker-, nieuws- of bedrijfstweets. Er wordt een accuraatheid gerapporteerd van 80% op basis van een SVM die gebruik maakt van unigrams. Ook probeert de auteur een onderscheid te maken tussen tweets die feiten bevatten en tweets die opinies bevatten. Hier zijn echter geen accuraatheidscijfers van bekend.

3.1.1.6 Video-annotatie op basis van Twitter

We eindigen deze sectie met het bespreken van 2 artikels die live-uitzendingen proberen te voorzien van sleutelwoorden op basis van Twitter. In [16] proberen de auteurs de principes die ze eerder beschreven in [9] nu toe te passen in een mash-up, Statler genaamd. In Figuur 3.1 wordt hier een schermafdruk van getoond. De auteurs maken gebruik van 2 indicatoren die aanduiden of er iets interessant is gebeurd tijdens een geanalyseerde

minuut. *Chatness* geeft aan hoeveel karakters per tweet gemiddeld gebruikt worden en *Importance* geeft aan hoeveel tweets een @-symbool bevatten. Verder extraheren ze voor elke minuut de meest belangrijke woorden die getweet worden om zo makkelijk te kunnen navigeren naar een bepaalde gebeurtenis in de video.

Hetzelfde idee wordt ook beschreven in [17]. De auteurs stellen een raamwerk voor dat het mogelijk maakt clips te annoteren en te archiveren op basis van een chatbox die beschikbaar is bij elke live-uitzending op UStream[4]. De gebruikers loggen hier in via een sociaal netwerk zoals Facebook of Twitter. Gebeurtenissen worden geïdentificeerd op basis van een similariteitsscore:

$$Sim(bericht_1, bericht_2) = \alpha * exp(-|tijdstip_1 - tijdstip_2|) + \beta * cos(vector_1, vector_2) \\ +\gamma * isVriend(gebruiker_1, gebruiker_2)$$

Hierbij modelleert de eerste term de gelijkheid van 2 berichten op basis van tijdstip van verzenden. De tweede term berekent de similariteit tussen 2 berichten op basis van de inhoud van de berichten door 2 vectoren te gebruiken die opgesteld zijn volgens het Term Frequency-Inverse Document Frequency (TF-IDF) principe. De derde term gaat het er vanuit dat als 2 personen vriend zijn van elkaar, de kans groter is dat ze over hetzelfde praten. De video kan vervolgens worden gesegmenteerd in clips. Gebruikers kunnen nadien via sleutelwoorden het archief raadplegen.

3.1.2 Sport

In de context van Twitteranalyse is er nog maar weinig literatuur beschikbaar dat zich focust op sportwedstrijden. Het eerste artikel dat we bespreken geeft een degelijke diepte-analyse van het realtime aspect van Twitter en sport. Het vormt tevens de basis van deze masterproef. We gaan vervolgens verder met 2 artikels die een uitgebreidere aanpak beschrijven voor het analyseren van sporttweets. We sluiten deze sectie af met 2 artikels die tweets gebruiken om videosamenvattingen te genereren. Tabel 3.1 geeft een overzicht van de 5 besproken artikels.

3.1.2.1 Realtimedetectie en -herkenning van gebeurtenissen

In [18] analyseren de auteurs het realtimekarakter van Twitter tijdens wedstrijden van de US National Football League 2010-2011. De auteurs proberen ondermeer een schatting te maken van de snelheid waarmee gebeurtenissen in een wedstrijd gerapporteerd worden. De kleinste vertraging bedraagt 13 seconden, de grootste vertraging 27 seconden en gemiddeld duurt het 17 seconden voordat een gebeurtenis voor het eerst getweet wordt. Mobiele

[4] http://www.ustream.tv/

Tabel 3.1: Overzicht van de 5 artikels die analyse van sportwedstrijden op Twitter behandelen.

Artikel	**Sport**	**Techniek (Algoritme + kenmerken)**
[18]	American Football (2010-2011 NFL)	Zoekvenster + sleutelwoorden
[19]	Cricket (ICC World Cup 2011)	Machinaal leren (niet gespecifieerd) + syntactische kenmerken, spelersnamen, sleutelwoorden
[20]	American Football (3 zondagen 2010-2011 NFL)	HMM + sleutelwoorden, spelersnamen
[21]	Voetbal (World Cup 2010)	Tweetvolume, TF-IDF
[22]	4x Voetbal + 4x Rugby	Tweetvolume

gebruikers zijn typisch enkele seconden trager en de meest interessante gebeurtenissen worden het snelst getweet. Verder stellen de auteurs een algoritme voor om 4 verschillende gebeurtenissen te detecteren en te identificeren. Het algoritme werkt met een zoekvenster om een sterke stijging in het Twitterverkeer te detecteren. Als $\frac{\#tweets\ in\ 2de\ helft\ zoekvenster}{\#tweets\ in\ 1ste\ helft\ zoekvenster} > grenswaarde$, wordt er een gebeurtenis gedetecteerd. Vervolgens wordt de gebeurtenis geïdentificeerd op basis van een lexicon. Elke tweet wordt aan een gebeurtenis gelinkt op basis van sleutelwoorden. De meest voorkomende gebeurtenis die bovendien een tweede grenswaarde overschrijdt, wordt als gebeurtenis aangenomen. De auteurs rapporteren een accuraatheid van 90% voor de meest voorkomende gebeurtenis, de *touchdown*, maar een zwakker resultaat voor de minst significante gebeurtenis, de *fumble*, namelijk 64%.

3.1.2.2 Analyse van sportgebeurtenissen

In [19] maken de auteurs als enige auteurs van Sectie 3.1.2 gebruik van machinaal leren om spelers en gebeurtenissen te detecteren in tweets van sportwedstrijden. Het eerste luik van het artikel draait rond het detecteren van een speler in een tweet. Hierbij maken de auteurs gebruik van de spelersnamen die verzameld worden via de teamwebsite en contextwoorden via Wikipedia[5] die gerelateerd zijn aan de sport. Vervolgens wordt er een vector opgesteld die uitsluitend binaire karakteristieken bevat: gebruik van voornaam, achternaam, initialen, contextwoord, hoofdlettergebruik, @spelersnaam en #spelersnaam. De classificatie geeft een precisie van 83% en 86% voor respectievelijk 'bevat speler' en 'bevat geen speler'. De recall is respectievelijk 70% en 90%. In het tweede luik proberen de auteurs tweets te identificeren die verwijzen naar een gebeurtenis binnen de wedstrijd. Hierbij maken ze geen onderscheid tussen de verschillende gebeurtenissen. Ze maken gebruik van contextwoorden alsook van syntactische kenmerken zoals hoofdlettergebruik. Ook wordt informatie van het interval waarin de tweet zich bevindt gebruikt. Zo wordt

[5]http://en.wikipedia.org

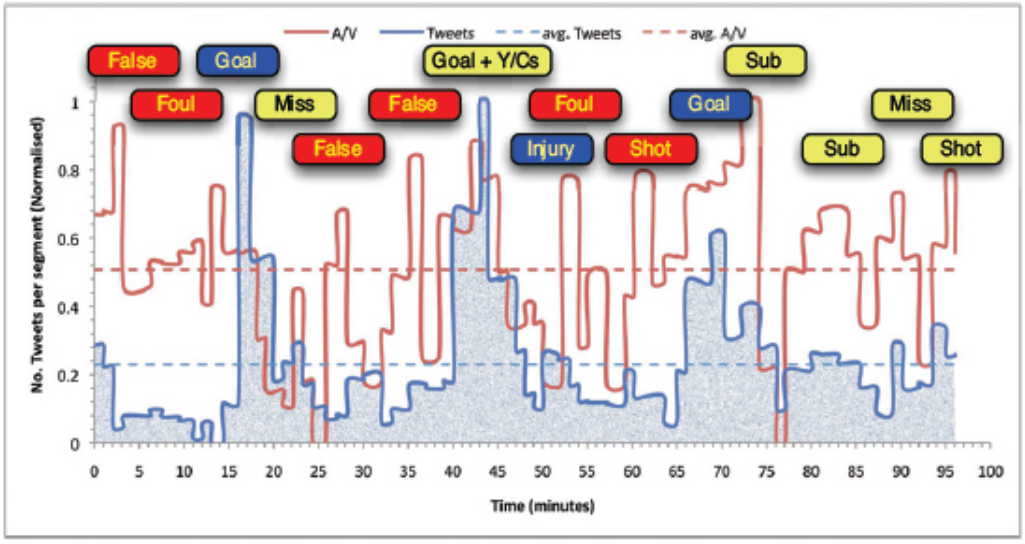
het tweetvolume in rekening gebracht per interval van 2 minuten en de verhouding retweets/gewone tweets. Tijdens een belangrijk moment komen er minder retweets voor dan in een interval waar er niets interessant gebeurt. De classificatie geeft een precisie van 85% en 87% voor respectievelijk tweets die een gebeurtenis beschrijven en tweets die geen gebeurtenis beschrijven. De recall is respectievelijk 86% en 89%.

De auteurs in [20] proberen niet de tweets afzonderlijk te classificeren maar stellen een systeem voor dat een tweetsamenvatting maakt van een volledige sportwedstrijd op voorwaarde dat een gebeurtenis gedetecteerd is. Hiervoor maken ze gebruik van een Hidden Markov Model (HMM). Ze onderscheiden 3 soorten probabiliteiten: een probabiliteit die specifiek is voor elke toestand maar gelijk is voor elke gebeurtenis, een probabiliteit die specifiek is voor een toestand en een wedstrijd (bv. spelersnamen) en een probabiliteit die een achtergronddistributie beschrijft om ruis op te vangen. Ter evaluatie wordt een lijst van 20, 30, 40, 50, 60 of 70 tweets opgevraagd en wordt er nagegaan welke tweets een bepaalde fase in de wedstrijd beschrijven en welke niet. Indien het aantal tweets beperkt is tot 30 zal de recall slechts 50% zijn, maar deze stijgt naar 90% indien men 70 tweets opvraagt over een *touchdown*. Opnieuw blijkt dat een *touchdown* makkelijk te analyseren valt dan andere gebeurtenissen [18].

3.1.2.3 Videosamenvattingen op basis van Twitter

Het tweetvolume blijkt een goede indicator te zijn voor interessante gebeurtenissen tijdens een sportwedstrijd [18, 19]. Dit maakt het mogelijk om op eenvoudige wijze een videosamenvatting te maken op basis van de hoogste pieken in tweetvolume. De auteurs in [22] vergelijken de prestatie van videosamenvattingen op basis van het tweetvolume tegenover videosamenvattingen op basis van audiovisuele karakteristieken. Het doel is in beide gevallen om de 10 belangrijkste momenten te identificeren en vervolgens een fragment te selecteren van minimaal 60 seconden door gebruik te maken van shotgrensdetectie. Verder wordt er aan elk fragment een sleutelwoord toegekend op basis van het meest voorkomende woord in het desbetreffende tijdsinterval. De evaluatie gebeurt op 3 voetbalwedstrijden waarbij wordt nagegaan of alle doelpunten en gele kaarten in de samenvattingen te vinden zijn. Alle samenvattingen bevatten de doelpunten. De Twittersamenvattingen bevatten slechts 1/3 van de gele kaarten terwijl de audiovisuele samenvattingen alle gele kaarten weten te lokaliseren. De audiovisuele samenvattingen genereren echter ook veel valse positieven (Een fragment wordt herkend als een gele kaart maar bevat geen gele kaart fase). Hierdoor ontstaat de kans dat niet alle meest significante gebeurtenissen zich in de top 10 zullen bevinden. Figuur 3.2 vergelijkt Twitterkarateristieken met audiovisuele karakteristieken en duidt met labels aan welke gebeurtenissen door de respectievelijke algoritmes geselecteerd worden.

In [21] stellen de auteurs naast samenvattingen op basis van tweetvolume ook gepersona-



Figuur 3.2: Vergelijking tussen samenvattingen op basis van Twitter (blauw) en op basis van audiovisuele karakteristieken (rood). Gele labels geven gebeurtenissen aan die door beide algoritmes geselecteerd werden [22].

liseerde samenvattingen voor. In tegenstelling tot het zonet besproken artikel [22] maken zij geen gebruik van shotgrensdetectie-algoritmes maar verdelen ze de video in clips van 1 minuut. De tweets worden op basis van hun tijdsstempel gealigneerd met de video en zo toegekend aan een clip. De tweets in 1 interval van 1 minuut worden beschouwd als 1 document en vervolgens geïndexeerd met een TF-IDF gewichtenalgoritme. De gebruiker kan vervolgens op basis van een sleutelwoord een gepersonaliseerde samenvatting laten genereren. Finaal werden de 2 soorten samenvattingen met elkaar vergeleken door middel van een gebruikersstudie waarbij de 13 kandidaten een lichte voorkeur uitspraken voor gepersonaliseerde samenvattingen van 5 minuten ten nadele van samenvattingen op basis van tweetvolume.

3.2 Hoogtepuntextractie uit sportvideo's op basis van audio-, video- en webtekstanalyse

Om videoarchieven en videoclips van sportwedstrijden efficiënter toegankelijk te maken, wordt er vaak gebruik gemaakt van de audio- en videokarakteristieken van de videobeelden tijdens welbepaalde gebeurtenissen om de videoclip te annoteren. Deze analyses zijn bijzonder rekenintensief en geven vaak valse positieven. Daarom wordt er soms ook gebruik gemaakt van tekstuele samenvattingen die terug te vinden zijn op sportwebsites. Het gebruik van Twitter voor deze annotatie sluit hier nauw bij aan.

We beginnen in Sectie 3.2.1 met een overzicht van de literatuur die focust op het extraheren

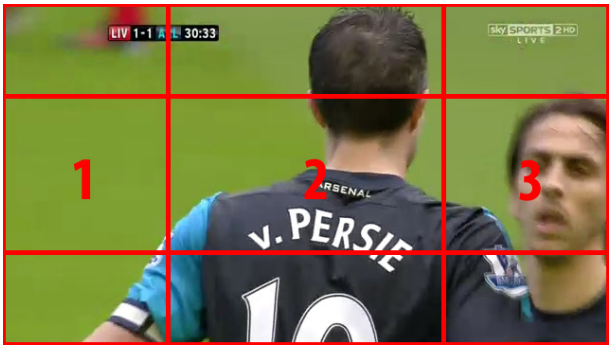
van hoogtepunten uit sportvideo's op basis van videoanalyse. In Sectie 3.2.2 bespreken we enkele alternatieve manieren van hoogtepuntextractie en vervolgens bespreken we in Sectie 3.2.3 enkele alternatieve toepassingen van audio- en videoanalyse. We eindigen in Sectie 3.2.4 met een overzicht van de literatuur die gebruik maakt van webtekstanalyse om hoogtepunten uit sportvideo's te extraheren. Tenzij anders vermeld, hebben alle besproken artikels betrekking op voetbalwedstrijden.

3.2.1 Hoogtepuntextractie op basis van videoanalyse

Om op automatische wijze hoogtepunten te kunnen extraheren, is het belangrijk de juiste technieken te gebruiken die het best hoogtepunten detecteren in een videosequenties. De meest voorkomende technieken zijn shotclassificatie en logodetectie.

In [23] wordt er voor het eerst een systeem voorgesteld dat op automatische wijze videosamenvattingen kan genereren van voetbalwedstrijden. Eerst wordt de dominante kleur van de videosequentie bepaald. Dit komt overeen met de groene grastint. Vervolgens worden de shotgrenzen gedetecteerd op basis van de dominante kleurratio en het verschil tussen de histogrammen in 2 opeenvolgende beelden. Daarna wordt shotclassificatie toegepast met behulp van een Naive-Bayes classificeerder. Een belangrijk begrip hieromtrent is de guldensnedecompositieregel afgebeeld in Figuur 3.3. Dit is een technniek gebruikt in tv-producties waarbij de protagonist afgebeeld wordt in de belangrijkste regio: regio 2. De dominante kleurratios van regio 1, 2 en 3 en het verschil in kleurratios tussen de regio's onderling worden gebruikt om beelden in te delen in één van de 4 belangrijkste shottypes: bovenaanzicht, middenaanzicht, close-up en buitenveldaanzicht (Figuur 3.4). Een doelpunt wordt opnieuw gedetecteerd op basis van tv-productieregels: indien een doelpunt gescoord wordt, zal er van een bovenaanzicht van de wedstrijd direct overgegaan worden op een close-up van enkele spelers, gevolgd door enkele herhalingen van het doelpunt. Deze techniek is één van de belangrijkste technieken om goals te detecteren in voetbalvideosequenties. Het voorgestelde algoritme kan in realtime werken als de video onderbemonsterd wordt naar 88x72 pixels.

In [24] wordt op dit idee verder gewerkt maar worden er een aantal extra eigenschappen van live-uitzendingen gebruikt. Bovenop shotclassificatie op basis van een dominante kleur wordt er gebruik gemaakt van herhalingsdetectie. Hierbij maken de auteurs gebruik van logodetectie. Logo's worden herkend op basis van een sjabloon van RGB-waarden specifiek voor dat type live-uitzending. Extractie van de dominante kleur wordt eveneens via een sjabloon van RGB en HSV-waarden uitgevoerd. Er wordt ook gebruik gemaakt van goalmonddetectie en scoreborddetectie. Doelpunten worden met 100% precisie en 100% recall gedetecteerd op een beperkte verzameling van 10 voetbalwedstrijden met 7 doelpunten.



Figuur 3.3: Illustratie van de guldensnedecompositieregel. Het beeld wordt zowel in de hoogte als in de breedte verdeeld volgens een 3:5:3-verhouding.

(a) Bovenaanzicht (b) Middenaanzicht (c) Close-up (d) Buitenveldaanzicht

Figuur 3.4: Illustratie van de 4 belangrijkste shottypes. Soms worden (c) en (d) als 1 type beschouwd omdat deze vanuit het oogpunt van de analyse dezelfde karakteristieken vertonen.

De aanpak voorgesteld in [25] ligt in dezelfde trend en maakt ook gebruik van dominante kleurratios en de guldensnedecompositieregel. De auteurs kiezen echter niet voor histogramanalyse om shotgrenzen te vinden maar beschrijven een aanpak die gebruik maakt van de Discrete Cosinus Transformatie (DCT). Verder introduceren de auteurs een Finite State Machine (FSM) om hoogtepunten te vinden in een videosequentie.

In [26] wordt een hierarchische aanpak voorgesteld om aan shotclassificatie te doen. Eerst wordt er logodetectie uitgevoerd om het onderscheid te maken tussen herhalingsshots en niet-herhalingsshots. Dit gebeurt op basis van sjablonen die eigen zijn aan de uitzending. Vervolgens worden de niet-herhalingsshots verder onderverdeeld in de 4 belangrijkste shottypes doormiddel van dominante kleurratio, gezichtsherkenning en de grootte van de objecten die in een beeld aanwezig zijn.

Alle voorgaande artikels maken gebruik van de dominante kleur van het voetbalveld om aan shotclassificatie te doen. Deze kleur wordt ofwel bepaald voor elke video apart, ofwel maakt men gebruik van een gemiddelde met een bepaalde marge. Bovendien is de menselijke kleurenperceptie verschillend met wat de histogrammen weergeven. De auteurs in [27] komen met een volledig nieuwe aanpak die gebruik maakt van een dominante set clusteringsalgortime [28]. Hierbij wordt niet op voorhand een dominante kleur bepaald, maar zal tijdens de analyse de dominante kleur van 4 regio's van de gulden snede bepaald worden (regio 1, 2, 3 en de regio onder 2). De gemiddelde kleur van de verschillende regio's en de grootte van deze clusters worden dan gebruikt om met behulp van een

SVM de shots te classificeren. De techniek geeft precisies van 88% tot 96% voor de 3 types shots, maar haalt slechts een recall van 62% voor middenaanzichten. De auteurs vergelijken hun methode met het algoritme voor shotclassificatie gebruikt in [23] en halen beduidend betere resultaten.

3.2.2 Alternatieve manieren van hoogtepuntextractie

In [29] ligt de focus minder op de videoanalyse maar maakt men voornamelijk gebruik van de bijhorende audiosequentie. Een uitbundige commentator of een uitbundig publiek duidt op een belangrijk moment in de wedstrijd. Dit wordt gecombineerd met de bewegingsvectoren aanwezig in de video en de gemiddelde lengte van een hoogtepunt.

In [30] maken de auteurs niet alleen gebruik van audio en video maar gaan ze ook op zoek naar heel specifieke markeringen. Zo zullen ze bijvoorbeeld op zoek gaan naar de palen van het doel in een voetbalsequentie of specifiek gaan vergelijken wanneer een bal gevangen wordt in een honkbalsequentie door middel van een reeks voorbeeldafbeeldingen. Parallel wordt de audio geanalyseerd. Een hoogtepunt is dus opnieuw een combinatie van 2 onafhankelijke analyses.

3.2.3 Alternatieve toepassingen van hoogtepuntextractie

In tegenstelling tot de voorgaande artikels waarbij hoogtepuntextractie voornamelijk het genereren van samenvattingen tot doel heeft, kunnen we hoogtepuntextractie ook gebruiken als middel om de bandbreedte- en batterijbeperkingen van mobiele apparaten te omzeilen [31, 32]. Het extraheren van hoogtepunten uit live-uitzendingen in realtime geeft de gebruiker van een mobiel toestel de mogelijkheid om de wedstrijd te volgen zonder een continue verbinding te hebben met een netwerk tijdens de wedstrijd. De hoogtepunten kunnen dan op aanvraag getoond worden. De meest interessante techniek hiervoor is om gebruik te maken van de reeds bestaande hoogtepuntfragmenten in een live-sportuitzending, namelijk herhalingen. De auteurs maken hiervoor gebruik van een logosjabloon dat eigen is aan de uitzending.

3.2.4 Webtekstanalyse

De verschillende aanpakken die gepresenteerd werden in de vorige secties maakten enkel gebruik van de video en in sommige gevallen van de bijhorende audio. Bovendien is het vereist dat de volledige videosequentie wordt verwerkt om de interessante gebeurtenissen te vinden. Op basis van de laagniveau-eigenschappen van de video (bijvoorbeeld: beeld, shot, type shot), proberen we hoogniveau-betekenisvolle informatie af te leiden uit de

video (bijvoorbeeld: “In de 87’ scoort Wayne Rooney de 1-0 voor Manchester United”). Verschillende auteurs proberen deze intensieve en moeilijke operatie te omzeilen door gebruik te maken van webteksten. Deze webteksten bevatten immers al de hoogniveau-betekenisvolle informatie waardoor de videoanalyse herleid wordt tot het selecteren van het juiste shot op basis van de absolute tijdsaanduiding die de webteksten bevatten.

In [33] presenteren de auteurs een raamwerk voor het detecteren van gebeurtenissen in webteksten van voetbalwedstrijden. Het systeem is in staat in realtime gebeurtenissen te detecteren. De auteurs focussen op gestructureerde webteksten. Om gebeurtenissen te classificeren maken ze gebruik van een verzameling sleutelwoorden per type gebeurtenis en de woorden voor en na het sleutelwoord. Er wordt vanuit gegaan dat de webteksten een minuut-seconden notatie gebruiken om de tijd van de gebeurtenis aan te duiden. Vervolgens maakt men gebruik van klokherkenning in de video om de gebeurtenissen beschreven in de webteksten terug te vinden. Merk op dat men slechts enkele beelden zal moeten analyseren om het juiste tijdstip terug te vinden. Op basis van shotdetectie en shotclassificatie kan het juiste hoogtepunt uit de videosequentie geselecteerd worden.

In [34] passen de auteurs dit raamwerk toe op basketbalsequenties. Dit is een moeilijkere sport om te analyseren aangezien het spel vaker stil ligt en er veel vaker afgewisseld tussen herhalingen en de echte wedstrijd. Voor het aligneren van de video en de webteksten maken ze dan ook gebruik van een HMM in plaats van een FSM. Gebaseerd op de technieken in [33, 34] presenteren dezelfde auteurs in [35] een generiek raamwerk dat in staat is geannoteerde hoogtepunten van sportwedstrijden op te vragen via een gebruikersinterface.

Een meer geavanceerde aanpak wordt in [36] gepresenteerd. De auteurs vertrekken van hetzelfde idee om webteksten te gebruiken om aan te geven wanneer een gebeurtenis begint. Dit is hier echter geen vereiste. De focust ligt op het extraheren van alle gerelateerde beelden. De belangrijkste videotechnieken die gebruikt worden zijn camerabeweging, shotclassificatie, middencirkel- en goaldetectie, klokherkenning en herhalingsdetectie. Om de tekst te analyseren wordt opnieuw gebruik gemaakt van sleutelwoorden. Om de tijdsinformatie optimaal te modelleren wordt gebruik gemaakt van Allen-algebra [37]. Deze techniek zal slechter presteren om gebeurtenissen te detecteren maar brengt een sterke verbetering om de randen van de gebeurtenis correct te detecteren. De precisie om het volledige fragment van een doelpunt te extraheren stijgt van 76% naar 92% tegenover de methode gepresenteerd in [33].

3.3 Machinaal leren

In Sectie 3.1 worden verschillende aanpakken gepresenteerd op basis van machinaal leren om een specifieke tweetclassificatie taak uit te voeren. Er worden 4 aanpakken vermeld:

een Support Vector Machine (SVM) [10, 11, 12, 15], Naive-Bayes [11, 13, 14], Gradient Boosted Decision Tree (GBDT) [6, 7] en MaxEnt [11].

Bij machinaal leren wordt over het algemeen verondersteld dat elk document kan voorgesteld worden door een vector $\vec{x} = (x_1, x_2, ..., x_n)$ en een geassocieerde klasse y. Indien elk element van de vector de aanwezigheid aanduidt van een term in een document spreken we van classificatie op basis van unigrams. Indien elk element van de vector de aanwezigheid van een opeenvolging van n termen voorstelt, spreken we van n-grams. Een andere manier om de vector op te bouwen is om elk element te laten samenvallen met een bepaalde karakteristiek van het document. Zo kunnen we bijvoorbeeld voor elk document het aantal uitroeptekens tellen of de lengte van het document in rekening brengen.

Op basis van een trainingsverzameling van documenten, hun bijhorende vectoren en klassen wordt er een model opgesteld. Het model is afhankelijk van het type classificeerder. Op basis van het model kan vervolgens voor elk nieuw document een klasse voorspeld worden.

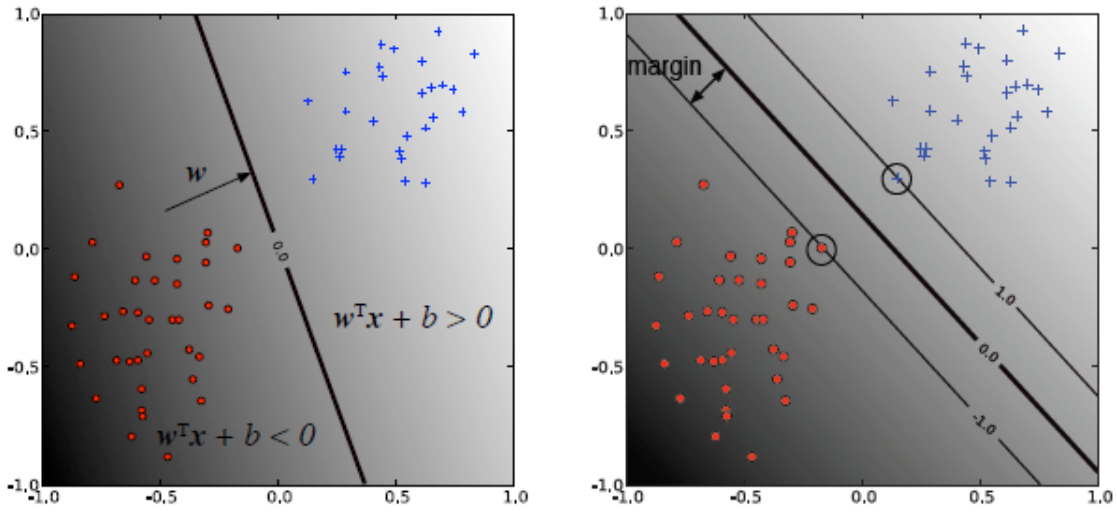
3.3.1 Naive-Bayes

De Naive-Bayes classificeerder vertrekt van het basisidee dat de aanwezigheid van een bepaalde term of een bepaalde karakteristiek onafhankelijk is van de aanwezigheid van een andere term of andere karakteristiek. Het is uiteraard vanzelfsprekend dat dit in werkelijkheid niet zo is. In een document dat over de specificaties van een auto gaat is de kans vrij groot dat de termen 'motor' en 'diesel' zich beiden in het document bevinden. De kans dat een document d tot de klasse c behoort is evenredig met volgende formule:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

Waarbij $\vec{t}$ een binaire vector voorstelt met n_d elementen en elk element t_k de aanwezigheid van een term in het document voorstelt. Het document d wordt geclassificeerd in de klasse c waarvoor $P(c|d)$ maximaal is [38].

Het grote nadeel van het Naive-Bayes model is dat de vector verondersteld wordt te bestaan uit nominale waarden en bijgevolg elke mogelijke term zich in de trainingsverzameling moet bevinden. Dit vormt een probleem voor numerieke waarden. Wanneer de documentlengte een karakteristiek is, zal in principe elke mogelijke documentlengte moeten opgenomen zijn in de trainingsverzameling.



(a) (b)

Figuur 3.5: Illustratie van lineaire SVM in 2D [40]. Links wordt het concept beslissingsgrens gedemonstreerd (a). Rechts worden de support vectors omcirkeld (b).

3.3.2 MaxEnt

MaxEnt is de afkorting van Maximale Entropie. MaxEnt zal net als Naive-Bayes gebruik maken van statistische modellering om de klasse van een vector te bepalen, maar gaat, in tegenstelling tot Naive-Bayes, ervan uit dat de elementen van de vector wel degelijk statistisch afhankelijk zijn. Dit zorgt ervoor dat het trainen van de classificeerder veel langer duurt omdat de verschillende gewichten moeten bepaald worden via een iteratief proces. Hierbij wordt er gebruik gemaakt van de Maximum A Posteriori (MAP) schatting omdat niet alle kansen gekend zijn. We gaan hier niet verder op in. Meer informatie kan ondermeer gevonden worden in [39].

3.3.3 Support Vector Machines

Een techniek die wel in staat is numerieke waarden te interpreteren is de Support Vector Machine [40]. De vector $\vec{x_i}$, geassocieerd met een document en een klasse y_i, wordt beschouwd als een punt in een hyperdimensionale ruimte. In Figuur 3.5(a) wordt het 2-dimensionale geval afgebeeld. We beschouwen 2 klassen, de rode klasse en de blauwe klasse. Deze worden gescheiden door een beslissingsgrens met normaalvector $\vec{w}$. De basisformule van een lineaire SVM is van de volgende vorm:

$$f(x) = \mathbf{w}^T \mathbf{x} + b$$

De vector $\vec{w}$ wordt ook wel de gewichtenvector genoemd en b is de bias. De term $\mathbf{w}^T\mathbf{x}$ wordt ook wel de kern $K(w;x)$ genoemd. In een hyperdimensionale ruimte stelt $\vec{w}$ de normaalvector van een hypervlak voor en b de verschuiving ten opzichte van de oorsprong. In het eenvoudige geval in Figuur 3.5(a) kunnen we de klasse van een document bepalen door bovenstaande formule toe te passen op de geassocieerde vector $\vec{x}$. Indien $f(x)$ kleiner is dan 0, behoort het document tot de rode klasse, anders tot de blauwe klasse.

In Figuur 3.5(b) wordt het concept marge uitgebeeld. Bij het bepalen van de normaalvector $\vec{w}$ is het van belang dat de beslissingsgrens zo ver mogelijk verwijderd is van de dichtste vectoren van elke klasse. Bovendien moet de afstand tot elke klasse even groot zijn. De support vectors zijn de vectoren die het dichtst bij de beslissingsgrens voor hun klasse liggen. Het gebied tussen de beslissingsgrens en de support vectors noemt men de marge en bevat geen vectoren. Het is van belang dat de marge zo groot mogelijk is omdat de trainingsverzameling slechts een beperkte deelverzameling is van alle bestaande vectoren. Hoe groter de marge, hoe kleiner de kans dat een onbekende vector in een verkeerde klasse zal ingedeeld worden. Het maximaliseren van de geometrische marge $\frac{1}{||\mathbf{w}||}$ is equivalent aan het minimaliseren van $||\mathbf{w}||^2$. Mathematisch kan het berekenen van de optimale beslissingsgrens uitgedrukt worden als:

$$Minimaliseer(\mathbf{w}, b): \ \frac{1}{2}||\mathbf{w}||^2$$

$$Onderhevig\ aan: y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1\ , \quad i = 1...n$$

Hierbij stelt $\mathbf{x}_i$ een vector voor van de trainingsverzameling en $y_i \in \{-1, 1\}$ de bijhorende klasse. In de praktijk blijkt dat de data niet altijd gemakkelijk lineair te verdelen is. Indien we toelaten dat sommige vectoren toch binnen de marges vallen, kunnen we veel grotere en gunstigere marges bereiken. We voegen spelingsvariabelen ϵ_i toe en een kostparameter C die bepaalt in hoeverre vectoren binnen de marge mogen liggen. De formule wordt nu:

$$Minimaliseer(\mathbf{w}, b): \ \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\epsilon_i$$

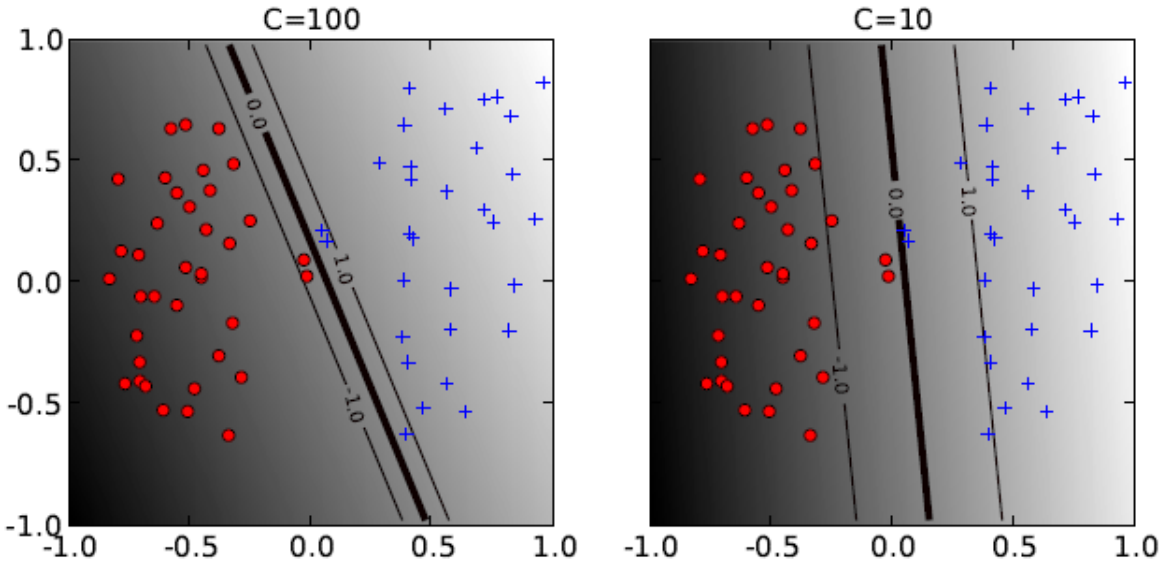
$$Onderhevig\ aan: y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \epsilon_i\ , \quad i = 1...n$$

$$\epsilon_i \ \geq \ 0$$

Deze marges noemen we zachte marges. De invloed van de kostparameter C wordt geïllustreerd in Figuur 3.6.

Soms blijkt een lineaire kern met zachte marges niet het gewenste resultaat te bereiken. Er bestaan 3 niet-lineaire basiskernen [41]:

- Polynomiaal: $K(\mathbf{x}_i; \mathbf{x}_j) = (\gamma\mathbf{x}_i^T\mathbf{x}_j + r)^d, \gamma > 0$



Figuur 3.6: Invloed van de kostparameter op de zachte marges bij een SVM met een lineaire kern [40].

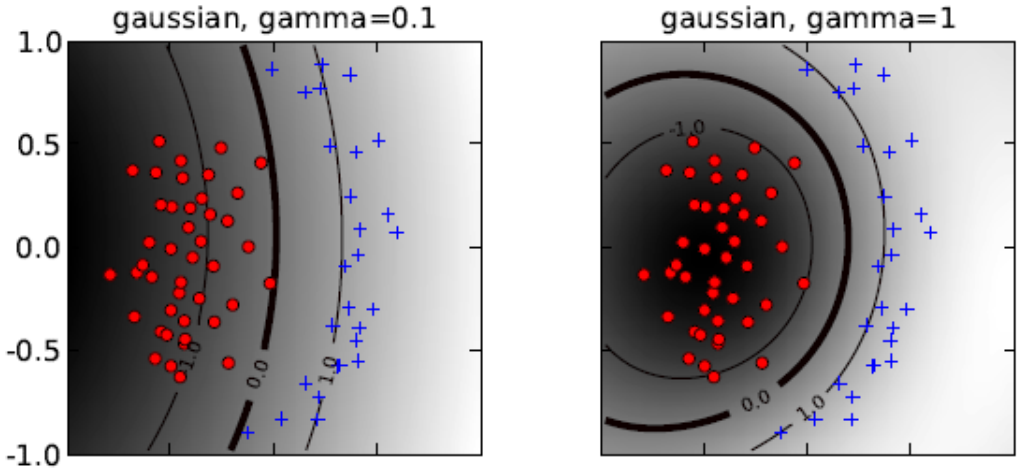
- Radiale basis functie (RBF): $K(\mathbf{x}_i; \mathbf{x}_j) = exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2), \gamma > 0$
- Sigmoid: $K(\mathbf{x}_i; \mathbf{x}_j) = tanh(\gamma \mathbf{x}_i^T \mathbf{x}_j + r)$

Hierbij zijn γ, r en d kernparameters. De keuze van de niet-lineaire kern hangt af van de toepassing maar ook van het aantal kernparameters. Over het algemeen wordt RBF verkozen voor zijn goede prestaties en beperkte verzameling parameters (C,γ). Figuur 3.7 toont de invloed van de parameter γ op de zachte marges.

Het mag duidelijk zijn dat een SVM wel kan omgaan met numerieke waarden die zich niet in de trainingsverzameling bevinden. De vectoren geassocieerd met een document zijn punten in een hyperruimte en de marges zorgen voor de afbakening van de verschillende klassen binnen de hyperruimte [40].

3.3.4 Gradient Boosted Decision Trees

Gradient Boosted Decision Trees (GBDT's) is een algoritme dat door meerdere iteraties de prestaties van een gekozen basisalgoritme probeert te verbeteren. Zoals de naam aanduidt, maken GBDT's gebruik van boomgebaseerde algoritmes. Meestal wordt als basisalgoritme Classification And Regression Trees (CART) gebruikt. In elke iteratie zal er een nieuwe boom gecreëerd worden die een bepaalde klasse van slecht geclassificeerde vectoren aanpakt. Het finale model bevat een ensemble van bomen die allemaal afgestemd zijn om een specifiek type vectoren te classificeren. Voor meer informatie wordt verwezen naar het werk van Friedman [42, 43].



Figuur 3.7: Invloed van γ op de zachte marges bij een SVM met een RBF kern [40].

3.3.5 Vergelijking van de verschillende technnieken

We merken op dat een eerste schifting van de technieken kan gebeuren op basis van het type vector. Als de vector slechts een eindig aantal waarden kan aannemen, kunnen alle technieken toegepast worden. Indien sommige elementen een oneindig bereik hebben of indien niet alle waarden zich in de trainingsverzameling bevinden, is het niet interessant om Naive-Bayes of MaxEnt te gebruiken omdat deze niet kunnen omgaan met numerieke waarden. Een Support Vector Machine kan dit wel. Bij GBDT's hangt dit af van het type basisalgoritme dat gebruikt wordt.

De verschillende technieken kunnen ook ingedeeld worden op basis van de populariteit en de ervaring met een algoritme bij het toepassen op tweets. MaxEnt werd slechts 1 maal toegepast op binaire vectoren. SVM's, GBDT's en Naive-Bayes werden meermaals succesvol toegepast.

De prestaties van de SVM en Naive-Bayes worden geëvalueerd in [44] en [45] op de Reuters-21578 collectie. In het eerste artikel wordt gebruik gemaakt van binaire vectoren en in het tweede artikel van vectoren op basis van de termfrequenties. In beide artikels concluderen de auteurs SVM veel beter presteert dan Naive-Bayes voor het classificeren van documenten.

Gradient Boosted Decision Trees (GBDT's) zijn een minder populair type classificeerder maar daarom niet minder goed. In [46] komt de auteur tot de conclusie dat GBDT's en de SVM evengoed presteren. De prestatie hangt vaak af van het type kern dat gebruikt wordt. Een SVM zal beter presteren indien men voldoende kennis heeft om een specifiek model te bouwen. GBDT's laat meer fouten toe dan een SVM. Bovendien heeft een GBDT een kleiner model en is het gemakkelijk decodeerbaar.

3.4 Besluit

De eerste stappen in het ontwikkelen van een systeem dat de hoogtepunten detecteert in een sportwedstrijd via Twitter zijn reeds gezet door verschillende auteurs [18, 19, 21, 22]. De meesten maken enkel nog gebruik van sleutelwoorden om gebeurtenissen te detecteren [18, 21, 22]. Slechts 1 poging werd ondernomen om in realtime hoogtepunten te detecteren tijdens sportwedstrijden in Twitter [18]. Meer algemene aanpakken maken echter al gebruik van machinaal leren om Twitterberichten te analyseren en te classificeren. We denken hierbij aan Support Vector Machines (SVM's) [10, 11, 12, 15], Naive-Bayes [11, 13, 14] en Gradient Boosted Decision Trees (GBDT's) [6, 7].

De literatuur die hoogtepunten detecteert in sportvideosequenties is een stuk uitgebreider. Heel wat techieken maken gebruik van enkel audio- en videoanalyse om belangrijke gebeurtenissen te detecteren [23, 24, 25, 26, 29, 30]. Deze technieken zijn rekenintensief en slagen er niet altijd in gebeurtenissen correct te detecteren. Een slimmere aanpak bestaat uit het gebruik van webteksten die een overzicht geven van de belangrijkste gebeurtenissen in een sportwedstrijd met hun respectievelijke tijdstippen [33, 34, 35, 36]. Dit laat ons toe accuraat de gebeurtenissen te lokaliseren en te annoteren. We merken echter op dat deze webteksten niet direct worden geüpdate en bovendien niet altijd beschikbaar zijn.

Het idee om Twitter te gebruiken om videosamenvattingen te maken werd al door enkele auteurs uitgewerkt in de literatuur [21, 22]. Het doel is hier echter om na afloop van de wedstrijd een samenvatting te kunnen genereren of videofragmenten te kunnen opvragen.

Tot op heden bestaat er nog geen systeem dat in realtime op basis van Twitter gebeurtenissen detecteert en de bijhorende fragmenten uit een videosequentie extraheert en annoteert. Zo een systeem zou echter de bandbreedte- en batterijbeperkingen van mobiele apparaten kunnen omzeilen. Dit idee werd reeds geopperd voor het ontstaan van Twitter maar kon in praktijk moeilijk uitgevoerd worden door de grote verwerkingsvereisten van audio- en videoanalyse [31, 32].

Hoofdstuk 4

Detectie van gebeurtenissen in Twitter tijdens voetbalwedstrijden

In dit hoofdstuk beschrijven we een detectiesysteem dat in staat is om in realtime belangrijke gebeurtenissen te detecteren in Twitter. Het doel is om enkele seconden na het gebeuren van een belangrijke gebeurtenis, de gebeurtenis te detecteren, te identificeren en eventueel de betrokken personen te vermelden. We focussen ons hierbij op de Engelse Premier League 2011-2012 die tot op heden de meest bekeken voetbalcompetitie ter wereld is. We beperken de identificatie tot de belangrijkste gebeurtenissen: een doelpunt, een fout (eventueel een gele of rode kaart), een strafschop, een wissel en het einde van de eerste helft en tweede helft. Het doel is om in realtime een gelijkaardige verslaggeving te kunnen verzorgen zoals sportwebsites als Soccerway [47], afgebeeld in Figuur 4.1.

Om de tweets te verzamelen maken we gebruik van de streaming API van Twitter [48]. De streaming API laat ons toe een continue verbinding op te zetten met Twitter waarbij we telkens alle tweets ontvangen met een zelfgekozen sleutelwoord of hashtag op voorwaarde dat dit niet meer is dan 1% van het totaal aantal verzonden tweets wereldwijd.

Voetbalsupporters maken gebruik van hashtags geassocieerd met hun ploeg om tweets te verzenden tijdens een wedstrijd. Tabel 4.1 geeft een overzicht van de belangrijkste hashtag per ploeg van de Engelse Premier League die supporters gebruiken om tweets te verzenden. Typisch aan deze competitie is dat de hashtag wordt gevormd door de

Figuur 4.1: Fragment van de live-verslaggeving op de website van Soccerway [47] tijdens de wedstrijd Aston Villa - Bolton Wanderers 21/04/12.

Tabel 4.1: Overzicht van de hashtags van de 20 voetbalploegen in de Engelse Premier League 2011-2012 [49].

Ploeg	**Hashtag**	**Ploeg**	**Hashtag**
Arsenal	#AFC	Newcastle United	#NUFC
Aston Villa	#AVFC	Norwich City	#NCFC
Blackburn Rovers	#BRFC	Queens Park Rangers	#QPR
Bolton Wanderers	#BWFC	Stoke City	#SCFC
Chelsea	#CFC	Sunderland (AFC)	#SAFC
Everton	#EFC	Swansea City	#SWANS
Fulham	#FFC	Tottenham Hotspurs	#THFC
Liverpool	#LFC	West Bromwich Albion	#WBAFC
Manchester City	#MCFC	Wigan Athletic	#WAFC
Manchester United	#MUFC	Wolverhampton Wanderers	#WWFC

initialen van de ploegnaam gevolgd door de letters FC (Engels: Football Club). We maken hiervan gebruik om tweets te verzamelen per ploeg per wedstrijd. Dit is echter geen vereiste. Het is perfect mogelijk om het voorgestelde detectiesysteem enkel toe te passen op een spelersnaam om de belangrijke gebeurtenissen geassocieerd met een speler te detecteren.

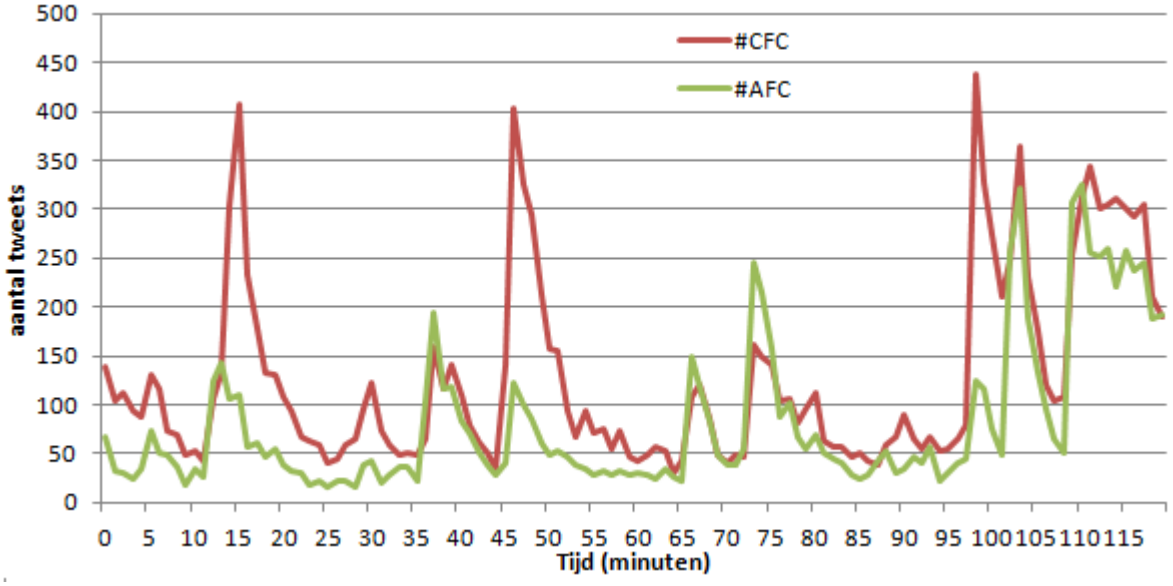
De aanpakken die werden voorgesteld in de literatuur maken voornamelijk gebruik van sleutelwoorden om gebeurtenissen te detecteren en te identificeren [18, 21, 22]. Dit zal echter niet altijd leiden tot robuuste detectie en identificatie. Een tweet die het woord 'goal' bevat kan ook verwijzen naar een gemist doelpunt. Bovendien is de hashtag die we gebruiken niet exclusief toegekend aan een ploeg.

We beginnen dit hoofdstuk in Sectie 4.1 met een algemene analyse van het tweetgedrag van supporters op Twitter tijdens voetbalwedstrijden. In Sectie 4.2 stellen we vervolgens het basis detectie- en identificatie-algoritme voor om gebeurtenissen te detecteren en te identificeren door gebruik te maken van één hashtag. We breiden dit algoritme uit in Sectie 4.3 door rekening te houden met het wedstrijdverloop en door gebruik te maken van de hashtags van beide ploegen. Finaal eindigen we in Sectie 4.4 met een besluit.

4.1 Analyse van het tweetgedrag tijdens voetbalwedstrijden

4.1.1 Pieken en dalen

Tijdens een voetbalwedstrijd zullen er fluctuaties zijn in het aantal tweets dat per tijdseenheid wordt verzonden. Niet elke fase is even interessant om verslag over uit te brengen. Figuur 4.2 toont het verloop van het aantal tweets dat per minuut werd verstuurd tijdens

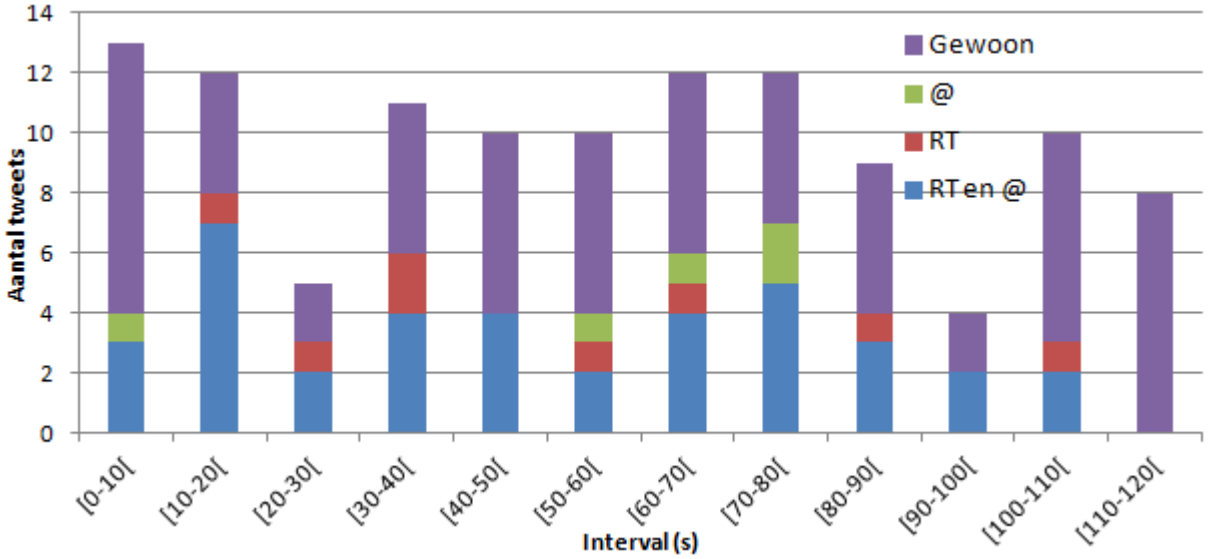


Figuur 4.2: Grafiek van het aantal berichtjes dat per minuut verzonden werd tijdens de wedstrijd Chelsea-Arsenal 29/10/2011 voor de respectievelijke hashtags #CFC en #AFC. De wedstrijd eindigde op 3-5.

de wedstrijd Chelsea - Arsenal 29/10/2011. Om verwarring te vermijden merken we op dat het tijdsverloop gerekend is vanaf het aanvangsuur van de wedstrijd tot het einde van de wedstrijd, inclusief de rustpauze tussen de 2 helften. De 46^{ste} minuut in de figuur komt overeen met minuut 45+1 in wedstrijdminuten en de 81^{ste} minuut komt ruwweg overeen met de 65^{ste} wedstrijdminuut. In wat volgt gebruiken we niet de wedstrijdminuten maar de minuten zoals gebruikt in Figuur 4.2 tenzij expliciet anders vermeld.

We zien duidelijk dat niet in elke minuut evenveel tweets werden verstuurd. De afgebeelde wedstrijd eindigde op 3-5. Een doelpunt is meestal de meest belangrijke gebeurtenis in de wedstrijd. Dit komt ook overeen met wat we zien in de figuur. De 8 hoogste pieken vallen samen met de 8 gescoorde doelpunten. Bovendien kan men ruwweg onderscheiden welke ploeg wanneer scoort. Bij 3 van de 8 pieken ligt de curve van Chelsea (#CFC) beduidend hoger dan de curve van Arsenal (#AFC). Wanneer Arsenal gescoord heeft, vallen de pieken van beide ploegen ongeveer samen. 2 pieken vallen hier ook toevallig samen met het einde van beide wedstrijdhelften.

Een stijging in het tweetvolume zal niet altijd met een voorgedefinieerde gebeurtenis gerelateerd zijn zoals een doelpunt of gele kaart maar kan ook veroorzaakt worden door een interessante fase voor één van beide ploegen. In minuut 30 zien we een duidelijke piek in de rode curve maar niet in de groene curve. In de desbetreffende minuut had Chelsea een zeer goede kans tot scoren maar miste. De supporters van Chelsea hadden hier duidelijk meer interesse in dan de supporters van Arsenal. (Opmerking: er kan geopperd worden dat de #AFC-curve een kleine piek vertoont in minuut 30 die evenredig is met het algemene tweetvolume. Het is echter zo dat een deel van de tweets zowel de



Figuur 4.3: Illustratie van de samenstelling van de 81^{ste} en 82^{ste} minuut van de wedstrijd Chelsea - Arsenal. Dit is een dalmoment.

hashtag #AFC als #CFC vertoont, vandaar deze minieme stijging.)

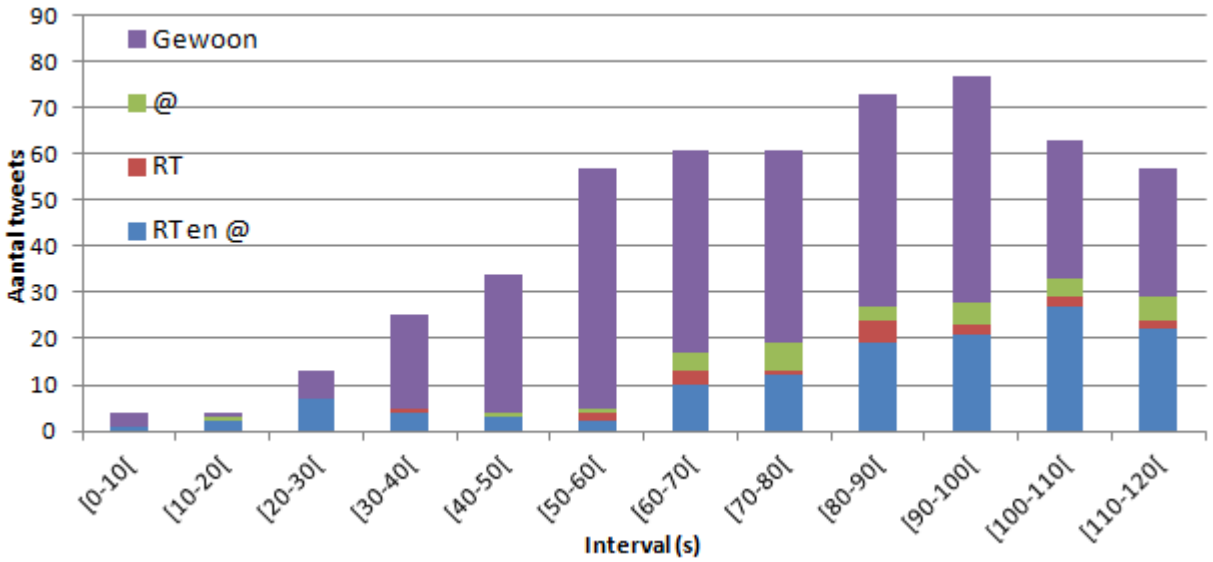
Een wedstrijd op Twitter is bijgevolg een opeenvolging van pieken en dalen waarbij een piek een belangrijke gebeurtenis aanduidt en een dal een tijdspanne aanduidt tussen 2 pieken waarin weinig of niets interessant is gebeurd.

4.1.2 Pieken versus dalen

Doordat in een dalmoment niets interessants is gebeurd, verwachten we dat het tweetgedrag van gebruikers verschillend zal zijn dan wanneer er een doelpunt werd gescoord. We onderscheiden 4 soorten tweets:

- RT: Retweets
- @-symbool: Antwoorden op tweets of vermeldingen van gebruikers
- RT en @-symbool: Een retweet van een antwoord
- Gewoon: Tweets die noch een retweet, noch een antwoord zijn, m.a.w. alle overige en gewone tweets

In Figuur 4.3 wordt een dalmoment afgebeeld waarbij de samenstelling bekeken wordt in intervallen van 10 seconden. De samenstelling van een interval van 10 seconden is telkens anders, maar over het algemeen kunnen we stellen dat tijdens een dalmoment er ongeveer evenveel 'gewone' tweets zijn als tweets met een @-symbool of retweets. Het gemiddeld aantal tweets is hier gelijk aan 9,67 per 10 seconden waarvan gemiddeld 5,42 tweets of 56% gewone tweets zijn. Dit is in contrast met de samenstelling tijdens piekmomenten.



Figuur 4.4: Illustratie van de samenstelling van de 45^{ste} en 46^{ste} minuut van de wedstrijd Chelsea - Arsenal. Dit is een piekmoment.

Tijdens een gebeurtenis kunnen we 4 verschillende fasen onderscheiden:

1. **Stilte voor de storm:** Deze fase wordt gekenmerkt door een algemeen tweetvolume dat lager is dan tijdens dalmomenten. Gebruikers nemen de belangrijke gebeurtenis waar en houden de vinger aan het toetsenbord.

2. **Om ter snelst:** Deze fase wordt gekenmerkt door een explosie van het aantal gewone tweets terwijl het aantal niet-gewone tweets laag blijft. De tweets in deze fase komen voornamelijk van gebruikers die als eerste de gebeurtenis willen rapporteren.

3. **Lees en deel:** In deze fase lezen de gebruikers de tweets geproduceerd tijdens vorige fase en zullen deze vervolgens retweeten, beantwoorden of becommentariëren. Deze fase wordt gekenmerkt door een sterke stijging in het aantal niet-gewone tweets.

4. **Herstel:** De gebeurtenis is over zijn hoogtepunt heen en het evenwicht gewone/niet-gewone tweets heeft zich terug hersteld. Het tweetvolume zal stilaan weer in een dalmoment terechtkomen.

Dit wordt geïllustreerd in Figuur 4.4 waar de overgang tussen een dal en een piek wordt weergegeven tijdens de 45^{ste} en 46^{ste} minuut van de wedstrijd Chelsea - Arsenal. In het interval [0-20[is het tweetvolume lager dan tijdens een dalmoment (fase 1). Na ongeveer 25 seconden in de 45^{ste} minuut wordt voor het eerst een doelpunt gerapporteerd. Vanaf het interval [30-40[vindt een ware explosie van het tweetvolume plaats. Het aantal gewone tweets neemt sterk toe in het interval [30-60[terwijl het aantal niet-gewone tweets gelijk blijft (fase 2). In het interval [60-80[stagneert het gewone tweetvolume maar vindt er een explosie plaats van het aantal niet-gewone tweets (fase 3). Het niet-gewone tweetvolume

stijgt gestaag tot een hoogtepunt in het interval [100-110[waarna het weer afneemt. Het hoogtepunt van het aantal gewone tweets werd reeds bereikt in het interval [50-60[. Een sterke afname vindt pas plaats in het interval [100-120[(fase 4). We merken op dat de samenstelling van de verschillende intervallen al snel evolueert naar het evenwicht dat we terugvinden bij dalmomenten.

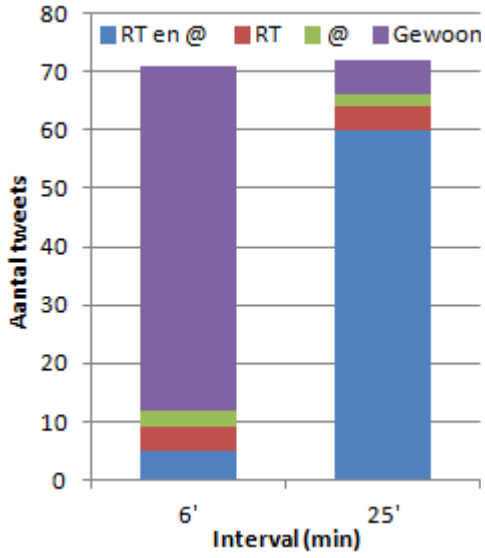
We kunnen concluderen dat het tweetvolume en de verhouding gewone/niet-gewone tweets een heel goede indicatie geven of we ons in een piek of een dal bevinden. We moeten echter enkele kanttekeningen plaatsen. Allereerst merken we op dat, doordat het totaal aantal verstuurde tweets sterk zal verschillen van wedstrijd tot wedstrijd, pieken en dalen enkel goed kunnen geïdentificeerd worden na een wedstrijd. Een tweede kanttekening heeft betrekking op de verhouding gewone/niet-gewone tweets. Niettegenstaande dat deze verhouding toelaat om de fase van een gebeurtenis gemakkelijk te bepalen, mogen we deze echter niet blindelings toepassen. Kijken we namelijk naar het interval [110-120[in Figuur 4.3, dan zien we dat dit interval enkel maar gewone tweets bevat.

Het idee dat het tweetvolume en het aantal retweets in een interval goede indicaties vormen om een gebeurtenis te detecteren tijdens een sportwedstrijd, werd reeds succesvol toegepast [19]. De auteurs berekenden echter de verhouding gewone/niet-gewone tweets op een interval van 2 minuten. Hierdoor wordt de 'om ter eerst'-fase die slechts 30 seconden duurt gemengd met intervallen die veel meer retweets bevatten. Bovendien werd de analyse pas achteraf uitgevoerd wanneer men de verhouding tussen de tweetvolumes van pieken en dalen kende.

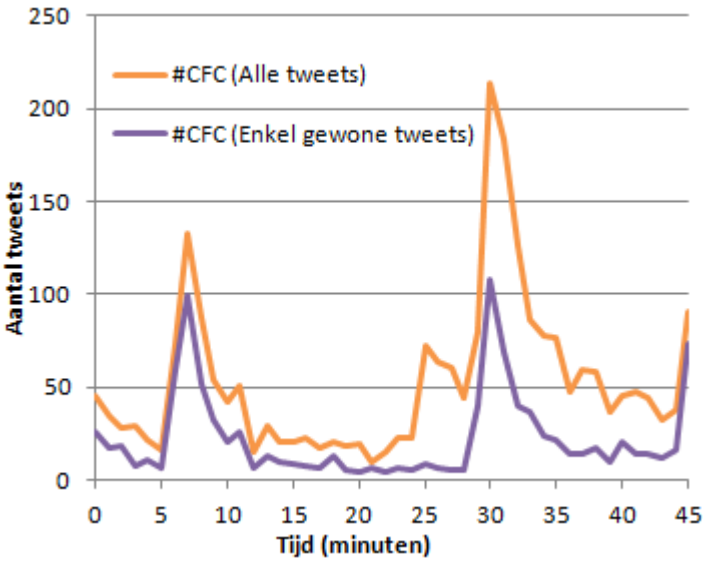
In algemene literatuur wordt gesteld dat het aantal tweets met @-symbolen daalt aan het begin van een gebeurtenis maar dat er niet direct een verband is tussen het aantal tweets met @-symbolen en het einde van de gebeurtenis [9]. In sportwedstrijden is het eerder zo dat een daling van het aantal retweets ten opzichte van het aantal gewone tweets het einde van een belangrijke gebeurtenis aangeeft. Dit kan verklaard worden doordat een belangrijke gebeurtenis in een sport slechts in een fractie van een seconde plaatsvindt waardoor het begin van de belangrijke gebeurtenis eigenlijk ook het einde is. Tweets met enkel @-symbolen komen veel minder voor.

4.1.3 Schaduwgebeurtenissen

Pieken in het tweetvolume geven aan wanneer er iets interessant gebeurt in een wedstrijd. Dat geldt echter niet voor elke piek. In sommige gevallen wordt een piek gecreëerd door een vertraagde reactie van de gebruikers of door een samenloop van omstandigheden. Een piek veroorzaakt door een gebeurtenis waarvoor er reeds in het verleden een piek was noemen we een schaduwgebeurtenis. De piek is immers een schaduw van een voorgaande piek die gecreëerd werd door dezelfde gebeurtenis.



Figuur 4.5: Samenstelling van de minuten 6 en 25 tijdens Chelsea-Wolverhampton 26/11/11.



Figuur 4.6: Verloop eerste helft Chelsea-Wolverhampton 26/11/11. 6' = 1-0, 25' = schaduwgebeurtenis, 29' = 2-0.

In Figuur 4.6 worden het aantal tweets per minuut tijdens de eerste helft van de wedstrijd Chelsea - Wolverhampton 26/11/11 afgebeeld. We kunnen 3 pieken onderscheiden in de oranje curve (alle tweets), namelijk in minuut 7, 25 en 30. In minuut 6 en 29 worden respectievelijk het eerste en tweede doelpunt gescoord. Wanneer we de samenstelling van minuut 25 bekijken merken we dat 52 van 60 retweets uit de categorie 'RT en @' allemaal dezelfde retweet zijn:

> "RT @chelseafc: GOAL: 1-0 to Chelsea and it's the captain John Terry who converts Juan Mata's corner to give the Blues the lead #CFC (SL)"

Figuur 4.5 toont de samenstelling van minuut 6 en 25. In minuut 6 zien we een duidelijke overheersing van gewone tweets terwijl in minuut 25 retweets overheersen. Deze schaduwgebeurtenis wordt dus gekenmerkt door een overheersing van retweets. Wanneer we alle retweets en tweets met @-symbolen verwijderen krijgen we een veel egalere curve. In Figuur 4.6 stelt de paarse curve het tweetvolume van enkel gewone tweets voor. Er blijven nu slechts 2 grote pieken over, namelijk diegenen die de 2 doelpunten voorstellen. De schaduwgebeurtenis is verdwenen.

De term 'schaduw' werd voor het eerst gebruikt in [22]. Hierbij werd het fenomeen beschreven dat bij een laat doelpunt het gemiddelde tweetvolume steeg ondermeer veroorzaakt door retweets en antwoorden ten gevolge van de euforie die gepaard ging met het doelpunt. Hierdoor werden pieken gecreëerd na het late doelpunt die hoger waren dan pieken van voor het late doelpunt. Als men er vanuit ging dat de 10 hoogste pieken overeen kwamen met de 10 belangrijkste gebeurtenissen, dan werden pieken geselecteerd die door toeval werden gecreëerd.

Tabel 4.2: De eerste tweets verzonden na het vallen van een doelpunt tijdens de wedstrijd Chelsea - Arsenal van 29/10/2011.

Tijdstip	**Gebruiker**	**Tweet**
45:31	TomYMASCFC	CAPTAIN JOHN TERRY! FUCK YES! #cfc
45:32	octoriodior	Goaaal!!!!! JT #CFC
45:32	bronzia88	CAPTAIN JT !!!!!!!!! #CFC
45:32	HusseinMarhoon	Terry. 2-1 #CFC
45:33	Tifosipazzi	And now it is Terry. Goal #CFC
45:33	RichardTarr	John Terry!!!!!!!!!Who else, seriously! #CFC
45:34	ChelseaActivity	Corner to Chelsea as we approach 45' mins. Terry converts!!!!!! 2-1! #CFC
45:34	MGehad7	2-1 Chelsea .. John Terry #cfc
45:34	alexeykravtsov	2-1 #cfc

4.1.4 Piekmomenten in detail

Naast het detecteren van gebeurtenissen willen we ze ook kunnen identificeren. Elk type gebeurtenis heeft zijn specifieke karakteristieken. In de volgende deelsecties bespreken we de gebeurtenissen die we willen kunnen detecteren met ons systeem en hun karakteristieken.

4.1.4.1 Het doelpunt

Doelpunten zijn de belangrijkste gebeurtenissen tijdens een voetbalwedstrijd en bepalen de winnaar van de wedstrijd. Dit zijn dan ook de gebeurtenissen waarover de gebruikers het meest tweeten. Aangezien we een realtime systeem willen bouwen, is het belangrijk dat we weten hoe de eerste tweets na het scoren van een doelpunt er uit zien. Tabel 4.2 bevat een fragment van de eerste tweets die verzonden werden na het scoren van een doelpunt tijdens de wedstrijd Chelsea - Arsenal van 29/10/2011.

Inhoudelijk valt meteen op dat de supporters de naam van de doelpuntenmaker tweeten en slechts 2 maal een sleutelwoord gebruiken. Een tweede kenmerk is het intensief gebruik van sentimentkenmerken zoals repetitie van letters, hoofdletters en uitroeptekens. Een derde kenmerk is de lengte van het bericht. Over het algemeen zijn deze berichten kort.

4.1.4.2 De strafschop

Een tweede gebeurtenis die voor heel wat opwinding zorgt is de strafschop omdat de kans vrij groot is dat er een doelpunt uit zal volgen. Een selectie van tweets net na het toekennen van de strafschop is te vinden in Tabel 4.3. Net als bij een doelpunt wordt er gebruik gemaakt van herhalingen en hoofdletters, maar in mindere mate. Er wordt echter geen gebruik gemaakt van spelersnamen zoals bij een doelpunt, maar voornamelijk van

Tabel 4.3: De eerste tweets verzonden na het toekennen van een strafschop tijdens Stoke City - Newcastle United 31/10/11.

Tijdstip	Gebruiker	Tweet
01:40:37	Ogwani	Penalty to Newcastle. What the F was that for? #SCFC #NUFC
01:40:37	billyjoelismint	lol! #nufc
01:40:37	mattys123	YES, PENALTY #NUFC
01:40:38	mickhedley10	Yes!!!!!!!!!!!!!!!!!! #nufc
01:40:38	ryan__marshall	The ref evens out his earlier poor decision!! #NUFC
01:40:39	LewiFive0	How the fuck can Pulis have the cheek to moan about us using the towel on throws!!!!!! #NUFC
01:40:39	McStay7	Pen to #nufc Haha! Brilliant! #MNF

Tabel 4.4: De eerste tweets verzonden na het toekennen van een rode kaart aan V. Kompany tijdens Manchester City - Wolverhampton Wanderers 29/10/11.

Tijdstip	Gebruiker	Tweet
01:31:36	jaymotty	Kompany sent off #mcfc
01:31:38	doggedtim	Huzzah - red for Kompany. Come on Wolves #mcfc #wolves
01:31:39	ryscaKP	Vincent Kompany Red Card :" #MCFC
01:31:40	ericeptional	Straight red for Vincent Kompany #mcfc
01:31:41	WeFollowFutbol	74' PENALTY Wolves! Kompany sent off and City down to 10. #epl #mcfc #wolves
01:31:45	O_I_T_H	I dont even mind the fact that he's in my fantasy football team #wolves #wwfc #mcfc
01:31:45	Tom_456	noooooooo kompany :(#mcfc
01:31:47	premierleague	RED CARD: Kompany 74 (MCI). #bpl #mcfc
01:31:48	THEAndyManning	Oh fuck city #MCFC

de sleutelwoorden 'penalty' en 'pen'. De tweets zijn ook langer. We merken verder op dat dit uiteraard slechts een fragment is en dat niet onmiddellijk veralgemeend mag worden.

4.1.4.3 De fout

Wanneer een speler een fout begaat op een andere speler kan een vrije trap of strafschop toegekend worden en eventueel een gele of rode kaart. Of een fout tot een piek in het tweetvolume leidt, zal sterk afhangen van het type fout en de gevolgen ervan. Een speler die een andere speler brutaal afstopt en hiervoor een rode kaart krijgt zal bijna altijd een piek genereren terwijl een lichte fout zonder kaart onopgemerkt zal blijven. Een fragment van een fout die een rode kaart tot gevolg heeft, wordt getoond in Tabel 4.4. Opnieuw merken we een tweetgedrag op dat verschillend is van de vorige gebeurtenissen. Alhoewel het gebruik van hoofdletters en herhalingen sterk gedaald is, is het niet volledig afwezig. We merken wel een terugkeer van het gebruik van spelersnamen. Ook wordt er gebruik gemaakt van sleutelwoorden zoals 'red card' en 'sent off'.

Tabel 4.5: Selectie van tweets verzonden na de wissel van Ramires voor Lukaku tijdens Chelsea - Arsenal 29/10/11.

Tijdstip	**Gebruiker**	**Tweet**
01:29:33	ChelseaActivity	Lukaku is coming on. Ramires is off. #CFC
01:29:34	chelseafc	Lukaku on for Ramires - 72 mins gone. #CFC(SL)
01:29:47	Naziho17	Ramires Off Lukaku On #CFC
01:29:47	kristi_eddie	Lukaku, do something for we nah. #cfc
01:29:48	AntoHo11and	Lukaku!! #cfc
01:29:49	BlueFromSpain	Ramires OFF, Lukaku ON. Super-offensive sub! #CFC
01:29:49	Apat_05	Come On Lukaku, do the best #CFC

Fouten zorgen echter ook voor controverse. Zo zal bijvoorbeeld bij een fout zonder kaart wel eens de uitspraak 'That should have been a yellow card' gedaan worden. Dit zorgt voor een extra moeilijkheid indien er op deze manier pieken worden gecreëerd.

4.1.4.4 De wissel

Tijdens een wedstrijd heeft elk team het recht om 3 spelers te wisselen. Net zoals bij fouten hangt veel af van de interesse van de supporters voor de wissel. Een wissel zal over het algemeen enkel een piek veroorzaken voor het team dat de spelers wisselt. Een selectie van tweets is weergegeven in Tabel 4.5. Een wissel is een gebeurtenis die typisch op een zeer neutrale wijze wordt gerapporteerd door gebruikers. De meeste tweets bevatten een vaste structuur van 2 spelersnamen en enkele sleutelwoorden zoals 'coming on for' of 'on/off'. Dit komt enerzijds doordat een wissel een minder spannende gebeurtenis is en anderzijds omdat deze gebeurtenis minder plots is. Gebruikers hebben de tijd om het bericht te typen omdat het even duurt totdat de vierde scheidsrechter zijn bordje heeft bovengehaald om de wissel aan te kondigen.

4.1.4.5 Het einde van de eerste en tweede helft

Het einde van de eerste en tweede helft verschillen op 2 vlakken van alle andere gebeurtenissen. Ten eerste zijn het vaste momenten in een wedstrijd en kunnen ze dus op voorhand verwacht worden. Ten tweede worden ze vaak enkel gerapporteerd door Twittergebruikers die accuraat verslag willen brengen van de wedstrijd. Een selectie van tweets is te vinden in Tabel 4.6. Alhoewel de meeste tweets in de selectie een score bevatten, zullen ook veel tweets commentaar bevatten op de afgelopen helft zoals bijvoorbeeld in de laatste tweet. Naast een score bevat een groot aandeel tweets ook verwijzigingen naar de afgelopen helft zoals 'half time' of 'FT'. Dit maakt deze tweets gemakkelijk te onderscheiden van andere tweets. Bovendien kan zo een gebeurtenis slechts eenmaal voorkomen.

Tabel 4.6: Selectie van tweets verzonden na het einde van de eerste helft (boven) en tweede helft (onder) tijdens Manchester City - Wolverhampton Wanderers 29/10/11.

Tijdstip	**Gebruiker**	**Tweet**
00:48:19	MCFC	Half time and it's #MCFC 0-0 #Wolves. Huge chorus of boos directed at the officials as they make their way off.
00:48:22	HusseinMarhoon	HT: Manchester City 0 - 0 Wolverhampton Wanderers. #EPL #MCFC #Wolves
01:54:05	bibinaditya	ft: man city 3-1 wolverhamton, nice play guys! #MCFC
01:54:10	EdenDassidy	FT #MCFC 3-1 #Wolves(Dzeko,Kolarov,A.Johnson-Hunt) Great football from city despite having a man sent off #EPL
01:54:15	slats96	Result nervy 3-1 win but take the 3 points #MCFC
01:54:16	araijaguar	Man City 3 - 1 Wolves #MCFC
01:54:57	65Rickz	Ugly win. We'll take it. #stillinfirst #mcfc

4.1.5 Besluit

We kunnen concluderen dat het tweetvolume sterk gecorreleerd is met de belangrijke gebeurtenissen in een voetbalwedstrijd. Niet elke piek verwijst echter naar een belangrijke gebeurtenis. Door retweets en tweets met een @-symbool te verwijderen, vermijden we schaduwgebeurtenissen en zullen bovendien echte gebeurtenissen beter tot uiting komen. Om sommige gebeurtenissen in realtime te detecteren zal het niet volstaan om enkel met sleutelwoorden te werken. De eerste tweets bevatten vaak enkel een spelersnaam zoals bijvoorbeeld bij een doelpunt. Om het onderscheid tussen de verschillende gebeurtenissen te maken kunnen we ondermeer gebruik maken van sleutelwoorden, spelersnamen, hoofdlettergebruik en herhalingen.

4.2 Een algoritme voor gebeurtenisdetectie en -identificatie

Het algoritme dat we zullen voorstellen in deze sectie heeft tot doel om in realtime verslaggeving te doen van een voetbalwedstrijd op basis van een door de gebruiker gekozen hashtag. We kiezen ervoor om ons terug te focussen op dezelfde 6 belangrijke gebeurtenissen in voetbalwedstrijden: het doelpunt, de strafschop, de fout, de wissel en het einde van de eerste en tweede helft.

Voor de ontwikkeling van het algoritme zijn we vertrokken van het enige bestaande werk in verband met de realtimeanalyse van sportwedstrijden. De auteurs in [18] stellen een algoritme voor dat de 4 belangrijkste gebeurtenissen in een American football-wedstrijd detecteert en identificeert. Het doel van deze auteurs is echter niet om aan verslaggeving

te doen, maar om een indicatie te geven aan supporters of adverteerders wanneer er iets belangrijk gebeurt in een wedstrijd.

Het algoritme kan op een semantische wijze onderverdeeld worden in 3 grote stappen:

1. **Filteren van de tweets:** In de eerste stap worden alle tweets verwijderd die geen meerwaarde zijn voor de volgende stappen of die fouten kunnen introduceren.

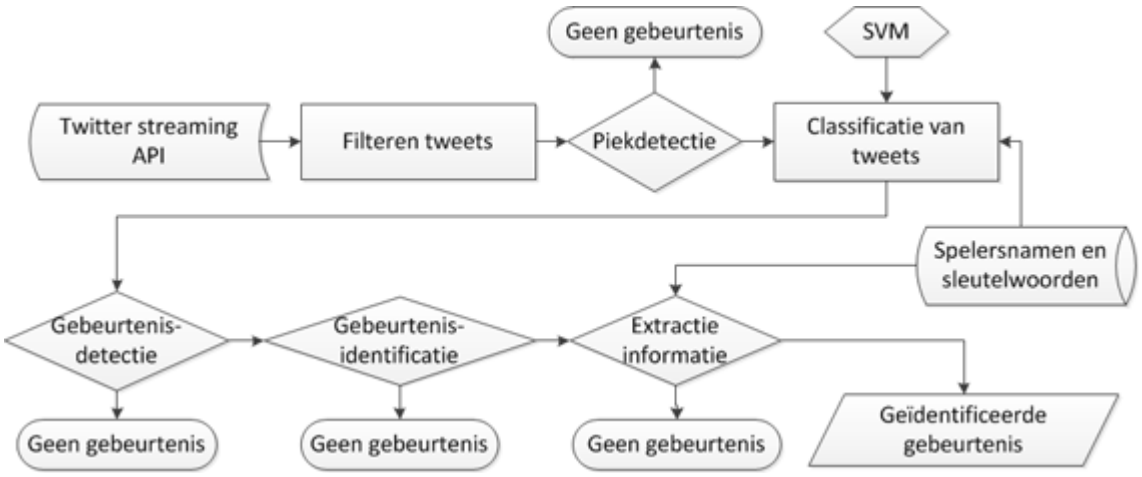
2. **Detectie:** In deze stap wordt gedetecteerd of er een gebeurtenis heeft plaatsgevonden onafhankelijk van het type gebeurtenis. Eerst vindt er een ruwe selectie plaats op basis van het tweetvolume om vervolgens een verfijnde analyse uit te voeren waarbij we kijken naar de samenstelling van de tweets op een bepaald moment in de tijd.

3. **Identificatie:** Nadat een gebeurtenis gedetecteerd is, vindt de identificatie plaats. Eerst wordt het type gebeurtenis geïdentificeerd om vervolgens gedetailleerde informatie te extraheren.

Een gedetailleerder overzicht van het algoritme kan gevonden worden in Figuur 4.7. Het algoritme wordt elke seconde uitgevoerd. Tweets worden in realtime ontvangen door het gebeurtenisdetectiesysteem via de Twitter streaming API. Eerst worden de tweets gefilterd. Vervolgens vindt de eerste fase van de detectie plaats: de piekdetectie. De piekdetectie is een ruwe selectie gebaseerd op het tweetvolume. Daarna worden alle tweets geclassificeerd met behulp van een Support Vector Machine (SVM). We maken hierbij gebruik van een verzameling sleutelwoorden en een verzameling spelersnamen zoals gevonden op de website van Soccerway [47]. De volgende stap is de gebeurtenisdetectie. Opnieuw wordt de beslissing gemaakt of we te maken hebben met een gebeurtenis of niet. Indien dit zo is, wordt er overgegaan tot de gebeurtenisidentificatie. Hier wordt de gebeurtenis geïdentificeerd als 1 van de 6 gebeurtenissen. Op basis van de geïdentificeerde gebeurtenis wordt in de laatste stap relevante informatie uit de tweets geëxtraheerd.

Voor de ontwikkeling van een robuust detectie- en identificatie-algoritme maken we gebruik van 11 verschillende wedstrijden uit de Engelse Premier League. Een uitgebreid overzicht wordt gegeven in Tabel A.1.

De indeling van Sectie 4.2 volgt de opbouw van het algortime zoals weergegeven in Figuur 4.7. In Sectie 4.2.1 bespreken we hoe de tweets gefilterd worden. De piekdetectie wordt besproken in Sectie 4.2.2. De classificatie van de tweets wordt besproken in Sectie 4.2.3. Vervolgens bespreken we in Sectie 4.2.4 de gebeurtenisdetectie en in Sectie 4.2.5 de gebeurtenisidentificatie. De extractie van interessante informatie wordt besproken in Sectie 4.2.6. We eindigen Sectie 4.2 met een besluit in Sectie 4.2.7.



Figuur 4.7: Schematische voorstelling van het voorgestelde algoritme om gebeurtenissen te detecteren in Twitter.

4.2.1 Filteren van de tweets

Om accuraat gebeurtenissen te kunnen detecteren is het belangrijk dat we enkel tweets verwerken die over de wedstrijd zelf gaan en die gebeurtenissen beschrijven op het moment dat ze plaatsvinden. We weten uit Sectie 4.1 dat retweets en tweets met een @-symbool aanleidingen kunnen geven tot schaduwgebeurtenissen en voornamelijk verstuurd worden op momenten dat er weinig te beleven valt of in de 'Lees en deel'-fase van een piekmoment. We verwijderen deze tweets op basis van de symbolen 'RT' en @.

Een tweede groep tweets die een correcte detectie kunnen verstoren zijn tweets met URL's. We identificeren 2 groepen tweets met URL's. De eerste groep zijn de spam tweets. De tweede groep zijn tweets die URL's bevatten naar live-uitzendingen van de wedstrijd op het internet of tweets van gebruikers die een belangrijke fase in de wedstrijd op een videowebsite zoals Youtube[1] hebben geplaatst. Beide groepen hebben gemeen dat ze nooit een gebeurtenis zullen beschrijven op het moment dat de gebeurtenis plaatsvindt. We verwijderen de tweets op basis van de protocolvermelding http:// en https://.

Na het verwijderen van oninteressante tweets pakken we de tweets zelf aan. We verwijderen alle vreemde tekens met uitzondering van het uitroepteken, de punt, het vraagteken, het koppelteken en de dubbele punt.

Een gereduceerde Twitterberichtenstroom gaat door naar de volgende stap.

4.2.2 Piekdetectie

Doordat we een systeem bouwen dat in realtime gebeurtenissen moet detecteren is het niet mogelijk om pieken te selecteren op basis van het absolute tweetvolume omdat we niet op

[1]http://www.youtube.com/

voorhand weten hoeveel tweets er zullen verstuurd worden. Dit hangt van verschillende factoren af zoals de ploeg, het tijdstip van de wedstrijd, de inzet van de wedstrijd en zelfs van welke doelpunt er gescoord wordt.

Om een sterke stijging te detecteren maken we naar analogie met [18] gebruik van een adaptief zoekvenster dat we elke seconde van de wedstrijd toepassen op de Twitterdata. Het zoekvenster heeft een grootte van 10, 20, 30 of 60 seconden. Om een stijging te detecteren splitsen we het zoekvenster in 2 gelijke helften en tellen we het aantal berichtjes in elke helft. Indien het aantal berichtjes in de tweede helft in verhouding tot het aantal berichtjes in de eerste helft groter is dan een grenswaarde α_{piek}, betekent dit een start van een piek. Dit wordt mathematisch uitgedrukt als volgt:

$$\frac{Aantal\ tweets\ in\ tweede\ helft\ van\ het\ zoekvenster}{Aantal\ tweets\ in\ eerste\ helft\ van\ het\ zoekvenster} > \alpha_{piek}$$

Het venster werkt op een adaptieve manier en zal starten met een venster van 10 seconden. Indien de verhouding kleiner dan of gelijk aan α_{piek} is, wordt het venster vergroot tot 20 seconden en vervolgens eventueel tot 30 en 60 seconden. Indien er geen piek wordt gedetecteerd binnen een zoekvenster van 60 seconden, wordt er verondersteld dat er geen belangrijke gebeurtenis heeft plaatsgevonden.

De meest optimale waarde voor α_{piek} werd experimenteel bepaald op 1,85 op basis van de ontwikkelingsverzameling. In de praktijk blijkt deze grenswaarde veel te gevoelig te zijn wanneer er zich een beperkt aantal tweets in het zoekvenster bevinden. Daarom voegen we zelf nog een tweede voorwaarde toe:

$$Aantal\ tweets\ tweede\ helft \geq \alpha_{min}$$

Hierbij is $\alpha_{min} = 10$. Indien er bij een zoekvenster van 10 seconden zich niet voldoende tweets in de tweede helft van het zoekvenster bevinden en er toch een gebeurtenis heeft plaatsgevonden, dan zal de gebeurtenis pas gedetecteerd worden bij een zoekvenster van 20, 30 of 60 seconden. We voorkomen echter het probleem dat bij wedstrijden waar er iets minder wordt getweet, een tweet gevolgd door 2 tweets als sterke stijging wordt aanzien.

Door gebruik te maken van de voorgestelde aanpak wordt het kaf van het koren gescheiden voordat we verder gaan met intensievere stappen. Het belangrijkste criterium om α_{piek} te bepalen is dat alle belangrijke gebeurtenissen door gaan naar de volgende stap en dat zo veel mogelijk niet-belangrijke gebeurtenissen worden tegengehouden. We merken op dat het niet vereist is dat op elke seconde van een gebeurtenis een piek moet gedetecteerd worden maar dat het eenmalig overschrijden van α_{piek} voldoende is om de gebeurtenis volledig te identificeren.

4.2.3 Classificatie van de tweets

Nadat een sterke stijging is gedetecteerd, willen we verifiëren of er wel degelijk een belangrijke gebeurtenis heeft plaatsgevonden. Hiervoor is een analyse van de tweets noodzakelijk. Uit Sectie 4.1.4 weten we dat gebeurtenissen door meer dan enkel sleutelwoorden gekarakteriseerd worden. Voor deze analyse beschouwen we enkel de tweets in de tweede helft van het zoekvenster aangezien deze tweets de stijging veroorzaakt hebben. We maken gebruik van machinaal leren om de tweets te classificeren.

In Sectie 3.3 werd reeds een introductie gegeven tot machinaal leren. Om tweets te kunnen classificeren door middel van machinaal leren moeten we 2 zaken bepalen. Een tweetvector $\overrightarrow{tweet} = (x_1, x_2, ..., x_n)$ die de representatie is van een tweet en een classificatiemodel dat ons toelaat voor elke tweet een klasse y te bepalen op basis van de tweetvector. Het type model kan pas gekozen worden als we de karakteristieken kennen van de tweetvector. We beginnen daarom met het opstellen van de tweetvector. Vervolgens bepalen we de classificatietechniek en bijgevolg ook het type model om daarna het classificatiemodel op te stellen.

4.2.3.1 Opstellen van de tweetvector

Binnen het domein van de natuurlijke taalverwerking maakt men al enkele decennia intensief gebruik van machinaal leren om documenten in klassen in te delen. Elk document kan voorgesteld worden als een binaire vector $\vec{x} = (x_1, x_2, ..., x_n)$ waarbij elke positie i overeenkomt met de aanwezigheid van een term t_i. Indien een document een term t_i bevat zal x_i de waarde 1 hebben, anders 0. Deze techniek staat beter betekend als het gebruik van unigrams (zie Sectie 3.3).

Het grote verschil tussen formele documenten en tweets is dat documenten de spellingsregels van de taal volgen terwijl op Twitter er geen vaste spellingsregels zijn. Dit zorgt voor een explosieve groei in het aantal mogelijke termen [50]. Bovendien worden verschillende talen door elkaar gebruikt. Het fenomeen wordt ook wel omschreven als de magerheid van data (Engels: Data Sparsity) [13]. Hoe meer elementen de vectoren bevatten, hoe trager de verwerking zal verlopen.

Een tweede mogelijkheid is het bepalen van een beperkte verzameling statistische kenmerken die eigen zijn aan de verschillende gebeurtenissen. We denken hierbij aan de lengte van een document of het aantal uitroeptekens in een document. De vector zal bijgevolg geen binaire vector meer zijn. Het gebruik van niet-binaire waarden stelt ook eisen aan de technniek die gebruikt wordt om documenten te classificeren.

Met betrekking tot Twitter en op basis van de literatuur [10, 12], kunnen we 4 grote groepen kenmerken onderscheiden:

- **Woordkenmerken:** Dit is het meest eenvoudige en meest gebruikte kenmerk. Een tweet bevat een term of een tweet bevat de term niet.

- **Meta-kenmerken:** Dit kenmerk vertolkt de functie van een woord: een werkwoord, een bijvoegelijk naamwoord, een zelfstandig naamwoord,... Dit staat beter bekend in het Engels als PoS (Part-of-Speech). Onder hetzelfde kenmerk kunnen ook andere functies gecreëerd worden zoals het vertolken van een negatieve of positieve functie van een term.

- **Syntactische kenmerken:** Syntactische kenmerken hebben betrekking op alles dat met de syntaxis van de tweet te maken heeft. Dit is onder andere het gebruik van leestekens, herhalingen en hoofdletters, en ook het gebruik van emoticons, hashtags en @-symbolen.

- **Statistische kenmerken:** Alle overige kenmerken die niet binair uitgedrukt kunnen worden. De belangrijkste voorbeelden hiervan zijn het aantal woorden in een tweet of het gemiddeld aantal karakters per term.

We merken op dat de eerste 3 kenmerken zowel binair als decimaal kunnen uitgedrukt worden. Statistische kenmerken daarentegen kunnen enkel decimaal uitgedrukt worden. Elk gebruikt kenmerk komt overeen met een element in de vector.

In Tabel 4.7 wordt een overzicht gegeven van de 7 klassen waarin we de tweets willen indelen. Dit zijn de 6 gebeurtenissen die we willen identificeren en een klasse voor tweets die geen voorgedefinieerde gebeurtenis beschrijven. In de kolommen wordt er telkens aangegeven voor welke kenmerken de aanwezigheid van dit kenmerk mogelijks bepalend kan zijn voor een klasse op basis van Sectie 4.1.4. Deze kenmerken vallen voornamelijk in de categorieën meta-kenmerken, syntactische kenmerken en statistische kenmerken. We maken gebruik van deze kenmerken om de tweetvector op te stellen.

Lengte tweet De eerste 2 elementen van onze vector zijn het aantal karakters van de tweet en het aantal termen in een tweet. Beiden worden geschaald naar een waarde tussen 0 en 1. Het aantal karakters kan nooit meer zijn dan 140 omdat de maximaal toegelaten lengte van een tweet 140 karakters is en het aantal woorden blijkt nooit meer dan 30 te zijn. Deze laatste waarde werd afgeleid op basis van de ontwikkelingsverzameling. De eerste versie van de tweetvector ziet er als volgt uit:

$$\overrightarrow{tweet} =< \#karakters, \#woorden >$$

Spelersnaam Bij vele gebeurtenissen worden ook spelersnamen vermeld. In [19] probeert de auteur tweets in te delen in 2 categorieën: tweets met spelersvermelding en tweets

Tabel 4.7: Overzicht van de kenmerken van een tweet die mogelijks bepalend zijn voor een klasse.

	Lengte tweet	spelersnaam	Hoofdletters en herhaling	Leestekens	Contextwoorden	Scorevermelding
Doelpunt	x	x	x	x	x	x
Strafschop	x	x	x	x	x	
Fout		x	x	x	x	
Wissel		x			x	
Einde 1^{ste} helft					x	x
Einde 2^{de} helft					x	x
Geen gebeurtenis						

zonder spelersvermelding. Slecht een deel van de voorgestelde kenmerken zijn toepasbaar: de vermelding van de voornaam, de vermelding van de achternaam en de initialen van de speler.

$$\overrightarrow{tweet} =< \#karakters, \#woorden, voornaam, achternaam, initialen >$$

Doordat we niet voor elke spelersnaam een apart element voorzien in de vector, maken we gebruik van meta-kenmerken in plaats van woordkenmerken. Het voordeel is dat de vectoren ploegonafhankelijk zijn en veel korter dan indien we voor elke speler een apart element voorzien. We maken gebruik van een databank van spelersnamen die samengesteld is uit de profielen die te vinden zijn op de website van Soccerway [47].

Omdat sommige initialen samenvallen met woorden uit de Engelse taal maken we gebruik van een stopwoordenlijst van WordNet [51]. Op deze manier worden de woorden niet als spelersinitialen geïnterpreteerd maar als een gewone term beschouwd.

Hoofdlettergebruik en herhalingen Voor supporters is het gebruik van woorden met hoofdletters of een herhaling van bepaalde letters een manier om uiting te geven aan hun vreugde of woede via Twitter [12]. We maken een onderscheid tussen herhalingen en hoofdlettergebruik bij spelersvermeldingen en bij andere termen. De reden is dat tijdens een doelpunt en een rode kaart meestal een spelersnaam wordt vermeld en tijdens een strafschop meestal niet. Als laatste kenmerk nemen we ook nog de procentuele graad van

hoofdlettergebruik in rekening:

$$\overrightarrow{tweet} =< \#karakters, \#woorden, voornaam, achternaam, initialen, hoofdlettergebruik, hoofdletters\_speler, hoofdletters\_woord, herhaling\_speler, herhaling\_woord >$$

Leestekens Op een zelfde manier als hoofdlettergebruik en herhalingen kunnen ook leestekens expressie geven aan een tweet. We beschouwen enkel een herhaling indien die gelijk is aan het 3 of meer keer opeenvolgend voorkomen van een leesteken. Als leestekens kiezen we voor de punctuatie, het vraagteken en het uitroepteken. Het gebruik van een enkelvoudig uitroepteken blijkt ook meer voor te komen tijdens belangrijke gebeurtenissen en wordt ook als kenmerk opgenomen. Andere leestekens blijken geen interessante functie te vervullen en werden reeds in de filterstap verwijderd.

$$\overrightarrow{tweet} =< \#karakters, \#woorden, voornaam, achternaam, initialen, hoofdlettergebruik, hoofdletters\_speler, hoofdletters\_woord, herhaling\_speler, herhaling\_woord herhaling\_uitroep, herhaling\_vraagteken, herhaling\_punt, aanwezigheid\_uitroep >$$

Contextwoorden Niet alle gebeurtenissen kunnen onderscheiden worden op basis van puur syntactische en statistische kenmerken. Het verschil tussen een tweet die een toegekende strafschop beschrijft en een doelpunt kan soms enkel bepaald worden op basis van een contextwoord: 'PENALTY!!!!' versus 'GOAL!!!!'. Fases zoals een wissel, een gewone fout en het einde van de eerste of tweede helft kunnen bijna enkel op basis van contextwoorden onderscheiden worden.

Doordat we het model niet te veel taalafhankelijk willen maken, maken we gebruik van semantisch gelijke onderwerpklassen [13]. Woorden die tot een zelfde onderwerp behoren, worden in 1 klasse opgedeeld. Als klassen beschouwen we de 6 belangrijkste gebeurtenissen die we willen classificeren, namelijk doelpunt, strafschop, fout, wissel, einde eerste helft en einde tweede helft. De woorden bepalen we door voor de 6 belangrijkste klassen de meest voorkomende contextwoorden uit de tweets te halen. In Tabel 4.8 wordt een overzicht gegeven van de contextwoorden per klasse. Voor elk van de 6 klassen zal een binair element toegevoegd worden die de aanwezigheid van een contextwoord uit die klasse aangeeft in een tweet. Dit zijn zogenaamde meta-kenmerken.

Net zoals we de herhaling van leestekens of letters van spelersnamen een interessant kenmerk vinden, vinden we ook de herhaling van letters in contextwoorden interessant. Hetzelfde geldt voor het hoofdlettergebruik. We noemen deze kenmerken *herhaling_context-*

Tabel 4.8: Overzicht van de contextwoorden die gebruikt worden voor elke klasse.

Klasse	**Contextwoord**
Doelpunt	goal, gol, gal, get in, score, scores, header, screamer, kick, assist, head, equalises, equalizes, own goal, OG
Strafschop	penalty, pen
Fout	yellow, red, booked, penalized, card, booking, foul, sent off
Wissel	on off, in out, coming on for, coming off for, comes of for, comes on for, in out, coming on for, coming off for, comes off for, comes on for, replaced, replaces, sub, substitution
Einde 1^{ste} helft	HT, half time, halftime
Einde 2^{de} helft	FT, full time, fulltime

woord en *hoofdletters_contextwoord*. De tweetvector ziet er nu als volgt uit:

$$\begin{aligned}\overrightarrow{tweet} =< &\#karakters, \#woorden, voornaam, achternaam, initialen, hoofdlettergebruik,\\ &hoofdletters\_speler, hoofdletters\_contextwoord, hoofdletters\_woord,\\ &herhaling\_speler, herhaling\_contextwoord, herhaling\_woord,\\ &herhaling\_uitroep, herhaling\_vraagteken, herhaling\_punt, aanwezigheid\_uitroep\\ &ctxt\_doelpunt, ctxt\_penalty, ctxt\_fout, ctxt\_wissel, ctxt\_einde1ste, ctxt\_einde2de >\end{aligned}$$

Scorevermelding Een speciaal geval is het gebruik van scores in tweets. Alhoewel een score ook als contextwoord van de klasse doelpunt kan aanzien worden, beschouwen we het toch als een apart kenmerk. In een scorevermelding zit een bepaald patroon wat verschillend is van een gewoon contextwoord.

De finale vector ziet er als volgt uit:

$$\begin{aligned}\overrightarrow{tweet} =< &\#karakters, \#woorden, voornaam, achternaam, initialen, hoofdlettergebruik,\\ &hoofdletters\_speler, hoofdletters\_contextwoord, hoofdletters\_woord,\\ &herhaling\_speler, herhaling\_contextwoord, herhaling\_woord,\\ &herhaling\_uitroep, herhaling\_vraagteken, herhaling\_punt, aanwezigheid\_uitroep\\ &ctxt\_doelpunt, ctxt\_penalty, ctxt\_fout, ctxt\_wissel, ctxt\_einde1ste, ctxt\_einde2de,\\ &score >\end{aligned}$$

4.2.3.2 Selectie van de classificatietechniek

Nadat de tweetvector bepaald is, kunnen we een classificatietechniek selecteren. In Sectie 3.3 werd reeds een studie gemaakt van de gebruikte technieken in de literatuur voor de classificatie van tweets in het algemeen. De meest voorkomende technieken waren Naive-Bayes, Support Vector Machines (SVM's) en Gradient Boosted Decision Trees (GBDT's).

Door het feit dat de tweetvector verschillende numerieke waarden bevat wordt het moeilijk om Naive-Bayes te gebruiken. Indien we de SVM en de GBDT vergelijken dan merken we 2 zaken op: GBDT's werden slechts door een beperkte groep auteurs gebruikt en de ondersteuning is beperkt. De bekende dataminingtool Weka [52] die we gebruiken voor het opstellen van het classificatiemodel, heeft nog geen werkende implementatie hiervoor. SVM's zijn wijdverspreid, kunnen met numerieke waarden werken en zijn in verschillende domeinen state-of-the-art. Onze keuze valt dus op Support Vector Machines (SVM's) als classificatietechniek.

4.2.3.3 Opstellen van het classificatiemodel

Een classificatiemodel wordt door de SVM gebruikt om ongekende tweets een klasse toe te kennen. Het classificatiemodel is afhankelijk van de gekozen classificatietechniek, een verzameling parameters die eigen zijn aan het model en een verzameling trainingsdata met bijhorende klasse. De gekozen classificatietechniek is een SVM en de trainingsdata een verzameling geannoteerde tweets.

Als trainingsdata van het classificatiemodel maken we gebruik van de 5 wedstrijden uit de trainingsverzameling weergegeven in Tabel A.2. In totaal werden 1324 tweets willekeurig maar evenredig geselecteerd over de 5 wedstrijden. Van deze 1324 tweets behoren 299 tweets tot de klasse 'doelpunt', 11 tot de klasse 'strafschop', 14 tot de klasse 'fout', 24 tot de klasse 'wissel', 16 tot de klasse 'einde eerste helft' en 16 tot de klasse 'einde tweede helft'. De overige 944 tweets behoren tot de klasse 'geen gebeurtenis'. De trainingsverzameling is duidelijk ongebalanceerd. We hebben daarom gebruik gemaakt van een gewichtenschema. Als de klasse 'geen gebeurtenis' gewicht 1 heeft, dan heeft de klasse 'doelpunt' gewicht 2 en de andere klassen hebben allemaal gewicht 5.

In Sectie 3.3.3 werden er 3 verschillende niet-lineaire SVM's geïntroduceerd. We hebben gekozen voor een SVM met een radiale basisfunctie wegens de goede algemene prestaties en het beperkt aantal parameters [40]. We moeten bijgevolg nog de waarde van 2 parameters bepalen: de parameter C (kost) en de parameter γ. Voor zowel C als γ hebben we gebruik gemaakt van de standaardwaarden: $C = 1$ en $\gamma = \frac{1}{Aantal\ kenmerken}$. In ons geval betekent dit dat γ gelijk is aan $\frac{1}{23}$. Door het wijzigen van deze 2 parameters konden we de prestatie van de classificatie op de trainingsverzameling verhogen maar dit had een negatief effectief op de classificatie van tweets uit wedstrijden die niet tot de trainingsverzameling behoorden.

Het opgestelde classificatiemodel kan vervolgens gebruikt worden door de SVM om nieuwe tweets een klasse toe te wijzen. We merken op dat de classificeerder slechts een voorspelling doet en dus niet altijd in staat is de juiste klasse toe te kennen op basis van de informatie in de tweetvector.

4.2.4 Gebeurtenisdetectie

Nadat de tweets geclassificeerd zijn, kunnen we overgaan tot de gebeurtenisdetectie. In deze stap willen we nagaan of er werkelijk een interessante gebeurtenis heeft plaatsgevonden en de detectie van een piek niet louter toeval was. We zijn enkel geïnteresseerd in de tweets in het tweede deel van het zoekvenster. Voor elk van deze tweets werd in de vorige stap een klasse voorspeld.

Een gebeurtenis vindt plaats als het aantal tweets dat tot de klasse 'geen gebeurtenis' behoort, kleiner is dan een grenswaarde:

$$\frac{Aantal\ tweets\ klasse\ geen\ gebeurtenis}{Totaal\ aantal\ tweets} < \beta_{detectie}$$

Op basis van de ontwikkelingsverzameling werd $\beta_{detectie} = 0,55$ bepaald als de beste grenswaarde.

We maken bewust gebruik van een geen-gebeurtenisgrenswaarde omdat sommige gebeurtenissen samengesteld zijn. Indien we hier regels invoeren om met samengestelde gebeurtenissen te kunnen omgaan, komen we echter terecht in het domein van de identificatie.

4.2.5 Gebeurtenisidentificatie

Na de gebeurtenisdetectie volgt de identificatie van de gebeurtenis. Reeds in een vorige stap werden de tweets binnen de tweede helft van het zoekvenster geclassificeerd. Om te bepalen welke gebeurtenis heeft plaatsgevonden tellen we het aantal tweets per klasse. De klasse met het meeste aantal tweets binnen de tweede helft van het zoekvenster wordt aangenomen als winnende gebeurtenis. Dit kan wiskundig uitgedrukt worden als:

$$\arg\max_{y_i} \frac{Aantal\ tweets\ met\ klasse\ y_i}{Totaal\ aantal\ tweets}, \quad i = 1...6$$

Waarbij aan elke klasse y_i een gebeurtenis is gekoppeld en we de klasse 'geen gebeurtenis' buiten beschouwing laten.

Bovendien moet voor de winnende klasse y_i gelden:

$$\frac{Aantal\ tweets\ met\ klasse\ y_i}{Totaal\ aantal\ tweets} \geq \beta_{identificatie}$$

waarbij $\beta_{identificatie}$ experimenteel bepaald werd als 0,4 op basis van de ontwikkelingsverzameling. Indien deze grenswaarde niet gehaald wordt, wordt dit moment alsnog als 'geen gebeurtenis' geïdentificeerd.

Doordat een doelpunt vaak kort voorafgegaan wordt door een strafschop, introduceren we ook nog een speciale voorwaarde. Indien zowel de klasse doelpunt als strafschop beiden

de grenswaarde van 0,4 niet halen, maar de som van beiden wel, dan wordt dit moment in de tijd beschouwd als een doelpunt. De reden hiervoor is tweeledig. De eerste reden is dat in de tijd een strafschop altijd vooraf zal gaan aan een doelpunt. Bijgevolg zal de strafschop reeds gedetecteerd zijn op een vroeger tijdstip. De tweede reden is dat deze regel ons in staat stelt een doelpunt sneller te detecteren. We moeten immers niet wachten tot wanneer de grenswaarde $\beta_{identificatie}$ overschreden wordt voor de klasse 'doelpunt'.

We merken op dat het niet altijd mogelijk zal zijn om elke gebeurtenis te identificeren. Om dit te illustreren hebben we voor de wedstrijd Sunderland - Aston Villa uit de ontwikkelingsverzameling alle tweets na het toepassen van de filterstap geannoteerd. De wedstrijd eindigde op 2-2. In totaal werden er 21 fouten toegekend waarvan 4 gele kaarten: 3 aan Aston Villa en 1 aan Sunderland. Er werden ook 4 spelers gewisseld waarvan 1 van Aston Villa. Tabel 4.9 toont de verdeling van de tweets per klasse. Onmiddelijk valt op dat ondanks dat er 4 maal gescoord is, er 21 fouten zijn toegekend waarvan 4 gele kaarten en 4 wissels zijn uitgevoerd, doelpunten procentueel gezien een veel groter aandeel innemen dan wissels en fouten. Dit bevestigt de veronderstelling dat doelpunten door supporters als de belangrijkste gebeurtenis worden aanzien. Opvallend is dat ondanks dat er 3 gele kaarten aan Aston Villa werden toegekend, slechts 1 gele kaart werd gerapporteerd. Bijgevolg kunnen we geen enkele 'gele kaart'-gebeurtenis identificeren aangezien de verhouding $\frac{Aantal\ tweets\ met\ klasse\ y_i}{Totaal\ aantal\ tweets}$ nooit hoger dan 0,4 zal worden bij een minimum van 10 tweets in de tweede helft van het zoekvenster (zie Sectie 4.2.2).

Tabel 4.9: Overzicht van de verdeling van de tweets per klasse tijdens de wedstrijd Sunderland - Aston Villa van 29/10/11 voor de hashtag #AVFC.

Klasse	**Aantal**	**Procentueel**
Geen gebeurtenis	626	70,65%
Doelpunt	239	26,98%
Strafschop	0	0,00%
Fout	1	0,11%
Wissel	4	0,45%
Einde eerste helft	6	0,68%
Einde tweede helft	10	1,13%
Totaal	**886**	**100,00** %

4.2.6 Extractie van relevante informatie

Nadat de gebeurtenis geïdentificeerd is, kunnen we overgaan tot de extractie van de relevante informatie. Voor elk type gebeurtenis kunnen we andere informatie afleiden. We overlopen de verschillende gebeurtenissen en sommen op wat wel en wat niet mogelijk is op basis van tweets. We merken op dat het ook hier niet altijd mogelijk zal zijn de correcte informatie te extraheren. Zelfs al kunnen we de gebeurtenis identificeren, toch

komt het vaak voor dat niet alle benodigde informatie aanwezig is in de tweets.

Aan elke gebeurtenis is er een tijdstip verbonden. Bij voetbal is het de gewoonte om telkens de minuut aan te duiden waarin de gebeurtenis heeft plaatsgevonden. We berekenen deze waarde voor de gebeurtenissen doelpunt, strafschop, fout en wissel als het moment dat de gebeurtenis gedetecteerd wordt, verminderd met de helft van de lengte van het zoekvenster en vervolgens naar boven afgerond op een minuut. De reden hiervoor is dat het midden van het zoekvenster samenvalt met de start van de stijging en bijgevolg verondersteld wordt als het moment dat de gebeurtenis heeft plaatsgevonden. Voor de gebeurtenissen einde eerste helft en einde tweede helft wordt deze waarde berekend als het moment dat de gebeurtenis gedetecteerd wordt, naar boven afgerond op een minuut. De reden voor deze 2 verschillende berekeningswijzen is dat de gebeurtenissen einde eerste helft en einde tweede helft door de supporters geanticipeerd worden. Hierdoor worden tweets die het einde van de eerste of tweede helft rapporteren voor het einde van de eerste of tweede helft verzonden. Bijgevolg ontstaat het fenomeen dat na toepassing van de correctiefactor we een tijdstip uitkomen dat voor het werkelijke tijdstip ligt.

4.2.6.1 Het doelpunt

Een doelpunt wordt gekenmerkt door een speler en een score. Door het sterke verband tussen de speler en de score beschouwen we bij de extractie in eerste instantie enkel de spelers die minstens eenmaal in een tweet voorkomen met een scorevermelding. De speler die het meest vermeld wordt in de tweets wordt als doelpuntenmaker beschouwd.

De geëxtraheerde score is pas geldig als het aantal tweets dat dezelfde score vermeldt groter is dan 10%:

$$\frac{Aantal\ tweets\ met\ scorevermelding}{Totaal\ aantal\ tweets} > 0,10$$

Bovendien moet een spelersnaam in meer dan 15% van de tweets vernoemd worden om aanvaard te worden als doelpuntenmaker:

$$\frac{Aantal\ tweets\ met\ spelersnaam}{Totaal\ aantal\ tweets} > 0,15$$

Hiermee willen we vermijden dat een andere speler als doelpuntenmaker wordt beschouwd. Het resultaat van een volledige doelpuntdetectie ziet er als volgt uit:

> 31': R.van Persie scores! (1-1)

Indien geen speler gevonden wordt, wordt de ploegnaam vermeld:

> 31': Arsenal scores! (1-1)

Het probleem met de scorevermelding is dat gebruikers niet altijd even accuraat tweeten over de tussenstand. Het grootste probleem is dat gebruikers de tussenstand omdraaien. Indien het team dat niet thuis speelt eerst scoort, is de stand 0-1 en niet 1-0. De oplossing is dat we gebruik maken van de spelersvermelding om de juiste tussenstand te bepalen. We veronderstellen dat we voor de wedstrijd weten welke 2 teams spelen en wie thuis speelt. Op basis van de speler die het doelpunt gemaakt heeft en de scorevermeldingen die we terugvinden in de tweets, kunnen we op accurate wijze de score bepalen.

Een speciaal geval is een owngoal. In zo een geval moeten we score omgekeerd aanpassen. Een speler heeft namelijk in zijn eigen goal gescoord. Om owngoals te detecteren maken we gebruik van sleutelwoorden zoals 'owngoal' en 'OG'. Het resultaat ziet er dan als volgt uit:

> 23': L.Koscielny scores! Own Goal. (1-0)

4.2.6.2 De strafschop

Een strafschop wordt over het algemeen beschouwd als een zo goed als gescoord doelpunt. De speler waarop de fout gemaakt wordt of de speler die de fout heeft gemaakt, blijkt in de praktijk van ondergeschikt belang. Zelfs al zou over één van beiden getweet worden, is het bijzonder moeilijk om na te gaan aan welke ploeg de strafschop is toegekend op basis van de spelersnaamvermeldingen.

Een tweede probleem is dat de echtheid van een strafschopgebeurtenis moeilijk gecontroleerd kan worden. Tijdens een live-uitzending wordt een strafschop vaak herhaald vele minuten nadat deze strafschop heeft plaatsgevonden waardoor discussies plots weer oplaaien en er opnieuw een strafschop gedetecteerd wordt. Een ander probleem stelt zich wanneer een speler geen strafschop toegekend krijgt omdat de scheidsrechter de fout niet gezien heeft. We krijgen dan te maken met tweets zoals:

> "That's a penalty!!!!!! God damn it stupid refs #MCFC"[2]

De eerste zin in deze tweet is bijna identiek aan een typische tweet die tijdens een strafschopmoment verstuurd wordt. Bijgevolg zullen we in dit systeem enkel de strafschopinformatie gebruiken als dit een doelpunt tot gevolg heeft. In Sectie 4.3 wordt dit verder uitgewerkt. Het finale resultaat is dan van de vorm:

> 7': W.Rooney scores! Penalty. (1-0)

[2]Verzonden door sam_pritch op 13:54:38 14/4/12 tijdens Norwich City - Manchester city.

4.2.6.3 De fout

Net zoals bij een strafschop kan een fout tot heel wat controverse leiden. Opnieuw stellen we vast dat supporters niet alleen de gebeurtenis rapporteren maar ook becommentariëren. We maken daarom gebruik van het democratisch principe dat het type fout dat het meest gerapporteerd wordt als waarheid wordt beschouwd. We maken hierbij gebruik van sleutelwoorden. We beschouwen 2 soorten fouten f_i: een gele kaart en een rode kaart. Er moet gelden dat:

$$\arg\max_{f_i} \frac{Aantal\ tweets\ met\ klasse\ f_i}{Totaal\ aantal\ tweets}, \quad i = 1...2$$

Om aan het probleem van willekeurige commentaren het hoofd te kunnen bieden moet voor de winnende klasse f_i gelden dat:

$$\frac{Aantal\ tweets\ met\ klasse\ f_i}{Totaal\ aantal\ tweets} > 0,15$$

Indien de grenswaarde niet bereikt wordt, wordt een fout gerapporteerd zonder een type kaart te vermelden. Een fout die niet tot een kaart heeft geleid, valt ook onder deze categorie.

Het resultaat van de volledige foutendetectie is in het geval van een gele kaart van de vorm:

> 20': Foul: Yellow Card.

4.2.6.4 De Wissel

Een wissel zal altijd bestaan uit 2 spelers van dezelfde ploeg. We kijken dus naar alle spelers die samenkomen in een tweet en tellen voor elke spelersnaam individueel hoeveel keer deze voorkomt. De 2 spelers die individueel het meest voorkomen en ook samen in een tweet worden vernoemd, worden als de gewisselde spelers beschouwd. Een mogelijk resultaat van de wisseldetectie ziet er als volgt uit:

> 61': Substitution : A.Young, L.Nani

4.2.6.5 Het einde van de eerste en tweede helft

Het einde van de eerste helft en het einde van de tweede helft zijn beiden gebeurtenissen die slechts eenmaal voorkomen. De enige informatie die typisch geassocieerd wordt met deze gebeurtenissen, is de score. De score komt heel vaak voor in tweets die het einde van een helft aangeven (zie Tabel 4.6). Opnieuw duikt hier echter de moeilijkheid op dat

gebruikers de score vaak omwisselen. Bij deze gebeurtenis kunnen we echter geen gebruik maken van spelersnamen aanwezig in de tweets om de score te verifiëren. Een alternatieve mogelijkheid is om gebruik te maken van reeds gedetecteerde doelpunten.

Het resultaat van de volledige detectie van het einde van de eerste helft is dan van de vorm:

> 46': End of first half.(2-0)

4.2.7 Besluit

In deze sectie hebben we een algoritme voorgesteld dat in staat is om op basis van een verzameling regels en een Support Vector Machine (SVM) gebeurtenissen in Twitter te detecteren en te identificeren tijdens een voetbalwedstrijd. We kunnen het onderscheid maken tussen 6 belangrijke gebeurtenissen, namelijk een doelpunt, een strafschop, een fout, een wissel, het einde van de eerste helft en het einde van de tweede helft. Het voorgestelde algoritme houdt echter geen rekening met wat reeds in de vorige gebeurtenissen is gedetecteerd en geïdentificeerd. Dit is het onderwerp van de volgende sectie.

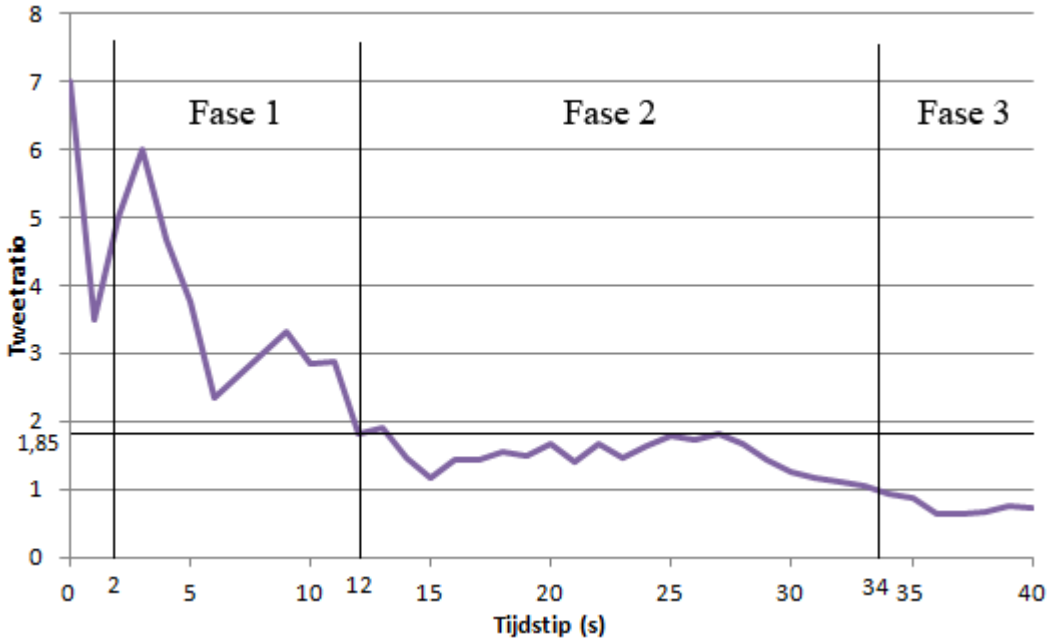
4.3 Geavanceerde gebeurtenisdetectie en -identificatie

Tot nu toe hebben we een algoritme dat enkel in staat is om elke seconde individueel te analyseren. Wanneer een gebeurtenis gedetecteerd wordt op tijdstip 33:00 met een zoekvenster van 30 seconden, is de kans groot dat deze gebeurtenis ook gedetecteerd zal worden op tijdstip 33:01 met een zoekvenster van 30 seconden door de overlapping in zoekvensters. In deze sectie breiden we het basis detectie- en identificatiealgoritme uit tot een volwaardig detectie- en identificatiesysteem.

4.3.1 Geavanceerde piekdetectie

In Sectie 4.2.2 werden 2 voorwaarden geïntroduceerd om een stijging in het tweetvolume vast te stellen, namelijk $\frac{Aantal\ tweets\ tweede\ helft}{Aantal\ tweets\ eerste\ helft} > \alpha_{piek}$ en het aantal tweets in de tweede helft van het zoekvenster moet groter dan of gelijk zijn aan α_{min} waarbij in de praktijk $\alpha_{piek} = 1,85$ en $\alpha_{min} = 10$. In Figuur 4.8 wordt het verloop van de tweetverhouding weergegeven bij een zoekvenster van 20 seconden. De eerste maal dat voldaan is aan beide voorwaarden is na 2 seconden.

Op basis van de definitie van een piek kunnen we 3 fasen onderscheiden:



Figuur 4.8: Illustratie van de tweetratio voor een zoekvenster van 20 seconden tijdens Chelsea - Arsenal 29/10/11 gerekend vanaf 45 minuten 30 seconden in de wedstrijd.

1. **De sterke stijging:** Een piek wordt altijd ingezet door een sterke stijging. In ons geval betekent dit dat de tweetratio groter is dan 1,85.

2. **Stagnatie:** De sterke stijging begint af te nemen en de tweetratio stagneert. De tweetratio is nog steeds groter dan 1 wat duidt op een stijging.

3. **Een geleidelijke daling:** Eens we over het hoogtepunt van de gebeurtenis heen zijn, wordt de daling ingezet. De tweetratio zal lager dan 1 worden.

De verschillende fasen worden geïllustreerd in Figuur 4.8.

In Sectie 4.1.1 hebben we vastgesteld dat met elke piek een gebeurtenis is gerelateerd. We verzachten daarom de voorwaarden van de tweetratio. Wanneer minstens eenmaal een sterke stijging gedetecteerd is (fase 1), is het voldoende dat de tweetratio hoger is dan 1 (fase 2). Vanaf wanneer de tweetratio lager wordt dan 1 dan betekent dit het einde van een gebeurtenis (fase 3). Bijgevolg zal opnieuw een sterke stijging nodig zijn die het begin van een nieuwe gebeurtenis aankondigt.

Door het invoeren van deze regel wordt nu niet langer op elke seconde gepoogd een gebeurtenis te detecteren en te identificeren, maar kan aan elke piek één gebeurtenis gelinkt worden. Vanaf wanneer een bepaalde gebeurtenis is geïdentificeerd dan kan deze niet meer opnieuw voorkomen binnen een piek.

4.3.2 Geavanceerde gebeurtenisdetectie en -identificatie

Een tweetratio hoger dan 1,85 zal niet altijd leiden tot een gebeurtenisdetectie. Meestal is dit omdat er geen gebeurtenis heeft plaatsgevonden, maar soms kan het gebeuren dat de drempel voor gebeurtenisdetectie net niet gehaald wordt.

In de praktijk komt het voor dat een speler met zijn bijnaam in plaats van zijn echte naam wordt genoemd. Indien de bijnaam niet in de spelersdatabank zit, zal de bijnaam als een gewoon woord geïnterpreteerd worden. Bijgevolg zal de SVM meerdere tweets in de verkeerde klasse indelen, meer bepaald in de klasse 'geen gebeurtenis'. Door te werken met 3 fasen in de piekdetectie, verhogen we de kans om een zoekvenster te vinden waarbij voldoende tweets een klasse krijgen toegewezen verschillend van de klasse 'geen gebeurtenis'. Hierbij doen we geen afbreuk aan het idee dat een piek gecorreleerd is met een gebeurtenis.

Een gelijkaardige redenering kan gemaakt worden voor de gebeurtenisidentificatie.

4.3.3 Geavanceerde extractie van relevante informatie

Doordat we een gebeurtenis op meerdere opeenvolgende seconden kunnen detecteren, zijn we niet verplicht om reeds de benodigde informatie bij de eerste detectie te extraheren. We kunnen bepaalde voorwaarden opleggen aan de geëxtraheerde informatie. Bovendien kunnen we ook gebruik maken van de tweets verzonden door beide partijen. We overlopen voor elke gebeurtenis de verbeteringen die geïntroduceerd werden.

4.3.3.1 Het doelpunt

Uit Sectie 4.2.5 weten we dat het doelpunt de meest belangrijke gebeurtenis is tijdens een voetbalwedstrijd. Hierdoor zal een doelpunt op verschillende opeenvolgende tijdstippen gedetecteerd worden. Niet op elk tijdstip zullen we zowel de score als de speler kunnen extraheren. Om de accuraatheid van het detectiesysteem te verhogen, voeren we enkele nieuwe regels in.

De eerste wijziging die we aanbrengen is een intern doelpuntensysteem. Doorheen het verloop van de wedstrijd wordt de score bijgehouden door gebruik te maken van de reeds gedetecteerde doelpunten. Indien een doelpunt wordt gedetecteerd en de bijhorende score kan geëxtraheerd worden, zal er gecontroleerd worden of een verhoging van de score wel mogelijk is. De assumptie die we hierbij maken is dat elke doelpunt van de eigen ploeg gedetecteerd kan worden. Bijgevolg kunnen de eigen gescoorde doelpunten slechts met 1 verhogen per gebeurtenis en kunnen meerdere detecties van hetzelfde doelpunt vermeden worden. Hiermee willen we onder andere vermijden dat gebeurtenissen uit andere

wedstrijden interferen met de wedstrijd die we analyseren. We merken op dat deze regel enkel geldt voor eigen gescoorde doelpunten omdat in de praktijk niet alle tegendoelpunten gedetecteerd worden. Doordat we tijdens een wedstrijd kunnen beschikken over de Twitterberichtstromen van beide ploegen kunnen we een systeem bouwen dat accuraat de score tijdens een wedstrijd kan weergeven.

Wanneer een piek als doelpunt geïdentificeerd wordt, zullen we in eerste instantie in de opeenvolgende tijdstippen op zoek gaan naar een scorevermelding die aan de regels voldoet uit Sectie 4.2.6.1. Vanaf dan gaan we pas op zoek naar een geldige spelersnaamvermelding. Wanneer beiden gevonden zijn, wordt de gebeurtenis finaal vrijgegeven. Als de gebeurtenis de derde fase van de piekdetectie betreedt en er nog steeds geen spelersnaam is gevonden, zal er wel een doelpunt gedetecteerd worden, maar dan zonder spelersnaam. In dat geval wordt de ploegnaam getoond.

4.3.3.2 De strafschop

Doordat we bij de detectie van strafschoppen vaak valse positieven hebben, gebruiken we de informatie van een strafschop enkel om doelpunten te annoteren. We gaan ervan uit dat indien binnen de 2 minuten voor de doelpuntdetectie een strafschopdetectie heeft plaatsgevonden, het doelpunt het gevolg van een strafschop is.

4.3.3.3 De fout

Bij het detecteren van fouten onderscheiden we 2 problemen: over sommige fouten wordt helemaal niet getweet en andere fouten veroorzaken meerdere detecties met soms meerdere minuten tussen 2 opeenvolgende detecties. Het eerste probleem werd reeds aangetoond in Sectie 4.2.5 waarbij het aantal verzonden tweets van de klasse 'fout' gelijk was aan 1 terwijl er in werkelijkheid in totaal 21 fouten waren toegekend. Het tweede probleem wordt veroorzaakt door de herhaling van de foutfase in de videosequentie. In een eerste golf van tweets wordt de fout gerapporteerd. Nadat de herhaling van de foutfase is afgespeeld, ontstaat er een tweede golf van tweets. Hierdoor worden 2 pieken gecreëerd.

Om de tweede piek te maskeren, voeren we de regel in dat er tussen elke 2 fouten een periode van 90 seconden moet zitten. Het is mogelijk dat er meerdere fouten plaatsvinden binnen een tijdsspanne van 90 seconden, maar we moeten de afweging maken tussen meerdere identificaties van dezelfde fout of het niet detecteren van meerdere fouten binnen een tijdsspanne van 90 seconden.

4.3.3.4 De wissel

Zoals reeds uitgelegd in Sectie 4.2.6.4 wordt elke wissel uniek gedefinieerd door 2 spelers. Bijgevolg kunnen we elke wissel controleren op basis van een lijst van voorgaande wissels.

4.3.3.5 Het einde van de eerste en tweede helft

Het einde van de eerste en tweede helft zijn beide gebeurtenissen die slechts eenmaal kunnen voorkomen. Door simpelweg bij te houden of deze gebeurtenis al heeft plaatsgevonden, vermijden we dat deze gebeurtenis meerdere keren plaatsvindt. Indien we beschikken over de Twitterberichtstromen van beide ploegen dan kan de stand op het einde van de eerste en tweede helft samengesteld worden uit de stand van eigen gescoorde doelpunten van beide ploegen.

4.4 Besluit

In dit hoofdstuk hebben we een systeem geïntroduceerd dat in staat is de 6 belangrijkste gebeurtenissen in een wedstrijd in realtime te detecteren, te identificeren en hierover extra informatie te extraheren.

We zijn vertrokken van het idee dat de pieken in het tweetvolume gecorreleerd zijn met deze belangrijke gebeurtenissen. Indien er een sterke stijging optrad gingen we de tweets in detail bestuderen. We maakten gebruik van een Support Vector Machine (SVM) en een verzameling regels om eerst de tweets te classificeren, en om vervolgens een gebeurtenis te detecteren en te identificeren. Zodra de gebeurtenis geïdentificeerd werd, probeerden we de bijhorende interessante informatie te extraheren.

Nadat we het basisalgoritme gedefinieerd hadden, hebben we dit gebruikt om een volwaardig detectiesysteem te bouwen dat rekening houdt met gebeurtenissen die reeds in het verleden geïdentificeerd werden. We merken op dat het moeilijk is om voor sommige gebeurtenissen een robuuste verzameling regels vast te leggen. Sommige gebeurtenissen blijken niet interessant genoeg voor Twittergebruikers en andere gebeurtenissen zijn niet uniek identificeerbaar. In hoofdstuk 6 wordt dit systeem geëvalueerd.

Hoofdstuk 5

Extractie van hoogtepunten uit voetbalvideosequenties in realtime

In dit hoofdstuk stellen we een systeem voor dat in realtime de hoogtepunten uit een live-voetbalvideosequentie haalt door gebruik te maken van Twitter. We focussen ons op de doelpunten omdat dit de belangrijkste gebeurtenissen zijn tijdens een voetbalwedstrijd. We maken gebruik van het systeem ontwikkeld in het vorige hoofdstuk om een indicatie te geven wanneer het doelpunt heeft plaatsgevonden. Hierdoor moeten we niet de volledige videosequentie analyseren. Na extractie van een videohoogtepunt, kunnen we dit ook annoteren met de naam van de speler die gescoord heeft, de ploeg, de minuut waarin er gescoord werd en de stand na het scoren van het doelpunt.

We beginnen dit hoofdstuk met de analyse van doelpuntfragmenten in voetbalvideosequenties in Sectie 5.1. Vervolgens ontwikkelen we in Sectie 5.2 enkele algoritmen die ons kunnen helpen bij een doeltreffende extractie van videofragmenten. Finaal presenteren we in Sectie 5.3 een algoritme dat op basis van Twitterberichten de juiste herhaling uit de videosequentie zal extraheren en annoteren. We eindigen met een besluit in Sectie 5.4.

5.1 Analyse van live-voetbaluitzendingen met betrekking tot doelpunten

5.1.1 Cinematografische karakteristieken van live-voetbaluitzendingen

5.1.1.1 Shottypes

Tijdens een voetbalwedstrijd maakt de regisseur gebruik van beelden van wel 35 camera's die op en rond het veld staan opgesteld. De regisseur zal tijdens de wedstrijd afwisselen

(a) Bovenaanzicht (b) Middenaanzicht (c) Close-up (d) Buitenveldaanzicht

Figuur 5.1: Illustratie van de 4 belangrijkste shottypes. Soms worden (c) en (d) als 1 type beschouwd.

tussen de verschillende camera's. Een groep opeenvolgende beelden die door 1 camera gemaakt zijn, noemen we een shot. De beelden die deze camera's maken kunnen we opdelen in 3 of 4 grote groepen [23]. Figuur 5.1 geeft hier een overzicht van.

Het eerste shottype is het bovenaanzicht. Zoals de naam al aangeeft zijn dit beelden die van bovenaf worden gefilmd. De beelden bestaan voornamelijk uit groen gras en miniscule spelers.

Het tweede type shot is het middenaanzicht. Dit beeld kan gemaakt worden door camera's aan de zijlijn maar ook door bovenaanzichtcamera's waarbij sterk is ingezoomd. Kenmerkend voor dit beeld is dat één of meerdere spelers volledig in beeld genomen worden waarbij ze bijna de volledige hoogte van het beeld innemen. Er is ook minder gras zichtbaar.

Het derde shottype is de close-up. Hierbij worden de spelers heel dicht in beeld gebracht waarbij nog slechts een deel van hun lichaam in beeld is.

Het laatste shottype is het buitenveldaanzicht. Hierbij wordt typisch het publiek in beeld gebracht. Sommige auteurs beschouwen het buitenaanzicht en de close-up als 1 shottype omdat ze dezelfde karakteristieken vertonen [23, 24, 25, 27].

5.1.1.2 De overgang tussen twee shots

Wanneer de regisseur een overgang wil maken tussen de beelden van 2 camera's tijdens een voetbalwedstrijd, worden er typisch 2 verschillende technieken gebruikt [53]. De eerste techniek is een abrupte overgang (Engels: cut) waarbij er van het ene op het andere beeld (Engels: frame) de beelden van een andere camera worden gebruikt. De tweede techniek is de graduele overgang (Engels: gradual transition) waarbij de beelden van 2 camera's in elkaar overvloeien over een aantal beelden. Abrupte overgangen worden als standaardtechniek gebruikt om te wisselen tussen camera's tijdens de wedstrijd. Graduele overgangen worden gebruikt tijdens herhalingen om verschillende shots in elkaar te laten overvloeien. Beide overgangen worden afgebeeld in Figuur 5.2.

Figuur 5.2: Illustratie van de 2 belangrijkste types overgang tussen 2 shots. Bovenaan wordt een abrupte overgang tussen 2 shots afgebeeld en onderaan een graduele overgang tussen 2 shots.

Figuur 5.3: Illustratie van een logotransitie gebruikt om een herhaling in te leiden tijdens Liverpool-Arsenal 3/3/12.

5.1.1.3 Herhalingen

Wanneer er een interessante gebeurtenis heeft plaatsgevonden tijdens een wedstrijd, zal de regisseur beslissen om de gebeurtenis te herhalen vanuit enkele andere camerastandpunten. In de meeste competities wordt er gebruik gemaakt van een logo om de herhaling in te leiden en af te sluiten. Het gebruikte logo is altijd eigen aan de competitie en hangt soms ook af van het type gebeurtenis. In Figuur 5.3 wordt een 2-delige logotransitie afgebeeld die gebruikt werd tijdens een doelpunt in een Engelse Premier League-wedstrijd tussen Liverpool en Arsenal tijdens het seizoen 2011-2012. We merken op dat het hier gebruikte logo specifiek voor een doelpunt is en bovendien het clublogo bevat van de ploeg die gescoord heeft.

5.1.2 Cinematografische technieken voor en na een doelpunt

In de periode vlak voor en na het doelpunt maakt de regisseur gebruik van een vaste opeenvolging van types shots [23]. Tijdens het moment dat vooraf gaat aan het doelpunt, is er ofwel veel beweging ofwel vindt er een stilstaande fase plaats. In beide gevallen zal er een bovenaanzicht gebruikt worden. Na het scoren, zal de regisseur direct overschakelen op close-upbeelden van de spelers en in het bijzonder de doelpuntenmaker. Daarna zal de herhaling van het doelpunt ingeleid worden door middel van een logotransitie. Vervolgens worden er een aantal opeenvolgende shots getoond die het doelpunt vanuit een

Figuur 5.4: Opeenvolging van de types shot voor en na het vallen van een doelpunt. Een groen balkje stelt een bovenaanzicht voor en een blauw balkje een close-up. Een rood balkje staat voor een logotransitie.

ander camerastandpunt in beeld brengen. Deze shots vloeien gradueel in elkaar over. De herhaling wordt beëindigd met het opnieuw verschijnen van een logo. Dit wordt schematisch voorgesteld in Figuur 5.4. We merken op dat de verschillende gekleurde balkjes in de figuur kunnen bestaan uit meerdere shots van hetzelfde type.

Soms gebeurt het dat een herhaling abrupt wordt afgebroken omdat er ondertussen al een andere interessante fase aan de gang is [32]. In dat geval kunnen we het einde van een herhaling herkennen doordat er een abrupte overgang is tussen 2 shots in plaats van een graduele.

5.2 Shotgrens- en logodetectie

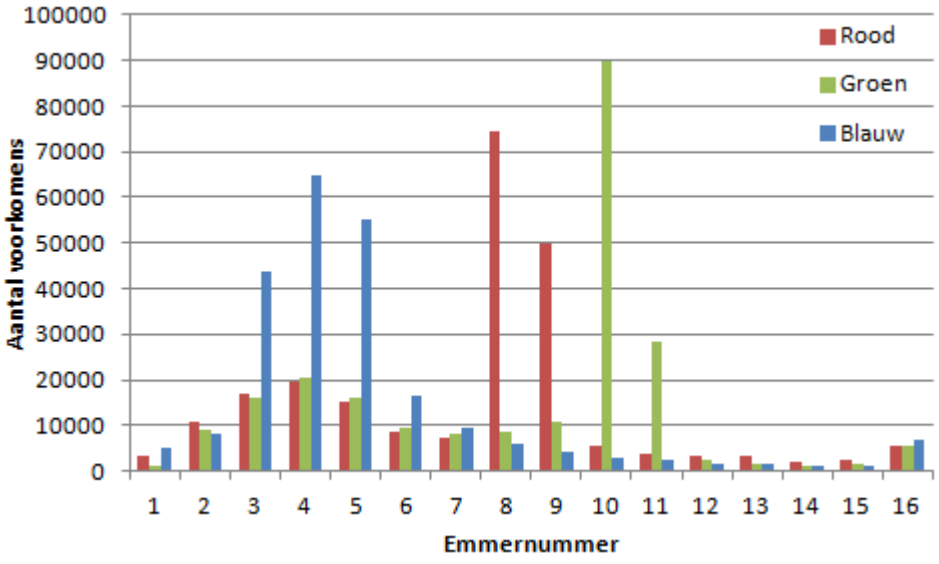
Uit de vorige sectie blijkt dat indien we de herhaling van een doelpunt uit een videosequentie willen halen, we in staat moeten zijn logotransities te detecteren en abrupte overgangen tussen 2 shots. Voor beide problemen zullen we gebruik maken van lokale histogrammen waarbij de afmetingen bepaald worden door de gulden snede.

Voor de ontwikkeling van zowel de shotgrensdetectie als de logodetectie, hebben we gebruik gemaakt van de ontwikkelingsverzameling weergegeven in Tabel A.3. De verzameling bevat 3 wedstrijden met 2 soorten logo's (zie Figuur 5.7) die eenzelfde beweging maken. 2 wedstrijden zijn in het MPEG-4 formaat en 1 wedstrijd in het MPEG-2 formaat. De resoluties zijn respectievelijk 640x360 pixels en 720x576 pixels.

5.2.1 Lokale histogrammen

Elke pixel in een beeld wordt gekenmerkt door een RGB-waarde die bestaat uit een rode, een groene en een blauwe component om de kleur van de pixel te representeren. Op basis van deze RGB-waarde kunnen we de afbeelding karakteriseren door voor elke component een histogram op te stellen. Een histogram is een representatie van de frequentieverdeling van zo een component in een aantal voorgedefinieerde klassen. Binnen de beeldanalyse

(a)

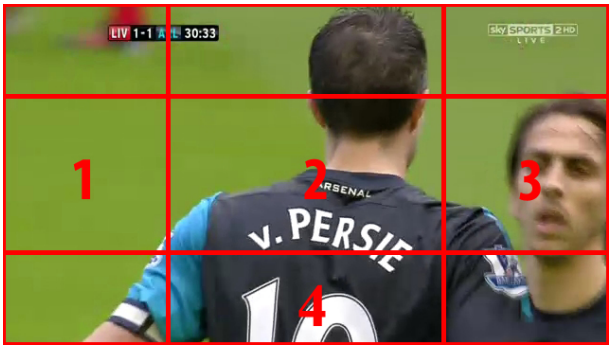


(b)

Figuur 5.5: Illustratie van een beeld uit een videosequentie (a) en het bijhorende histogram van RGB-waarden (b). De afbeelding komt uit de wedstrijd Liverpool-Arsenal 3/3/12.

maken we hiervoor gebruik van emmers (Engels: bins). Elke component heeft een waardebereik van 0 tot 255 en een emmer vertegenwoordigt een deel van dit waardebereik. Het waardebereik van een emmer kan berekend worden als $\frac{256}{\#emmers}$. Indien het aantal emmers een macht van 2 is met een maximum van 256, zal het waardebereik van een emmer altijd een geheel getal zijn.

In Figuur 5.5(b) wordt het histogram weergegeven van de afbeelding in Figuur 5.5(a). We hebben de RGB-waarden verdeeld in 16 gelijke emmers. Emmer 1 komt overeen met de waarden in het interval [0-15] en emmer 16 met de waarden in het interval [240-255]. Op basis van het histogram kunnen we de dominante kleuren afleiden van de afbeelding. In ons geval is dit het groen van het speelveld en het donkerblauw van het truitje van de speler. Dit is ook duidelijk zichtbaar in het histogram. Het speelveld wordt gekarakteriseerd door een rode piek in emmer 8 en 9 en een groene piek in emmer 10 en 11. Het truitje van de speler kan herkend worden aan de blauwe pieken in emmer 3, 4 en 5.



Figuur 5.6: Illustratie van de guldensnedecompositieregel. Het beeld wordt zowel in de hoogte als in de breedte verdeeld volgens een 3:5:3-verhouding.

In voetbalwedstrijden maken camera-operatoren vaak gebruik van de guldensnedecompositieregel [23, 25, 27]. Hierbij wordt het beeld opgedeeld in 9 grote delen volgens een 3:5:3-verhouding zoals afgebeeld in Figuur 5.6. De camera-operator zal proberen het belangrijkste onderdeel van de actie altijd weer te geven in regio 1, 2 en 3. Daarom kennen we elke regio een eigen lokaal histogram toe in plaats van gebruik te maken van een globaal histogram. Omdat het moeilijk is om op basis van deze 3 regio's een onderscheid te maken tussen een close-up en een middenaanzicht introduceren we een vierde regio onder regio 2 [27]. Bij een middenaanzicht zal de dominante kleur in deze regio typisch groen zijn terwijl dit bij een close-up de kleur van de kleding van een persoon zal zijn.

Het voordeel van lokale histogrammen is dat we bij shotgrensdetectie enkel de belangrijkste regio's vergelijken en bijgevolg het aantal berekeningen halveren. Een tweede voordeel is dat een bovenaanzicht, een middenaanzicht en een close-up accurater kunnen gekarakteriseerd worden. Als we teruggrijpen naar Figuur 5.1, dan zien we dat zowel figuur (a), (b) als (c) groen als dominante kleur hebben waardoor de histogrammen sterk gelijkend zullen zijn. Het derde voordeel is dat logotransities accurater kunnen gedetecteerd worden. Tijdens een logotransitie zal een logo bijvoorbeeld van links naar rechts verschuiven. De samenstelling van de lokale histogrammen zal continu sterk veranderen terwijl een globaal histogram slechts licht zal wijzigen tijdens een logotransitie.

5.2.2 Shotgrensdetectie

Om abrupte overgangen tussen 2 shots te detecteren, kunnen we de lokale histogrammen van elke 2 opeenvolgende beelden in de videosequentie met elkaar vergelijken. Dit doen we door voor elke component van elk lokaal histogram H_l het verschil te nemen tussen de corresponderende emmers.

$$D_l(i, i+1) = \sum_{j=1}^{3} \sum_{k=1}^{n} |H_{l,i}(j,k) - H_{l,i+1}(j,k)|$$

Hierbij is i het beeldnummer, j de componentindex, k het emmernummer en n het aantal emmers. $D_l(i, i+1)$ is bijgevolg het verschil tussen het lokaal histogram van regio l in beeld i en beeld $i+1$.

Het totale verschil tussen de 2 opeenvolgende beelden kan dan als volgt berekend worden:

$$D(i, i+1) = \sum_{l=1}^{4} D_l(i, i+1)$$

Om deze techniek te kunnen toepassen op videosequenties met een verschillende resolutie normaliseren we het resultaat als volgt:

$$D_n(i, i+1) = \frac{D(i, i+1)}{2 * \sum_{l=1}^{4} height_l * width_l}$$

Een abrupte shottransitie zal plaatsgevonden hebben wanneer

$$D_n(i, i+1) > 0,275$$

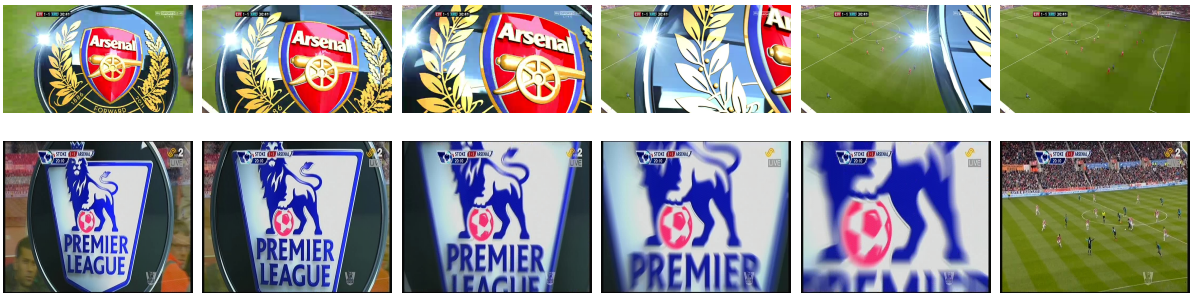
Hierbij hebben we de waarde 0,275 experimenteel bepaald door gebruik te maken van videosequenties uit de ontwikkelingsverzameling weergegeven in Tabel A.3.

5.2.3 Logodetectie

Het detecteren van een logo in een voetbalsequentie is op zich niet zo moeilijk aangezien een logo sterk verschillend is van alle andere uitgezonden beelden tijdens een voetbalwedstrijd [24, 26]. Omdat de ontwikkelingsverzameling wedstrijden bevat waarbij het logo wijzigt per ploeg, is een aanpak die gebruik maakt van een logosjabloon dat op voorhand afgeleid is, moeilijker [24].

In deze masterproef stellen we een aanpak voor die niet gebaseerd is op de RGB-waarden van het logo zelf, maar gebruik maakt van de beweging die een logo maakt tijdens een logotransitie. Figuur 5.7 toont de 6 laatste beelden tijdens 2 verschillende logotransities. De beweging van het logo vertrekt vanuit het midden en zal vervolgens het volledige scherm innemen om dan plots te verdwijnen.

In vergelijking met een abrupte shottransitie sluit een logotransitie eerder aan bij een graduele transitie. Hierbij is de overgang gespreid over meerdere beelden met een kleinere wijziging in de histogrammen. Daarom maken we gebruik van een zoekvenster met een grootte van 11 opeenvolgende beelden. We maken opnieuw gebruik van $D_n(i, i+1)$.



Figuur 5.7: Illustratie van de 6 laatste beelden van 2 verschillende logotransities.

Wanneer voor minstens 5 van de 10 overgangen geldt:

$$D_n(i, i+1) > 0,125$$

en dat er bovendien bij één overgang geldt:

$$D_n(i, i+1) > 0,5$$

dan hebben we een logotransitie gedetecteerd. De tweede voorwaarde onderscheidt een logotransitie van een gewone graduele transitie en wordt veroorzaakt door de abrupte wijziging tussen de 2 laatste beelden in Figuur 5.7. De grenswaarden werden bepaald op basis van de wedstrijden in de ontwikkelingsverzameling weergegeven in Tabel A.3.

5.3 Het combineren van Twitter en videosequenties

Langs de ene kant beschikken over technieken om herhalingen van doelpunten uit videosequenties te extraheren op basis van logo- en shotgrensdetectie. Langs de andere kant beschikken we over een systeem dat in realtime doelpuntmomenten kan detecteren in Twitter, ons kan vertellen wie gescoord heeft, in welke minuut dit was en wat de stand werd na het scoren van dit doelpunt. Bijgevolg moeten we enkel nog beide combineren. In deze sectie stellen we een algoritme voor dat op basis van een detectie van een doelpunt in Twitter de bijhorende herhaling uit de videosequentie haalt.

In Sectie 5.3.1 wordt de vertraging besproken tussen de live-uitzending en de Twitterberichtenstroom. Op basis van deze informatie wordt in Sectie 5.3.2 het finale algoritme voorgesteld.

5.3.1 Vertragingen in Twitter en live-uitzendingen

Vertragingen zijn een belangrijk onderdeel van elk systeem dat in realtime werkt. We onderscheiden in totaal 5 soorten vertraging. De eerste vertraging is de live-uitzendingsvertraging. Met betrekking tot het Twitterdetectiesyteem kunnen we 3 soorten vertraging onderscheiden: een menselijke vertraging, een Twittervertraging en een verwerkingsvertraging [18]. Wanneer een doelpunt gedetecteerd is in Twitter moet ook nog het bijhorende videofragment geëxtraheerd en geannoteerd worden. De laatste vertraging is dus de video-extractievertraging.

5.3.1.1 Live-uitzendingsvertraging

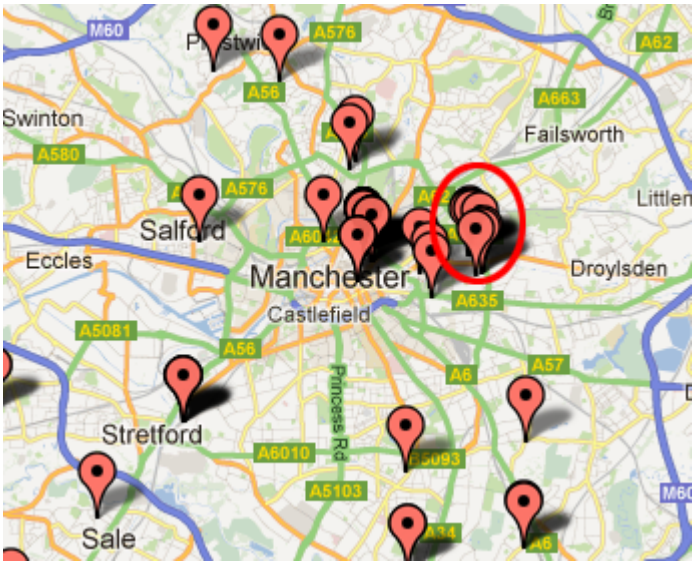
Een live-uitzendingsvertraging wordt door een tv-station en de regisseur geïntroduceerd om op onverwachte gebeurtenissen te anticiperen. Hierdoor kan een regisseur tijdig ingrijpen [18]. De auteurs in [18] schatten deze vertraging op 7 tot 12 seconden. Supporters kunnen zich echter ook in het stadion bevinden en tweets versturen. In dat geval is deze vertraging niet van toepassing.

In Figuur 5.8 wordt een kaart afgebeeld van de regio Manchester. De rode icoontjes zijn locaties van mobiele supporters die een tweet verzonden tijdens de wedstrijd Manchester City - Manchester United van 30/04/12 met de hashtag #MCFC. Tijdens deze wedstrijd hebben 362 supporters hun locatie vrijgegeven op een totaal van 21557 verschillende supporters. De rode cirkel duidt de locatie aan van het stadion van Manchester City. 5 verschillende gebruikers verzonden een tweet vanuit het stadion. We stellen vast dat dit slechts een marginaal deel van de tweets is en dat bijgevolg het overgrote deel van de tweets onderhevig is aan een live-uitzendingsvertraging.

5.3.1.2 Menselijke vertraging

Wanneer een gebeurtenis gebeurd is tijdens een voetbalwedstrijd en zichtbaar is in de live-uitzending zullen mensen hierover een tweet versturen. Het moment tussen het waarnemen van deze gebeurtenis en het verzenden van de tweet is de menselijke vertraging.

We onderscheiden hier een aantal types tweets. De allereerste tweets die verzonden worden na een doelpunt bevatten vaak enkel het woord 'goal'. Deze tweets worden waarschijnlijk verzonden door gebruikers die op voorhand hun tweet schrijven. Enkele seconden later vindt pas de echte explosie van tweets plaats die kort zijn en de spelersnaam bevatten. Het is deze explosie waarvan het Twitterdetectiesysteem gebruik maakt. Later evolueert de korte verslaggeving naar commentaren op het doelpunt en langere tweets.



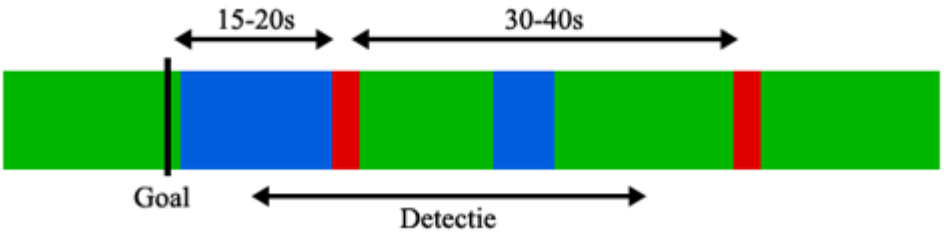
Figuur 5.8: Kaart van de regio Manchester. De rode icoontjes zijn locaties van mobiele gebruikers en de rode cirkel geeft de locatie van het stadion van Manchester City weer.

5.3.1.3 Twittervertraging

De Twittervertraging is de tijd die nodig is om een verzonden tweet te verwerken en door te sturen naar de computer waar het detectiesysteem op draait. In praktijk blijkt deze vertraging verwaarloosbaar te zijn op voorwaarde dat we gebruik maken van goede hashtags [18]. Twitter zal bij de verwerking voorrang geven aan tweets die een belangrijke gekende hashtag gebruiken. Tweets met minder gekende hashtags kunnen een vertraging oplopen tot 30 seconden. Indien we de hashtags uit Tabel 4.1 gebruiken blijkt de vertraging verwaarloosbaar klein.

5.3.1.4 Verwerkingsvertraging

Wanneer de tweets het detectiesysteem bereiken, kan de verwerking beginnen. Zoals we weten uit Hoofdstuk 4 maken we een gebruik van een zoekvenster met grootte 10, 20, 30 of 60 seconden. Impliciet werd hier de veronderstelling gemaakt dat de gebeurtenis zal plaatsvinden in het midden van het zoekvenster. De verwerkingstijd van de tweets zelf is verwaarloosbaar klein. Bijgevolg zal de vertraging geïntroduceerd door het zoekvenster 5, 10, 15 of 30 seconden zijn. Hierbij houden we echter geen rekening met een detectievertraging. De tweets moeten immers een explosie veroorzaken en bovendien moeten voldoende tweets geclassificeerd worden als behorende tot een gebeurtenis. Bijgevolg zal de verwerkingsvertraging gelijk zijn aan de detectievertraging plus de helft van de lengte van het gebruikte zoekvenster.



Figuur 5.9: Opeenvolging van de types shots voor en na het vallen van een goal. Een groen balkje stelt een bovenaanzicht voor en een blauw balkje een close-up. Een rood balkje staat voor een logotransitie.

5.3.1.5 Video-extractievertraging

Wanneer een doelpunt gedetecteerd is in de Twitterberichten, moeten we nog het juiste fragment selecteren. Indien deze detectie gebeurt voor of tijdens de herhaling, moet er gewacht worden tot de herhaling is afgelopen om het fragment op te leveren. Indien de detectie na de herhaling plaatsvindt, hangt dit af van het tijdsverschil tussen de herhaling en de detectie.

5.3.2 Een algoritme om Twitter en videosequenties te combineren

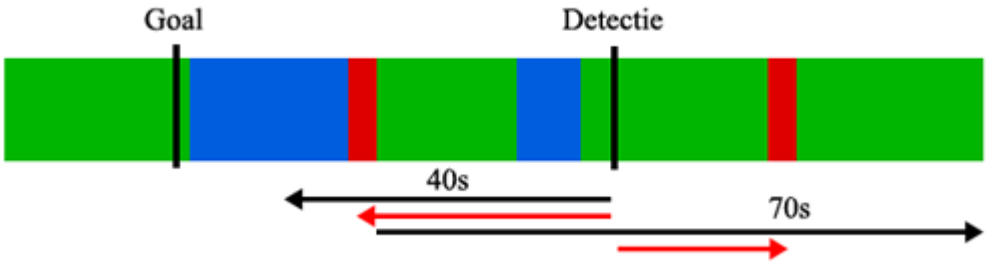
Voor het combineren van de gebeurtenisdetectie in Twitter en de herhalingsdetectie in de videosequentie schatten we de verschillende vertragingen aan de hand van de eerste 2 wedstrijden uit de ontwikkelingsverzameling in Tabel A.3. Een overzicht wordt in Figuur 5.9 gegeven. Voor de videosequentie leiden we af dat de vertraging tussen het scoren van het doelpunt en het starten van de herhaling tussen de 15 en 20 seconden ligt. De herhaling zelf duurt 30 tot 40 seconden. De detectie in Twitter vindt plaats tussen de 10 en de 40 seconden nadat het doelpunt plaatsvindt in de videosequentie. Bijgevolg kunnen we 2 gevallen onderscheiden. De detectie vindt plaats tijdens de herhaling of de detectie vindt plaats voor de start van de herhaling.

5.3.2.1 Geval 1: De doelpuntdetectie vindt plaats tijdens de herhaling

Wanneer een doelpuntdetectie in Twitter plaatsvindt tijdens een herhaling, gaan we eerst op zoek naar de start van de herhaling. We weten dat een doelpunt maximaal 40 seconden na het plaatsvinden ervan gedetecteerd wordt. Doordat er nog minstens 15 seconden zijn tussen het scoren van het doelpunt en de start van de herhaling, is het voldoende om maximaal 40 seconden terug te gaan in de tijd. Dit is ook de maximale lengte van een herhaling. We doen dit telkens in intervallen van 5 seconden.

Wanneer het eerste logo gedetecteerd wordt, gaan we op zoek naar het tweede logo dat het einde van de herhaling aanduidt. We werken opnieuw met intervallen van 5 seconden. Aangezien we met een live-uitzending te maken hebben, moeten we nu telkens wachten tot 5 seconden beeldmateriaal ontvangen is. Om te anticiperen op een herhaling die langer is dan 40 seconden, beschouwen we een interval van 70 seconden gerekend vanaf de start van de herhaling. Het werkelijk nog te onderzoeken interval is dan 70 seconden verminderd met het reeds geanalyseerde deel om het eerste logo te vinden. Wanneer beide logo's gedetecteerd zijn, kan de herhaling uit de video geëxtraheerd worden.

Het algoritme wordt visueel voorgesteld in Figuur 5.10. De zwarte pijlen tonen het bereik van het algoritme en de rode pijlen tonen welk deel daarvan werkelijk zal geanalyseerd worden.



Figuur 5.10: Illustratie van het algoritme om een herhaling in een videosequentie te vinden op basis van een gebeurtenisdetectie in Twitter. De gebeurtenisdetectie vindt plaats tijdens de herhaling.

Soms wordt een herhaling vroegtijdig afgebroken omdat zich opnieuw een belangrijke gebeurtenis aandient. Het tweede logo wordt dan niet getoond in de videosequentie. Indien na 70 seconden nog altijd geen logo gevonden is, maken we gebruik van het shotgrensdetectie-algoritme. Het moment dat de herhaling afgebroken wordt zal gepaard gaan met een abrupte overgang. Aangezien ook het logodetectie-algoritme gebruik maakt van dezelfde berekeningen, moeten we enkel maar de eerste shotgrensdetectie opvragen sinds de detectie van het eerste logo.

5.3.2.2 Geval 2: De doelpuntdetectie vindt plaats voor de start van de herhaling

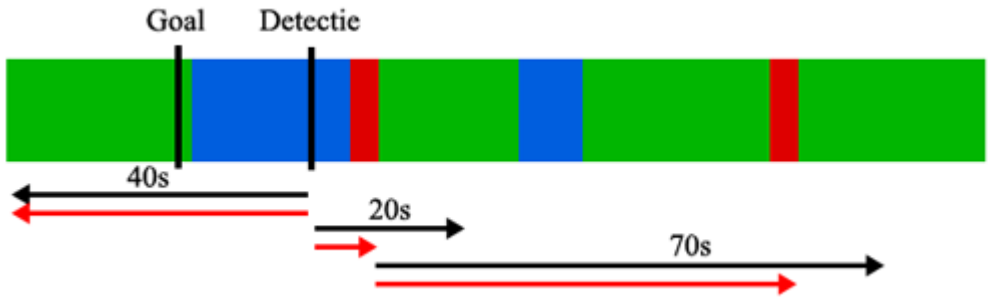
De gebeurtenisdetectie in Twitter vindt niet altijd plaats tijdens een herhaling maar zal meestal plaatsvinden voor de start van de herhaling (zie Sectie 6.4).

Wanneer de gebeurtenisdetectie plaatsvindt voor de herhaling zal het algoritme zoals in het eerste geval, 40 seconden teruggaan in de tijd op zoek naar het logo dat de start van de herhaling aangeeft. Dit keer zal er echter geen logo gevonden worden. Het algoritme

gaat er vervolgens vanuit dat de herhaling nog niet gestart is en zal op zoek gaan naar een logo dat na het moment van de doelpuntdetectie komt. Hiervoor definiëren we een interval van 20 seconden. De lengte is bepaald op basis van de tijd tussen het plaatsvinden van de gebeurtenis in de live-uitzending, de start van de herhaling en de tijd tot de eerste detectie in Twitter.

Het tweede logo zal op dezelfde wijze als in geval 1 gedetecteerd worden. Doordat we in realtime werken kan de analyse niet sneller uitgevoerd worden dan dat de herhaling duurt.

Het algoritme wordt visueel voorgesteld in Figuur 5.11. De zwarte pijlen tonen opnieuw het bereik van het algoritme en de rode pijlen tonen welk deel daarvan werkelijk zal geanalyseerd worden.



Figuur 5.11: Illustratie van het algoritme om een herhaling in een videosequentie te vinden op basis van een gebeurtenisdetectie in Twitter. De gebeurtenisdetectie vindt plaats voor de start van de herhaling.

We kijken altijd eerst in het verleden en dan pas in de toekomst omdat indien de doelpuntdetectie plaatsvindt op het einde van een herhaling, de kans bestaat dat we het tweede logo zouden kunnen detecteren en foutief veronderstellen dat dit het eerste logo van een herhaling is.

5.3.2.3 Finale algoritme

Gedurende de wedstrijd zullen er continu berichtjes geanalyseerd worden. Pas wanneer er een doelpunt gedetecteerd wordt in Twitter zal er gebruik worden gemaakt van de videosequentie. Bij zo een doelpuntdetectie zal er eerst maximaal 40 seconden in het verleden worden gekeken op zoek naar een logo. Indien er geen logo gevonden is, zal er 20 seconden in de toekomst gekeken worden. Indien er dan geen eerste logo is gedetecteerd, wordt er verondersteld dat er geen herhaling heeft plaatsgevonden. Indien er wel een eerste logo gedetecteerd wordt, zal er vanaf het tijdstip van het eerste logo op zoek gegaan worden naar een tweede logo met een maximum van 70 seconden. Wanneer een tweede

logo gevonden wordt, zal de doelpuntherhaling uit de videosequentie worden gehaald en geannoteerd worden met de naam van de speler die het doelpunt scoorde, het tijdstip en de stand na het scoren van dit doelpunt. Indien geen tweede logo gevonden wordt na 70 seconden maken we gebruik van het shotgrensdetectie-algoritme om toch een einde van de herhaling te vinden.

Omdat we over de videobeelden beschikken maken we gebruik van de werkelijke start van de wedstrijd om de minuut van scoren te bepalen en niet het officiële startuur. In praktijk blijkt dit namelijk al snel een minuut te verschillen. Deze taak kan ook geautomatiseerd worden door gebruik te maken van startherkenning in voetbalwedstrijden [33]. Het gebruik van klokherkenning in videosequenties [54] kan zelfs helpen bij het verbeteren van de accuraatheid van het tijdstip waarop de gebeurtenis heeft plaatsgevonden.

Het beschreven algoritme kan ook uitgebreid worden door gebruik te maken van shotclassificatie. Zodoende zouden we ook het fragment kunnen extraheren waarin het doelpunt werkelijk gescoord is op basis van de cinematografische technieken geïntroduceerd in Sectie 5.1.2.

5.4 Besluit

In dit hoofdstuk hebben we een algoritme voorgesteld dat in staat is om herhalingen van doelpunten te extraheren uit videosequenties op basis van Twitter. We hebben hiervoor gebruik gemaakt van de cinematografische kenmerken van live-uitzendingen van voetbalwedstrijden zoals het gebruik van logotransities en abrupte shotovergangen. Het resultaat van het algoritme is een herhalingsfragment van een doelpunt geannoteerd met de naam van de speler die het doelpunt gescoord had, wat de stand was in de wedstrijd na het scoren van het doelpunt en het tijdstip waarop dit doelpunt gescoord werd.

Hoofdstuk 6

Evaluatie

In dit hoofdstuk evalueren we de verschillende algoritmes en hun deelcomponenten die in de vorige 2 hoofdstukken geïntroduceerd werden. We bekijken niet alleen de prestaties in termen van precisie en recall, maar evalueren ook de vertraging die geïntroduceerd wordt.

We definiëren de precisie als het percentage correct gedetecteerde instanties ten opzichte van alle gedetecteerde instanties:

$$precisie = \frac{Aantal\ correct\ gedetecteerde\ instanties}{Totaal\ aantal\ gedetecteerde\ instanties}$$

De recall wordt gedefinieerd als het percentage correct gedetecteerde instanties in een verzameling ten opzichte van alle correcte instanties in die verzameling:

$$recall = \frac{Aantal\ correct\ gedetecteerde\ instanties}{Totaal\ aantal\ correcte\ instanties}$$

Voor de evaluatie maken we gebruik van een uitgebreide, onafhankelijke testverzameling van 15 voetbalwedstrijden uit de Engelse Premier League 2011-2012. Een overzicht hiervan is te vinden in Tabel A.4. Voor elke wedstrijd beschikken we over een aparte verzameling tweets per ploeg en een videosequentie van de wedstrijd. De videosequenties werden verzameld via live-streaming via het internet en via Telenet digitale televisie. Een videosequentie werd met de Twitterberichten gesynchroniseerd door een tweet te verzenden wanneer de opname startte. Alle tweets hebben een tijdsstempel die aangeeft wanneer de Twitterserver de tweet heeft ontvangen. Wanneer we verwijzen naar één van deze wedstrijden zullen we in de rest van dit hoofdstuk enkel nog de namen gebruiken van de 2 ploegen die de wedstrijd spelen. Zij vormen in de testverzameling een unieke combinatie.

We beginnen in Sectie 6.1 met een evaluatie van de individuele stappen van het Twittergebeurtenisdetectiesysteem. In Sectie 6.2 volgt de globale evaluatie van dit systeem. Vervolgens evalueren we in Sectie 6.3 de individuele componenten van het hoogtepunt-

extractiealgoritme. We eindigen met een globale evaluatie van dit algoritme in Sectie 6.4.

6.1 Evaluatie van de individuele stappen van het Twittergebeurtenisdetectiesysteem

In Hoofdstuk 4 introduceerden we een algoritme dat in staat is gebeurtenissen te detecteren in Twitter. Het algoritme bestaat uit 6 grote deelstappen: het filteren van de tweets, piekdetectie, classificatie van tweets, gebeurtenisdetectie, gebeurtenisidentificatie en extractie van gedetailleerde informatie. In de volgende secties evalueren we het filteren van tweets, classificatie van tweets, gebeurtenisdetectie en gebeurtenisidentificatie. De laatste stap wordt niet individueel geëvalueerd maar wordt pas onder de loep genomen in Sectie 6.2 waar we het globale systeem evalueren. Verder zullen we ook de piekdetectie niet individueel evalueren omdat dit zou vereisen dat we voor elk van de 15 wedstrijden voor elke seconde nagaan of er een gebeurtenis wordt beschreven in een venster van 10, 20, 30 of 60 seconden.

6.1.1 Filteren van tweets

De eerste stap die plaatsvindt is het filteren van de tweets. We hebben de tweets die de letters 'RT', een @-symbool of een URL bevatten verwijderd. De resultaten van deze reductie worden weergegeven in Tabel 6.1. Voor elke individuele hashtag wordt het aantal tweets na filtering gemiddeld gereduceerd tot 53% van het oorspronkelijk aantal tweets.

6.1.2 Classificatie van tweets

De gebeurtenisdetectie en -identificatie maken beiden gebruik van het resultaat van de classificatiestap om een gebeurtenis te detecteren en vervolgens te identificeren. We evalueren daarom eerst de gekozen kenmerken van de tweetvector om daarna de prestaties van de Support Vector Machine (SVM) te evalueren als onderdeel van de gebeurtenisdetectie en -identificatie.

In Sectie 4.2.3.1 hebben een tweetvector gedefinieerd die gebruikt wordt door de SVM om elke tweet in 1 van de 7 klassen in te delen: 'doelpunt', 'strafschop', 'fout', 'wissel', 'einde eerste helft', 'einde tweede helft' en 'geen gebeurtenis'. De verschillende kenmerken werden gekozen op basis van de karakteristieken die een tweet, die tot 1 van de 7 klassen behoorde, al dan niet vertoonde.

Tabel 6.1: Overzicht van het totaal aantal verzonden tweets per wedstrijd voor en na de filterstap over een tijdsspanne van twee uur vanaf de start van de wedstrijd.

	Voor		**Na**	
Thuis - Uit	**Thuis**	**Uit**	**Thuis**	**Uit**
Fulham - Wolverhampton Wanderers	2621	1951	1536	1305
Tottenham Hotspurs - Manchester United	7403	31359	3791	19099
Arsenal - Chelsea	6519	16944	3915	6330
Arsenal - Wigan Athletic	9175	5315	6100	3558
Blackburn Rovers - Manchester United	3499	35445	1634	21800
Chelsea - Queens Park Rangers	38977	13729	16579	5071
Chelsea - Wigan Athletic	11932	891	4813	582
Everton - Sunderland	1193	1584	733	750
Fulham - Chelsea	2841	12238	1664	6318
Manchester City - Manchester United	31832	40977	16655	22784
Manchester City - Sunderland	9130	4961	3977	2169
Manchester United - Aston Villa	33896	12382	17789	4867
Manchester United - Queens Park Rangers	30049	11360	16244	5208
Stoke City - Arsenal	744	3018	469	1852
Tottenham Hotspurs - Blackburn Rovers	3164	2193	1787	1187
Totaal	**387322**		**200566**	
Gemiddeld	**12910,73**		**6685,53**	

We maken gebruik van Weka [52] om de kenmerken te evalueren. We willen de kenmerken rangschikken volgens bijdrage tot de correcte classificatie van tweets door net zoals Barbosa en Feng [12] gebruik te maken van de informatiewinst. De informatiewinst blijkt een uitermate geschikt criterium te zijn voor het evalueren van de prestaties van een SVM [55, 56]. Omdat we zowel binaire kenmerken hebben als kenmerken met een groot aantal numerieke waarden kan dit bij de evaluatie een vertekend beeld geven [57]. We maken daarom gebruik van de winstratio om de kenmerken te rangschikken. Dit is de genormaliseerde versie van de informatiewinst [57]:

$$Winstratio(T, a) = \frac{H(T) - H(T|a)}{-\sum_{i=1}^{n} \frac{T_i}{T} * log_2 \frac{T_i}{T}}$$

Hierbij is T een verzameling van tweets met hun tweetvector en a een kenmerk van de tweetvector. De functie H is de entropiefunctie. Het kenmerk a kan n verschillende waarden aannemen binnen de verzameling T. T_j is bijgevolg een deelverzameling van T waar a slechts 1 waarde aanneemt.

De winstratio na toevoeging van een kenmerk wordt berekend door gebruik te maken van 10-wegse kruisvalidatie (Engels: 10-fold cross-validation). We maken hiervoor gebruik van de trainingsverzameling van de SVM die 1324 geannoteerde tweets bevat uit de wedstrijden in Tabel A.2.

De rankschikking van de kenmerken volgens de winstratio is te vinden in Tabel 6.2. De beste kenmerken zijn de contextwoorden die tot de zeldzamere klassen behoren. Daarna volgen verschillende sentimentkenmerken zoals herhaling en hoofdlettergebruik in contextwoorden of spelersnamen. Een scorevermelding blijkt ook een sterk kenmerk te zijn. Het gebruik van spelersnamen blijkt niet zo een sterk kenmerk te zijn tenzij de spelersnaam wordt uitgedrukt in hoofdletters. Dit kan verklaard worden doordat een vermelding van een spelersnaam zonder sentimentkenmerken een kenmerk is van meerdere klassen. De laagst gerangschikte kenmerken zijn opnieuw kenmerken die in meerdere klassen thuishoren. De lengte van de tweet in woorden of karakters blijkt een doorsnee kenmerk te zijn.

Tabel 6.2: Rangschikking van de verschillende kenmerken volgens bijdrage tot de classificatie van de tweets. Als criterium werd de winstratio gebruikt. Op rang 1 staat het meest onderscheidende kenmerk en op rang 23 het minst onderscheidende kenmerk.

	Kenmerk		**Kenmerk**
1	Contextwoord einde tweede helft	13	Hoofdlettergebruik
2	Contextwoord wissel	14	Herhaling letters in spelersnaam
3	Contextwoord einde eerste helft	15	Lengte tweet
4	Contextwoord fout	16	Gebruik uitroepteken
5	Herhaling letters in contextwoord	17	Gebruik achternaam van speler
6	Hoofdlettergebruik in contextwoord	18	Gebruik voornaam van speler
7	Contextwoord strafschop	19	Gebruik initialen van speler
8	Score vermeld	20	Herhaling vraagtekens
9	Hoofdlettergebruik in spelersnaam	21	Herhaling letters in andere woorden
10	Herhaling uitroepteken	22	Hoofdlettergebruik in andere woorden
11	Contextwoord doelpunt	23	Herhaling punctuaties
12	Aantal woorden per tweet		

6.1.3 Gebeurtenisdetectie

Om de gebeurtenisdetectie correct te evalueren, zouden we in het ideale geval alle tweets van de 15 wedstrijden uit de testverzameling moeten annoteren en nagaan of hetzelfde resultaat behaald wordt met een SVM. Omdat dit zou betekenen dat we bijna 200.000 tweets moeten annoteren hebben we besloten slechts een deel van de tweets te annoteren, verspreid over de 15 wedstrijden. We hebben telkens op semiwillekeurige wijze 5 minuten per wedstrijd geselecteerd en geannoteerd zodat elke gebeurtenis minstens eenmaal voorkwam. Dit geeft een totaal van 4.500 te evalueren seconden. Een overzicht is te vinden in Tabel 6.3. De aangegeven start en einde worden uitgedrukt in wedstrijdminuten. Telkens werden de gefilterde tweets van de thuisploeg geannoteerd met uitzondering van de wedstrijd Arsenal - Chelsea waar de hashtag van Chelsea werd gebruikt. In de kolom 'gebeurtenis' staat telkens aangegeven welke gebeurtenis geïdentificeerd werd tijdens

de gebeurtenisidentificatie. Er vonden geen niet-geïdentificeerde gebeurtenissen plaats tijdens de geselecteerde minuten. In totaal werden 6.757 tweets geannoteerd.

Tabel 6.3: Overzicht van de geselecteerde minuten per wedstrijd voor de gebeurtenisdetectie en -identificatie. In de kolom 'gebeurtenis' worden de geïdentificeerde gebeurtenis weergegeven.

Thuis - Uit	**Start**	**Einde**	**Gebeurtenis**
Arsenal - Chelsea	80	85	Fout (Gele kaart)
Arsenal - Wigan Athletic	80	85	
Blackburn - Manchester United	80	85	Doelpunt
Chelsea - QPR	90	95	Einde tweede helft
Chelsea - Wigan Athletic	90	95	Doelpunt
Everton - Sunderland	90	95	
Fulham - Wolverhampton	70	75	
Fulham - Chelsea	70	75	Wissel
Manchester City - Manchester United	70	75	
Manchester City - Sunderland	5	10	
Manchester United - Aston Villa	5	10	Strafschop + doelpunt
Manchester United - QPR	5	10	
Stoke City - Arsenal	45	50	
Tottenham Hotspurs - Blackburn Rovers	45	50	Einde eerste helft
Tottenham Hotspurs - Manchester United	45	50	

Voor de evaluatie van de gebeurtenisdetectie beschouwen we 2 klassen: 'gebeurtenis' en 'geen gebeurtenis'. We maken geen onderscheid tussen detecties die plaatsvinden met een zoekvenster van 10, 20, 30 of 60 seconden. De resultaten worden weergegeven in Tabel 6.4. Hierbij stelt 'aantal' respectievelijk het aantal piekdetecties voor die niet leiden tot een gebeurtenisdetectie en het aantal piekdetecties die wel tot een gebeurtenisdetectie leiden.

Tabel 6.4: Prestatie van de gebeurtenisdetectie voor de testverzameling uit Tabel 6.3.

	Testverzameling	
	Geen gebeurtenis	**Gebeurtenis**
Aantal	1927	209
Fout	20	31
Gemist	31	20
Precisie	98,96%	85,17%
Recall	98,40%	89,90%

Door het overwicht aan piekdetecties die niet tot een gebeurtenis leiden, is zowel de precisie als recall hoger dan 98%. Voor de gebeurtenisdetectie zelf bekomen we een precisie van ongeveer 85% en een recall van bijna 90%. Dit resultaat ligt in lijn met de bevindingen in [19] waar de auteurs gebeurtenissen proberen te detecteren in een cricketwedstrijd na afloop van de wedstrijd. De bedoeling was hier echter om individuele tweets te classificeren. Wij proberen verzamelingen tweets in te delen in deze 2 klassen. De auteurs rapporteren

een precisie en recall van respectievelijk 87% en 89% in het geval van 'geen gebeurtenis' en een precisie en recall van respectievelijk 85% en 86% voor het geval 'gebeurtenis'.

We kunnen besluiten dat de gebeurtenisdetectiestap succesvol gebeurtenissen kan detecteren met een precisie hoger dan 85% en een recall van bijna 90%.

6.1.4 Gebeurtenisidentificatie

Voor de gebeurtenisidentificatie maken we gebruik van dezelfde testverzameling als voor de gebeurtenisdetectie weergegeven in Tabel 6.3. We gaan er vanuit dat een gebeurtenis reeds correct gedetecteerd is. Dit betekent dat de gebeurtenis gedetecteerd wordt door zowel het detectiesysteem dat gebruik maakt van een SVM om de tweets te classificeren als het detectiesysteem dat gebruik maakt van de geannoteerde tweets. Slechts 6 wedstrijden voldoen aan dit criterium. Een overzicht wordt gegeven in Tabel 6.5. We merken op dat de 'gele kaart'-gebeurtenis ook verwijderd is. Dit komt omdat bij de gebeurtenisdetectie met behulp van de SVM, enkele tweets fout waren geclassificeerd waardoor er toch een gebeurtenis werd gedetecteerd.

Tabel 6.5: Overzicht van de geselecteerde minuten per wedstrijd voor de gebeurtenisdetectie en -identificatie. In de kolom gebeurtenis worden de geïdentificeerde gebeurtenissen weergegeven.

Thuis - Uit	**Start**	**Einde**	**Gebeurtenis**
Blackburn - Manchester United	80	85	Doelpunt
Chelsea - QPR	90	95	Einde tweede helft
Chelsea - Wigan Athletic	90	95	Doelpunt
Fulham - Chelsea	70	75	Wissel
Manchester United - Aston Villa	5	10	Strafschop + doelpunt
Tottenham Hotspurs - Blackburn Rovers	45	50	Einde eerste helft

In totaal werden 178 gebeurtenissen correct gedetecteerd. Hiervan werden 177 gebeurtenissen correct geïdentificeerd als 1 van de 6 gebeurtenissen en 1 gebeurtenis werd correct geïdentificeerd als 'geen gebeurtenis'. Een overzicht van de verdeling wordt gegeven in Tabel 6.6. Voor de verschillende klassen leidt dit tot een precisie en recall van 100%.

Tabel 6.6: Verdeling van de geïdentificeerde klassen in de gebeurtenisidentificatiestap.

Klasse	**Aantal**
Geen gebeurtenis	1
Doelpunt	125
Strafschop	27
Fout	0
Wissel	18
Einde eerste helft	6
Einde tweede helft	1
Totaal	**178**

Een geval van falen Alhoewel de gebeurtenisidentificatie een recall en precisie van 100% heeft, konden we één geval van falen terugvinden. In de wedstrijd Manchester United - Queens Park Rangers vindt er een strafschop plaats waarbij een speler een rode kaart ontvangt en de strafschop bovendien tot een doelpunt leidt. Dit zijn 3 verschillende gebeurtenissen door elkaar. Er worden 62 gebeurtenissen gedetecteerd met een precisie en recall van respectievelijk 90,32% en 98,25%. De identificatie faalt echter. Slechts 12 van deze 62 gebeurtenissen konden toegewezen worden aan 1 van de 3 gebeurtenissen. Dit falen wordt veroorzaakt doordat deze meervoudige gebeurtenis niet geanticipeerd was bij de ontwikkeling. Er was enkel rekening gehouden met het tegelijk voorkomen van doelpunten en strafschoppen. Door het vaak voorkomen van tweets die als rode kaart worden geclassificeerd, halen slechts weinig gebeurtenissen de vereiste drempel van 0,4. Alle 3 de gebeurtenissen werden echter minstens eenmaal correct geïdentificeerd waardoor dit in praktijk geen probleem vormt. Dit toont enerzijds de kracht van het identificatiealgoritme, maar toont anderzijds ook aan dat er nog ruimte is voor verbetering op het vlak van meervoudige gebeurtenissen.

6.2 Globale evaluatie van het Twittergebeurtenisdetectiesysteem

In deze sectie evalueren we de globale prestaties van het Twittergebeurtenisdetectiesysteem. We evalueren voor de 6 gebeurtenissen telkens de algemene detectie- en identificatieprestaties als de vertraging tussen de detectie en het gebeuren van de gebeurtenis in de videosequentie. Voor de evaluatie gebruiken we de data van de website Soccerway [47] als controledata. De vertraging wordt gemeten door gebruik te maken van de opgenomen videosequenties die gesynchroniseerd zijn met de Twitterberichtstromen.

6.2.1 Het doelpunt

De testverzameling van 15 wedstrijden bevat in totaal 47 doelpunten. Van deze 47 doelpunten worden alle 47 in minstens 1 van de 2 Twitterberichtstromen gedetecteerd. We leggen ons daarom toe op de correcte extractie van de informatie. Het is immers van cruciaal belang dat een doelpunt niet enkel gedetecteerd wordt, maar dat er ook kan geïdentificeerd worden welke ploeg gescoord heeft. Het resultaat van een doelpuntdetectie is van de vorm:

> "Gescoorde minuut: naam speler of ploeg (stand na scoren)".

De spelersnaam of ploegnaam wordt bepaald op basis van de stand. We beginnen daarom met de evaluatie van de score of, met andere woorden; het doelpunt aan de juiste ploeg

werd toegekend. Daarna evalueren we of telkens de correcte speler en de correcte minuut worden weergegeven. We eindigen met een evaluatie van de detectievertraging.

Doordat we van elke wedstrijd de tweets hebben verzameld voor de hashtag van beide ploegen laat dit ons toe om beide Twitterberichtstromen te combineren. We evalueren daarom 4 doelpuntdetectiestrategieën:

- **Alle:** Dit is de algemene evaluatiemethode waarbij we geen onderscheid maken tussen welke berichtenstroom welk doelpunt gedetecteerd heeft, maar elke detectie apart op zijn correctheid evalueren.
- **Eerste:** Wanneer we deze strategie toepassen beschouwen we enkel de eerste detectie van een doelpunt als de correcte detectie en besteden geen aandacht aan een eventuele tweede detectie.
- **Eigendoelpunt:** Bij deze strategie wordt enkel gekeken naar de doelpunten die gescoord werden door de ploeg waarvan we de berichtenstroom beschouwen. In het geval van de wedstrijd Arsenal - Chelsea zal een doelpunt dat gedetecteerd werd als gescoord door Chelsea, genegeerd worden als dit doelpunt gedetecteerd werd in de berichtenstroom van Arsenal.
- **Tegendoelpunt:** Dit is de complementaire strategie van de 'eigendoelpunt'-strategie. In dit geval wordt enkel rekening gehouden met een doelpunt gedetecteerd in de berichtenstroom van de ploeg die het doelpunt tegen krijgt.

6.2.1.1 Evaluatie van de scorevermelding

In Tabel 6.7 wordt een overzicht gegeven van de resultaten voor doelpuntdetectie voor de 4 verschillende strategieën. Een doelpunt wordt pas als correct beschouwd indien de stand correct wordt weergegeven. In totaal vonden 67 verschillende doelpuntdetecties plaats waarbij 2 detecties een foute stand aangaven en 1 doelpunt in geen van beide stromen kon gedetecteerd worden met een *juiste* stand.

De foutgedecteerde doelpunten en het gemiste doelpunt werden veroorzaakt door een foutgelopen eerste doelpuntdetectie tijdens de wedstrijd Arsenal - Wigan Athletic in de berichtenstroom van Wigan Athletic. Het gebeurtenisdetectiesysteem gaat er vanuit dat elk eigen gescoord doelpunt correct gedetecteerd kan worden en dat eigen doelpunten enkel puntsgewijs kunnen verhoogd worden. In de berichtenstroom van Wigan Athletic werd een doelpunt fout gedetecteerd doordat een zeer goede kans voor Arsenal gevolgd werd door een doelpunt van Wigan Athletic. Hierdoor was het zoekvenster waarop de analyse gebeurde, gevuld met tweets die de speler Y. Benayoun vernoemden en heel korte tweets die voornamelijk de score vermeldden. Bijgevolg werd een eerste doelpunt gedetecteerd.

Tabel 6.7: Prestatie van het finale gebeurtenisdetectiesysteem op doelpunten voor 4 verschillende strategieën waarbij een doelpunt correct gedetecteerd wordt indien de weergegeven stand correct is.

	Detectiestrategie			
	Alle	**Eerste**	**Eigendoelpunt**	**Tegendoelpunt**
Doelpuntdetecties	67	48	45	22
Fout	2	2	0	2
Gemist	29	1	2	27
Correct	65	46	45	20
Precisie	97,01%	95,83%	100,00%	90,91%
Recall	56,76%	97,87%	95,74%	42,55%

Deze eerste detectie zorgde ervoor dat het onmogelijk werd om het tweede doelpunt dat door Wigan Athletic gescoord werd ook toe te wijzen aan Wigan Athletic. In dat geval zou de stand van 1-0 naar 0-2 wijzigen. Beide doelpunten werden bijgevolg fout gedetecteerd. We merken op dat doordat het doelpunt toegekend werd aan de ander ploeg dit geen fout gedetecteerd eigendoelpunt is maar een fout gedetecteerd tegendoelpunt. Dit is wel een gemist eigendoelpunt. Het resultaat van de doelpuntdetectie is te vinden in Tabel 6.8. De 0-2 kan in geen van beide berichtstromen correct gedetecteerd worden.

Tabel 6.8: Resultaat van de doelpuntdetectie in de wedstrijd Arsenal - Wigan Athletic.

Ploeg	**Fout**	**Correct**
Wigan Athletic	6': Y. Benayoun scores! (1-0)	7': F. Di Santo scores! (0-1)
Arsenal	7': F. Di Santo scores! (0-1)	7': F. Di Santo scores! (0-1)
Wigan Athletic	9': Arsenal scores! (2-0)	8': J. Gomez scores! (0-2)
Arsenal	21': T. Vermaelen scores! (1-2)	21': T. Vermaelen scores! (1-2)

De precisie en recall van de strategie 'eerste' en de strategie 'eigendoelpunt' is in beide gevallen hoger dan 95%. De strategie 'tegendoelpunt' heeft een beduidend lagere recall. Dit is niet onlogisch aangezien niet altijd in beide berichtstromen het doelpunt gedetecteerd werd.

6.2.1.2 Evaluatie van de spelersvermelding

Op basis van de ploeg die het doelpunt krijgt toegekend, wordt ook een spelersnaam of ploegnaam afgeleid. Bij alle 65 correct gedetecteerde doelpuntdetecties werd telkens de correcte spelersnaam geëxtraheerd.

6.2.1.3 Evaluatie van de minuutvermelding

De laatste component die we kunnen evalueren is de minuut waarin het doelpunt gescoord wordt. De weergegeven minuut wordt bepaald onafhankelijk van het doelpunt. We verge-

lijken voor elk doelpunt de weergegeven minuut met de minuut opgegeven door de website Soccerway [47]. Het verschil beperkt zich in ons geval tot geen, 1 minuut, 2 minuten of 3 minuten verschil. Omdat foutgedetecteerde doelpunten geen controledata hebben, worden deze hier niet beschouwd. Voor de berekening van de minuut werd telkens gebruik gemaakt van de werkelijke start van de wedstrijd in de videosequentie.

In Tabel 6.9 wordt het resultaat van deze evaluatie weergegeven voor alle 4 de strategieën. De eerste rij geeft telkens weer hoeveel doelpuntdetecties de correcte minuut weergeven of 1 minuut, 2 minuten of 3 minuten verschil tonen. De tweede rij geeft de fractie weer dat een verschil vertegenwoordigt van de correct gedetecteerde doelpunten.

Tabel 6.9: Prestatie van de minuutberekening bij de weergave van een gescoord doelpunt.

	Aantal minuten verschil			
	0	**1**	**2**	**3**
Alle	50 76,92%	14 21,54%	0 0,00%	1 1,54%
Eerste	39 84,78%	7 15,22%	0 0,00%	0 0,00%
Eigendoelpunt	36 80,00%	9 20,00%	0 0,00%	0 0,00%
Tegendoelpunt	14 70,00%	5 25,00%	0 0,00%	1 5,00%

We stellen vast dat de eerste doelpuntdetectiestrategie een beduidend beter resultaat vertoont dan de andere strategieën. Bij bijna 85% van de correct gedetecteerde doelpunten, wordt ook de correcte minuut vermeld. Dit kan verklaard worden doordat de snelheid van detectie bij het weergeven van de correcte minuut, de bepalende factor is. De overige 15% vertoont bovendien slechts 1 minuut verschil. In tegenstelling tot de strategie 'alle' en de strategie 'tegendoelpunt' bevatten de strategieën 'eerste' en 'eigendoelpunt' geen weergaven die 2 of 3 minuten verschillen met de controledata.

6.2.1.4 Evaluatie van de snelheid van de doelpuntdetectie

Om de snelheid van de detectie te meten hebben we gebruik gemaakt van de bijhorende videosequenties. Voor elk doelpunt hebben we gemeten wanneer het doelpunt gescoord werd in de wedstrijd, wanneer het eerste bericht dat het doelpunt rapporteerde verscheen op Twitter, wanneer er een explosie van berichten plaatsvond en wanneer finaal de detectie plaatsvond. We hebben dit voor alle 4 de strategieën gedaan. De resultaten hiervan worden getoond in Tabel 6.10 en Tabel 6.11. Telkens wordt het verschil aangegeven tussen het tijdstip dat het doelpunt in de videosequentie zichtbaar was en het verschijnen van het eerste bericht, de explosie of de detectie in Twitter.

Tabel 6.10: Statistiek van de vertraging ten opzichte van het scoren van het doelpunt in de videosequentie voor de strategieën 'alle' en 'eerste'. Alle waarden zijn uitgedrukt in seconden.

	Alle			**Eerste**		
	Eerste bericht	**Explosie**	**Detectie**	**Eerste bericht**	**Explosie**	**Detectie**
Gemiddelde	7,25	14,85	24,29	6,20	12,80	20,91
Mediaan	6	16	22	5	10	18
Standaard afw.	7,78	11,52	15,59	7,81	12,44	15,64
[0%,75%]	13	22	38	13	18,75	32
[0%,90%]	16,6	27,4	46,6	15,5	25	43
Min	-9	-8	-5	-9	-8	-5
Max	25	54	56	25	54	56

Tabel 6.11: Statistiek van de vertraging ten opzichte van het scoren van het doelpunt in de videosequentie voor de strategieën 'eigendoelpunt' en 'tegendoelpunt'. Alle waarden zijn uitgedrukt in seconden.

	Eigendoelpunt			**Tegendoelpunt**		
	Eerste bericht	**Explosie**	**Detectie**	**Eerste bericht**	**Explosie**	**Detectie**
Gemiddelde	6,04	12,42	21	9,95	20,3	31,7
Mediaan	4	9	18	11	21	29
Standaard afw.	7,83	12,27	15,83	7,12	7,32	12,44
[0%,75%]	13	19	32	16	25	40
[0%,90%]	15,6	23,6	44	17,4	30	47,4
Min	-9	-8	-5	-4	9	12
Max	25	54	56	22	33	55

Opnieuw blijken de beste strategieën de strategie 'eerste' en 'eigendoelpunt'. De gemiddelde detectievertraging is in beide gevallen rond de 21 seconden ten opzichte van 24 en 31 seconden voor respectievelijk alle doelpuntdetecties en de tegendoelpuntdetecties. 90% van de doelpuntdetecties vindt plaats minder dan 45 seconden nadat het doelpunt heeft plaatsgevonden in de videosequentie voor de eerste 3 strategieën. De vertraging tussen het eerste bericht en de doelpuntdetectie is voor de eerste en de eigendoelpuntdetectiestrategie in beide gevallen ongeveer 15 seconden en de vertraging tussen de explosie van berichten die het doelpunt rapporteren en de doelpuntdetectie is telkens ongeveer 9 seconden. Voor de tegendoelpuntdetectiestrategie blijkt de late detectie grotendeels te wijten aan de latere explosie. We kunnen immers pas een doelpunt detecteren als ook voldoende berichten dit doelpunt rapporteren.

6.2.1.5 Besluit

We kunnen besluiten dat de beste techniek om doelpunten te detecteren de techniek is die gebruik maakt van beide Twitterberichtstromen en telkens de eerste detectie als correct beschouwd. Deze strategie presteert even goed in het detecteren van doelpunten als de eigendoelpuntenstrategie, maar detecteert meer correcte doelpunten en heeft een betere tijdsaanduiding. Op het vlak van snelheid presteert deze techniek even goed als de eigendoelpuntdetectiestrategie. Bijgevolg is de eerste doelpuntdetectiestrategie de beste strategie. De strategie 'alle' is slechts nuttig vanuit een theoretisch standpunt. De strategie 'tegendoelpunt' presteert ondermaats in het detecteren van doelpunten.

6.2.2 De strafschop

De testverzameling wedstrijden bevat in totaal 4 verschillende strafschoppen. Alle 4 de strafschoppen werden gedetecteerd door de ploeg die de strafschop kreeg toegekend en 1 strafschop werd ook gedetecteerd door de tegenpartij. De resultaten van alle strafschopdetecties in de volledige testverzameling worden weergegeven in Tabel 6.12. De lage precisie wordt veroorzaakt door 3 strafschopdetecties waar een dubieuze fase geen strafschop tot gevolg had en 3 andere foute strafschopdetecties werden gedetecteerd nadat de strafschop had plaatsgevonden.

Om het hoofd te bieden aan de lage precisie hadden we reeds in Sectie 4.3.3.2 de regel ingevoerd dat we enkel rekening houden met strafschopdetecties die binnen de 2 minuten tot een doelpunt leiden. Om deze regel te evalueren hebben we de tijd gemeten tussen elke strafschopdetectie en de eerstvolgende gebeurtenis. Bij de 5 correcte detecties waren er gemiddeld 58 seconden tussen het toekennen van de strafschop en het scoren van de strafschop met een minimum van 42 seconden en met een maximum van 88 seconden. Bij de 6 foute strafschopdetecties kwam er in 2 gevallen geen gebeurtenis na de strafschop.

Tabel 6.12: Prestatie van de strafschopdetectie.

	Alle
Detecties	11
Fout	6
Gemist	0
Correct	5
Precisie	45,45%
Recall	100,00%

Gemiddeld duurde dit 39 minuten en 53 seconden met een minimum van 5 minuten 5 seconden en een maximum van 55 minuten 12 seconden. Na invoering van deze 2-minutenregel werden bijgevolg alle strafschoppen correct toegekend aan een doelpunt met een precisie en recall van 100%.

We hebben ook de snelheid waarmee strafschoppen gedetecteerd werden, gemeten. Opnieuw hebben we de vertraging gemeten van het eerste bericht dat de strafschop rapporteert, de vertraging van de explosie van berichten die de strafschop rapporteert en de vertraging waarmee de strafschop gedetecteerd werd ten opzichte van het toekennen van de strafschop in de videosequentie. De resultaten zijn te vinden in Tabel 6.13.

Tabel 6.13: Statistiek van de vertraging ten opzichte van het toekennen van een strafschop in de videosequentie van alle correct gedetecteerde strafschoppen. Alle waarden zijn uitgedrukt in seconden.

	Eerste	**Explosie**	**Detectie**
Gemiddelde	1,8	6,6	11,8
Mediaan	0	6	10
Standaard afwijking	5,26	5,55	7,05
[0%,75%]	1	6	19
[0%,90%]	7	12	19
Min	-2	2	3
Max	11	16	19

We merken op dat zowel de vertraging van de eerste berichten als de vertraging van de explosie en de detectie kleiner is dan bij de doelpunten. Dit komt ondermeer omdat we hier als tijdstip in de videosequentie het tijdstip namen waarin voor het eerst de scheidsrechter zichtbaar was die de strafschop toekende. Vele tweets werden al verzonden wanneer de fout in het strafschopgebied begaan werd zonder dat men zeker was dat er een strafschop werd toegekend.

We kunnen besluiten dat de strafschopdetectie bijzonder succesvol is na toepassing van de 2-minutenregel. Alle strafschoppen werden in dat geval succesvol gedetecteerd zonder fout-positieven. We hebben echter geen oplossing voor het zeldzame geval waarbij een strafschop gemist wordt.

6.2.3 De fout

De evaluatie van de foutdetectie bestaat uit 2 vragen: "Wordt de fout correct gedetecteerd?" en "Is de geëxtraheerde informatie correct?". Bij de foutdetectie omvat deze informatie de minuut waarin de fout plaatsvond en het type fout: een fout zonder kaart, een gele kaart of een rode kaart. In de volledige testverzameling werden er 17 fouten gedetecteerd. Dit staat in schril contrast met de 306 fouten die in de 15 wedstrijden werden gefloten. Van deze 306 fouten hebben 50 fouten tot een gele kaart geleid en 1 tot een rode kaart. Het heeft dus niet veel zin om telkens de recall te evalueren. De oorzaak hiervan werd reeds aangetoond in Sectie 4.2.5. Twittergebruikers tweeten enkel over wat ze zelf interessant vinden. Dit werd ook vastgesteld door de auteurs van [18] waar de minst belangrijke gebeurtenis ook beduidend slechter kon gedetecteerd worden dan de meest belangrijke gebeurtenis.

In Tabel 6.14 wordt een overzicht gegeven van de prestaties van de foutdetectie. De eerste kolom bevat de prestaties van alle gedetecteerde fouten. Elke gedetecteerde fout was gerelateerd aan een fout. Bijgevolg is de precisie 100%. In Sectie 4.3.3.3 hebben we een 90-secondenregel geïntroduceerd waarbij slechts 1 fout kan gedetecteerd worden over een tijdsspanne van 90 seconden. Wanneer na deze 90-seconden nog een fout wordt gedetecteerd die dezelfde fout rapporteert, wordt deze detectie als incorrect beschouwd.

We evalueren we 2 strategieën:

- **De individuele 90-secondenregel:** We passen de 90-secondenregel toe op elke individuele berichtenstroom.
- **De gecombineerde 90-secondenregel:** We passen de 90-secondenregel toe wanneer in één van beide berichtstromen een fout is gedetecteerd.

De resultaten worden weergegeven in Tabel 6.14. Beide strategieën presteren evengoed en maken beiden dezelfde fout.

Tabel 6.14: Prestatie van de foutdetectie voor de algemene foutdetectie en 2 verschillende strategieën.

	Foutdetectie		
	Alle	**Individuele 90s-regel**	**Gecombineerde 90s-regel**
Foutdetecties	17	12	10
Fout	0	1	1
Correct	17	11	9
Precisie	100,00%	91,67%	90,00%

Wanneer de fout correct gedetecteerd is kunnen we overgaan tot de evaluatie van de identificatie van het type fout. De resultaten worden weergegeven in Tabel 6.15. De

gecombineerde detectie vertoont een fout minder dan de individuele detectie. We merken op dat indien de juiste kaart niet kan afgeleid worden, geen kaart wordt getoond. Bijgevolg zijn de 3 fouten in de kolom 'geen kaart' alle 3 fouten waarvoor niet kon afgeleid worden dat het om een gele kaart ging. De enige echte identificatiefout is een fout zonder kaart die als fout met een gele kaart wordt aanzien bij de individuele 90-secondenregel. Dit wordt veroorzaakt door tweets zoals "That should have been a yellow card!".

Tabel 6.15: Prestatie van de identificatie van het type fout.

	Foutidentificatie (Kaart)					
	Individuele 90s-regel			**Gecombineerde 90s-regel**		
	Geen	**Geel**	**Rood**	**Geen**	**Geel**	**Rood**
Aantal	4	6	1	4	4	1
Fout	3	1	0	3	0	0
Correct	1	5	1	1	4	1
Precisie	25,00%	83,33%	100,00%	25,00%	100,00%	100.00%

De tweede component van de identificatie die we kunnen evalueren is de minuut waarin de fout plaatsvond. De resultaten voor beide strategieën worden weergegeven in Tabel 6.16. In beide gevallen is het verschil beperkt tot maximaal 1 minuut. De gecombineerde 90-secondenregel presteert beter dan de individuele 90-secondenregel. Dit resultaat kan opnieuw verklaard worden door het feit dat we bij de gecombineerde aanpak telkens de snelste foutdetectie gebruiken.

Tabel 6.16: Prestatie van de minuutberekening bij de weergave van een gedetecteerde fout.

	Individuele 90s-regel		**Gecombineerde 90s-regel**	
	0	**1**	**0**	**1**
Aantal	8	3	7	2
Percentage	72,73%	27,27%	77,78%	22,22%

Het tweede luik van de evaluatie van de foutdetectie omvat de evaluatie van de snelheid waarmee gebeurtenissen gedetecteerd worden. De resultaten zijn te vinden in Tabel 6.17. De foutdetectie door middel van de gecombineerde 90-secondenregel is duidelijk sneller dan de foutdetectie door middel van de individuele 90-secondenregel. Dit is logisch aangezien de kans tot detectie hoger is. De snelheid van detectie is in beide gevallen trager dan bij doelpuntdetectie. Supporters vinden immers een doelpunt een belangrijkere gebeurtenis dan een fout.

We kunnen besluiten dat foutdetectie door middel van de gecombineerde 90-secondenregel zeer goede resultaten geeft met een hoge accuraatheid. Door het combineren van beide berichtstromen kunnen foute identificaties geëlimineerd worden. Ondanks dat we gebonden zijn aan de wil tot tweeten van de Twittergebruikers kan de recall en precisie verhoogd worden door meer geavanceerde analyses uit te voeren op de tweets.

Tabel 6.17: Statistiek van de vertraging ten opzichte van het bekeuren door de scheidsrechter in de videosequentie. Alle waarden zijn uitgedrukt in seconden.

	Individuele 90s-regel			**Gecombineerde 90s-regel**		
	Eerste	**Explosie**	**Detectie**	**Eerste**	**Explosie**	**Detectie**
Gemiddelde	10,55	18,91	31,64	9,44	16,33	27
Mediaan	11	17	31	9	16	26
Standaard afw.	6,19	7,88	12,75	6,31	5,39	7,92
[0%,75%]	15,5	23	36,5	11	17	31
[0%,90%]	17	27	47	17,6	22,2	35,6
Min	0	8	12	0	8	12
Max	20	36	58	20	27	38

6.2.4 De wissel

In de volledige testverzameling werden slechts 2 wissels gedetecteerd op een totaal van 80 wissels. In beide gevallen werden zowel de correcte spelers geëxtraheerd als het tijdstip correct berekend. Om de vertragingen te meten werd het moment dat de vierde scheidsrechter het wisselbord in de lucht stak als wisseltijdstip genomen. De vertraging van het eerste bericht, de explosie en de wisseldetectie zijn respectievelijk 2, 7 en 19 seconden voor de eerste wissel en respectievelijk 5, 5 en 7 seconden voor de tweede wissel.

6.2.5 Het einde van de eerste helft

Om de detecties van het einde van de eerste helft te evalueren leiden we uit de videosequentie af hoeveel extra tijd er bij elke wedstrijd werd verder gespeeld. In totaal werd 18 keer het einde van de eerste helft gedetecteerd in 12 verschillende wedstrijden. Indien we geen rekening houden met wanneer het einde van de eerste helft gedetecteerd werd is de precisie 100%. Cruciaal bij deze gebeurtenis is dat de opgegeven minuut correct is. Voor de evaluatie beschouwen we net zoals bij de foutdetectie, 2 strategieën. De eerste strategie is de individuele detectiestrategie. Aangezien het detectiesysteem altijd slechts eenmaal het einde van de eerste helft zal detecteren per berichtstroom komt dit overeen met alle detecties. De tweede strategie is de gecombineerde strategie. Hierbij combineren we de resultaten van beide berichtstromen.

De resultaten van de evaluatie van beide strategieën worden weergegeven in Tabel 6.18 en Tabel 6.19. We zien dat bij een gecombineerde aanpak 92% van de detecties maximaal 1 minuut verschillen met de werkelijke minuut, terwijl dit voor de individuele aanpak slechts 83% is. Opnieuw blijkt het voordelig beide berichtstromen te combineren.

Net zoals bij alle voorgaande evaluaties moet ook de vertraging geëvalueerd worden. De resultaten worden weergegeven in Tabel 6.20. Dit keer ligt de gemiddelde vertraging veel hoger dan bij de andere gebeurtenissen. Dit komt ondermeer omdat we 2 uitschieters

Tabel 6.18: Prestatie van de minuutberekening bij de weergave van het einde van de eerste helft voor de strategie individuele detectie.

	Individuele detectie				
	0	**1**	**2**	**4**	**9**
Aantal	13	2	1	1	1
Fractie	72,22%	11,11%	5,56%	5,56%	5,56%

Tabel 6.19: Prestatie van de minuutberekening bij de weergave van het einde van de eerste helft voor de strategie gecombineerde detectie.

	Gecombineerde detectie				
	0	**1**	**2**	**4**	**9**
Aantal	9	2	0	0	1
Fractie	75,00%	16,67%	0,00%	0,00%	8,33%

detecteren van 4 en 9 minuten. Het is bijgevolg interessanter naar de mediaan te kijken. De mediaan is gelijk aan 23 en 24 seconden. Dit ligt in lijn met de andere gebeurtenissen.

Tabel 6.20: Statistiek van de vertraging ten opzichte van het beëindigen van de eerste helft door de scheidsrechter in de videosequentie. Alle waarden zijn uitgedrukt in seconden.

	Individuele detectie			**Gecombineerde detectie**		
	Eerste	**Explosie**	**Detectie**	**Eerste**	**Explosie**	**Detectie**
Gemiddelde	1,94	14,72	72,00	2,65	15,65	76,29
Mediaan	-1	17,5	23	-1	18	24
Standaard afw,	10,88	13,45	133,35	10,78	13,27	136,17
[0%,75%]	12	22,25	52,5	14	23	57
[0%,90%]	16,3	33	155,4	16,4	33	170,2
Min	-13	-6	-5	-13	-6	-5
Max	23	38	545	23	38	545

We kunnen besluiten dat de detectie van de eerste helft al zeer goed presteert. Er is echter nog ruimte voor verbetering. Sommige detecties zijn veel later dan andere. Dit wordt ondermeer veroorzaakt door het beperkt aantal kenmerken dat gebruikt wordt door de SVM om een tweet de klasse 'einde eerste helft' toe te kennen. Nu wordt er voornamelijk gefocust op tweets die termen zoals 'HT' en 'half-time' bevatten.

6.2.6 Het einde van de tweede helft

De laatste gebeurtenis die we evalueren is het einde van de tweede helft. Dit doen we op dezelfde wijze als voor het einde van de eerste helft. Het einde van de tweede helft werd 8 keer gedetecteerd in 7 wedstrijden. Opnieuw maken we gebruik van 2 strategieën: de individuele detectie en de gecombineerde detectie. In beide gevallen werd een precisie van 100% bereikt voor zowel de detectie als de minuutberekening. De recall van de gecombineerde aanpak is 47%. Deze lage recall is te wijten aan dezelfde reden als de

trage detectie van het einde van de eerste helft. De focus ligt te veel op een beperkte verzameling kenmerken. Dit komt nog sterker tot uiting bij het einde van de tweede helft omdat dan veel Twittergebruikers hun emoties uitdrukken over de overwinning of de nederlaag die hun ploeg geboekt heeft.

We evalueren ook de snelheid waarmee de gebeurtenis gedetecteerd wordt. De resultaten worden weergegeven in Tabel 6.21. De mediaan van de gecombineerde detectie ligt beduidend lager dan de mediaan van de gecombineerde detectie van de eerste helft. Voor de individuele detectie vinden we een gelijkaardige detectiesnelheid terug.

Tabel 6.21: Statistiek van de vertraging ten opzichte van het beëindigen van de tweede helft door de scheidsrechter in de videosequentie. Alle waarden zijn uitgedrukt in seconden.

	Individuele detectie			**Gecombineerde detectie**		
	Eerste	**Explosie**	**Detectie**	**Eerste**	**Explosie**	**Detectie**
Gemiddelde	-6,5	5,5	16,5	-7,29	3,71	15
Mediaan	-5,5	6,5	20	-6	1	14
Standaard afw,	6,35	12,29	17,37	6,42	12,11	18,19
[0%,75%]	-1	16,5	27,75	-3	14	28
[0%,90%]	-1	18	33	-1	16,8	34
Min	-18	-15	-11	-18	-15	-11
Max	-1	18	40	-1	18	40

We kunnen besluiten dat indien het einde van de tweede helft gedetecteerd wordt, de weergegeven minuut met zeer hoge waarschijnlijkheid correct is. Dit is in contrast met de detectie van het einde van de eerste helft. Daarentegen is de recall lager dan 50%. Er is dus zeker nog ruimte voor verbetering.

6.2.7 Besluit

In deze sectie hebben we de 6 verschillende gebeurtenissen die het gebeurtenisdetectiesysteem in Twitter kan detecteren en identificeren uitgebreid geëvalueerd. We hebben zowel de prestaties van de gebeurtenisdetectie en -identificatie geëvalueerd als de snelheid waarmee dit gebeurt ten opzichte van de live-uitzending.

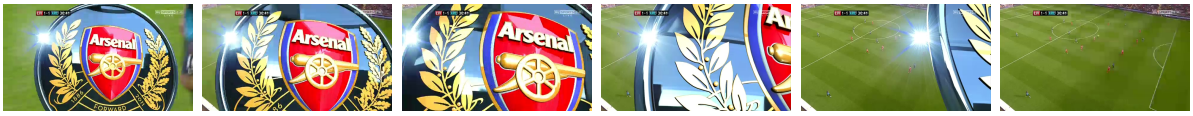
Over alle gebeurtenissen heen blijkt een aanpak die beide Twitterberichtstromen combineert de beste aanpak te zijn voor zowel het detecteren en identificeren van een gebeurtenis als de snelheid waarmee dit gebeurt. Bij elke gebeurtenisdetectie werd een precisie van 90% of hoger bereikt.

Voor de doelpuntdetectie en -identificatie werden uitmuntende resultaten behaald met een recall en precisie van meer dan 95%. De grootste verbeteringen kunnen bereikt worden bij de gebeurtenissen 'fout', 'einde eerste helft' en 'einde tweede helft' door het toevoegen van extra kenmerken aan de tweetvector of gebruik te maken van geavanceerde analyses. Toch

moeten we opmerken dat we altijd gebonden zijn aan de tweets die de Twittergebruikers verzenden. We verwachten daarom niet veel verbetering bij de detecties van wissels omdat deze zelden voldoende gerapporteerd worden.

6.3 Evaluatie van de individuele componenten van het hoogtepuntextractiealgoritme

In deze sectie evalueren we de prestaties van het shotgrensdetectiealgoritme en het logo-detectiealgoritme. De testverzameling bevat 3 soorten videosequenties. 2 types sequenties bevatten een logo en 1 type sequentie bevat geen logo. De 2 type logo's worden afgebeeld in Figuur 6.1 en Figuur 6.2. Tabel 6.22 geeft een overzicht van het videoformaat en de gebruikte logo's per wedstrijd: videosequenties verzameld via het internet zijn in MPEG-4 en videosequenties verzameld via Telenet digitale televisie in MPEG-2.



Figuur 6.1: Illustratie van de 6 laatste beelden van het eerste type logo uit de testverzameling. Het logo is aangepast aan de ploeg die een doelpunt heeft gemaakt.

Figuur 6.2: Illustratie van de 6 laatste beelden van het tweede type logo uit de testverzameling.

6.3.1 Shotgrensdetectie

Voor de evaluatie van de shotgrensdetectie selecteren we uit 3 wedstrijden van de testverzameling elk 5 minuten video, namelijk minuut 25 tot 30. We selecteren van elk type videosequentie een wedstrijd. De wedstrijd Tottenham Hotspurs - Manchester United (wedstrijd 1) bevat het logo uit Figuur 6.1, de wedstrijd Arsenal - Wigan Athletic (wedstrijd 2) bevat het logo uit Figuur 6.2 en de wedstrijd Everton - Sunderland (wedstrijd 3) bevat geen logo's.

Voor de verschillende sequenties worden de shotgrenzen manueel bepaald. In totaal bevatten de videosequenties samen 71 abrupte overgangen en 8 graduele overgangen. Er

Tabel 6.22: Overzicht van het gebruikte logo per wedstrijd van de testverzameling (Tabel A.4).

Thuis - uit	**Logo type**	**Formaat**	**Score**
Arsenal - Chelsea	Logo type 2	MPEG-2	0 - 0
Arsenal - Wigan Athletic	Logo type 2	MPEG-2	1 - 2
Blackburn Rovers - Manchester United	Logo type 1	MPEG-2	0 - 2
Chelsea - Queens Park Rangers	Logo type 2	MPEG-2	6 - 1
Chelsea - Wigan Athletic	Geen logo	MPEG-2	2 - 1
Everton - Sunderland	Geen logo	MPEG-2	4 - 0
Fulham - Chelsea	Logo type 1	MPEG-2	1 - 1
Fulham - Wolverhampton Wanderers	Logo type 1	MPEG-4	5 - 0
Manchester City - Manchester United	Logo type 2	MPEG-2	1 - 0
Manchester City - Sunderland	Geen logo	MPEG-2	3 - 3
Manchester United - Aston Villa	Logo type 2	MPEG-2	4 - 0
Manchester United - Queens Park Rangers	Logo type 1	MPEG-2	2 - 0
Stoke City - Arsenal	Logo type 2	MPEG-2	1 - 1
Tottenham Hotspurs - Blackburn Rovers	Logo type 2	MPEG-2	2 - 0
Tottenham Hotspurs - Manchester United	Logo type 1	MPEG-4	1 - 3

wordt geen rekening gehouden met de logotransities omdat dit niet relevant is voor het finale algoritme. De resultaten van de shotgrensdetectie worden weergegeven in Tabel 6.23.

Tabel 6.23: Prestatie van het shotgrensdetectiealgoritme op fragmenten uit 3 verschillende wedstrijden.

	Wedstrijd 1	**Wedstrijd 2**	**Wedstrijd 3**	**Totaal**
Gedetecteerd	28	14	32	74
Fout	3	0	1	4
Gemist	1	0	0	1
Correct	25	14	31	70
Precisie	89,29%	100,00%	96,88%	94,59%
Recall	96,15%	100,00%	100,00%	98,59%

Op de in totaal 74 gedetecteerde shotgrenzen worden er slechts 4 overgangen tussen beelden foutief als shotgrens aanzien. 2 fouten worden veroorzaakt door het wijzigen van de lichtreclame bij een close-up op een speler, de derde fout wordt veroorzaakt door de flits van een fototoestel en de vierde fout wordt veroorzaakt door het snel passeren van een been van een speler wanneer de camera volledig is ingezoomd op een speler die veel verder verwijderd is van de camera. Slechts 1 shotgrens wordt niet gedetecteerd. De oorzaak hier is de overgang tussen 2 close-ups van een speler in een bijna identieke scène. Geen enkele graduele overgang wordt als abrupte overgang gedetecteerd.

In [24] stellen de auteurs een shotgrensdetectiealgoritme voor dat gebruik maakt van het verschil in dominante kleurratios in de RGB-ruimte en het verschil van de histogrammen in de HSV-ruimte. De auteurs evalueren de prestaties van hun algoritme op 4 wedstrijden

en rapporteren een precisie van 94,2% en een recall van 97,1%. Het shotgrensdetectiealgoritme dat we hier voorstellen vergt een stuk minder berekeningen en maakt slechts gebruik van 1 kleurruimte. Toch behalen we op onze beperkte testverzameling betere resultaten dan in de literatuur. Een gelijkaardige aanpak wordt voorgesteld in [23], maar hier wordt er gebruik gemaakt van de HSI-ruimte in plaats van de HSV-ruimte. De precisie is in dit geval 91,7% en de recall 97,3%.

We kunnen besluiten dat ons lichtgewicht shotgrensdetectiealgoritme dat gebruikt maakt van lokale histogrammen zeer mooie resultaten neerzet die gelijkaardig zijn aan algoritmen in de literatuur [23, 24].

6.3.2 Logodetectie

Om de logodetectie te evalueren, selecteren we alle doelpuntherhalingen uit de testverzameling. Uit Tabel 6.22 leren we dat 12 van de 15 wedstrijden een logo gebruiken en dat er in totaal 34 doelpunten werden gemaakt. Bijgevolg bevat onze testverzameling 68 logotransities. Om het werk te vereenvoudigen maken we gebruik van de herhalingsdetecties uit de globale systeemevaluatie en vullen dit aan met individuele evaluaties van de logotransities van de gefaalde herhalingsdetecties. In zo een geval bakenen we telkens een interval van 30 seconden af rond de logotransitie. Het resultaat van de evaluatie is te vinden in Tabel 6.24.

Tabel 6.24: Prestatie van het logodetectiealgoritme op 2 types logo's.

	Logo 1	**Logo 2**	**Totaal**
Gedetecteerd	27	38	65
Fout	0	0	0
Gemist	3	0	3
Correct	27	38	65
Precisie	100,00%	100,00%	100,00%
Recall	90,00%	100,00%	95,59%

Voor beide logo's halen we een precisie van 100%. Voor het eerste type logo halen we slechts een recall van 90% terwijl we voor het tweede type logo een recall van 100% halen. Wanneer we de oorzaak van falen proberen te achterhalen, merken we op dat alle 3 de gemiste logo's voorkomen bij herhalingen van doelpunten gescoord door Manchester United. Om een logotransitie te detecteren moet aan 2 voorwaarden voldaan zijn. Uit Sectie 5.2.3 weten we dat bij minstens 5 van de 10 overgangen moet gelden dat:

$$D_n(i, i+1) > 0,125$$

en bovendien moet bij 1 overgang gelden dat:

$$D_n(i, i+1) > 0,5$$

Een analyse van de $D_n(i, i+1)$-waarden van de verschillende opeenvolgende beelden leert ons dat de piekwaarde van 0,5 nooit overschreden wordt. De maximale waarde die normaal bereikt wordt op het einde van de logotransitie schommelt bij de 3 gemiste logo's tussen 0,42 en 0,49. De oorzaak hiervan is zichtbaar in Figuur 6.3. In tegenstelling tot de transitie in Figuur 6.1 gaat de logotransitie over in een close-up van een speler van Manchester United in plaats van een bovenaanzicht van het veld. Zowel het logo als het truitje bevat veel rood waardoor het centrale lokale histogram slechts beperkt wijzigt in de 2 opeenvolgende beelden. Het verlagen van de piekgrenswaarde van 0,5 naar 0,4 zou hier een oplossing kunnen bieden. We hebben dit echter niet geïmplementeerd.

Figuur 6.3: Illustratie van de oorzaak van een falende logodetectie.

We vergelijken opnieuw de behaalde prestaties met resultaten uit de literatuur. In [24] detecteren de auteurs logo's in voetbalsequenties aan de hand van voorgedefinieerde RGB-waarden. De karakteristieke RGB-waarden van de logo's worden op voorhand afgeleid en vervolgens wordt elke beeld in de videosequentie met dit RGB-sjabloon vergeleken. De auteurs rapporteren een precisie van 96,3% en een recall van 100%. De prestaties van ons logodetectiealgoritme dat gebruik maakt van lokale histogrammen zijn vergelijkbaar. Het voorgestelde logodetectiealgoritme is echter generieker doordat het niet afhangt van de kleuren van het logo, maar toepasbaar is op alle logo's die eenzelfde beweging maken.

6.4 Globale evaluatie van het hoogtepuntextractiealgoritme

Het finale doel van deze masterproef is om hoogtepunten te extraheren uit videosequenties en te annoteren op basis van Twitter. Als toepassing hebben we er voor gekozen om ons te beperken tot doelpunten. Vanuit praktisch oogpunt werd reeds gekozen voor het combineren van de 2 berichtstromen bij de ontwikkeling van het algoritme op basis van de ontwikkelingsverzameling. Het is perfect mogelijk om de berichtstromen elk apart te analyseren maar de videoanalyse vergt gemiddeld 2 seconden per 5 seconden video.

Bijgevolg zou het niet meer mogelijk zijn om in realtime de hoogtepuntextractie uit te voeren indien we de analyse tweemaal zouden moeten uitvoeren. We maken bijgevolg gebruik van de eerste doelpuntdetectiestrategie die werd voorgesteld in Sectie 6.2.1.

In totaal bevat de testverzameling 12 wedstrijden die gebruik maken van een logo om een herhaling in en uit te leiden. Een overzicht is te vinden in Tabel 6.22. Deze deelverzameling van de testverzameling bevat 34 doelpunten. Voor elke wedstrijd hebben we 2 Twitterberichtstromen die gesynchroniseerd zijn met de voetbalsequenties. Voor de evaluatie simuleren we een realtime omgeving waarbij de 2 berichtstromen en de videosequentie naast elkaar worden afgespeeld. Wanneer een doelpunt gedetecteerd wordt in 1 van de 2 berichtstromen zal het hoogtepuntextractiealgoritme proberen om de herhaling uit de videosequentie te extraheren en te annoteren. Indien tijdens deze extractiefase een tweede doelpuntdetectie plaatsvindt, wordt deze genegeerd. Uit Sectie 5.3.2 weten we dat er 2 gevallen kunnen plaatsvinden: ofwel vindt de doelpuntdetectie in Twitter plaats voordat de herhaling begint ofwel vindt de doelpuntdetectie plaats tijdens de herhaling. Tijdens het evalueren van de testverzameling kwamen we echter eenmaal tegen dat de detectie pas plaatsvond 10 seconden na het beëindigen van de herhaling.

Het resultaat van de herhalingsextratieoperatie is te vinden in Tabel 6.25. Er worden 3 onderverdelingen gemaakt. In de eerste kolom wordt het totaalresultaat weergegeven. In de tweede en derde kolom worden de resultaten onderverdeeld per type logo en in de vierde, vijfde en zesde kolom worden de resultaten onderverdeeld per tijdstip van doelpuntdetectie in Twitter, namelijk voor, tijdens of na de herhaling.

Tabel 6.25: Prestatie van het hoogtepuntextractiealgoritme op de testverzameling.

	Herhalingen					
	Alle	**Logo 1**	**Logo 2**	**Voor**	**Tussen**	**Na**
Extracties	32	13	19	18	13	1
Fout	2	0	2	1	0	1
Gemist	2	2	0	2	0	0
Correct	30	13	17	17	13	0
Precisie	93,75%	100,00%	89,47%	94,44%	100,00%	0,00%
Recall	93,75%	86,67%	100,00%	89,47%	100,00%	100,00%

In totaal vonden 32 herhalingsextracties plaats, 2 daarvan waren fout. Zoals hierboven beschreven werd 1 foute extractie veroorzaakt door de late doelpuntdetectie in Twitter. Hierdoor werd het einde van de herhaling beschouwd als het begin van de herhaling. De tweede foute extractie werd veroorzaakt door de korte opeenvolging van 2 logo's. Vlak voor het scoren van het doelpunt werd een herhaling beëindigd. Het doelpunt werd in Twitter gedetecteerd voor de start van doelpuntherhaling. Hierdoor viel het logo van de vorige herhaling binnen de 40 seconden na de doelpuntdetectie en werd bijgevolg niet de herhaling geëxtraheerd, maar de periode tussen de 2 herhalingen. Bij toeval werd echter

de doelpuntfase zelf geëxtraheerd in plaats van de herhaling van het doelpunt.

Beide fouten kunnen vermeden worden door een gelijkaardige controle. Karakteristiek aan een herhaling is dat tijdens de herhaling enkel graduele overgangen worden gebruikt. Indien we 2 logo's terugvinden waartussen gebruik wordt gemaakt van abrupte overgangen, dan kan het gevonden fragment geen herhaling zijn.

2 herhalingen werden niet gevonden. Voor het hoogtepuntextractiealgoritme is het cruciaal dat het eerste logo gedetecteerd wordt. Voor het tweede logo hebben we immers een back-upplan dat gebruik maakt van shotgrensdetectie. De reden voor falen werd reeds uitgelegd in Sectie 6.3.2. Een logotransitie vloeit over in een close-up van een speler van Manchester United waardoor de piekgrenswaarde niet gehaald wordt.

Net zoals bij de evaluatie van het globale gebeurtenisdetectiesysteem evalueren we ook hier de snelheid waarmee de extractie plaatsvindt. De gemiddelde lengte van een herhaling in de testverzameling is 32 seconden. De minimale lengte is 16 seconden en de maximale is 45 seconden. Gemiddeld zijn er 7 seconden tussen het einde van de herhaling en het beschikbaar zijn van de geannoteerde herhaling. De finale vertraging tussen het scoren van het doelpunt in de live-uitzending en het beschikbaar zijn van de herhaling wordt weergegeven in Tabel 6.26. Gemiddeld is de herhaling beschikbaar na 56 seconden en in 90% van de gevallen is dit na 68 seconden. We merken op dat hier de vertraging sterk afhankelijk is van wanneer de herhaling eindigt en in mindere mate van de snelheid waarmee het doelpunt in Twitter gedetecteerd wordt. Enkel indien het tweede herhalingslogo niet kan gedetecteerd worden en we terug moeten vallen op shotgrensdetectie, krijgen we te maken met een buitensporige vertraging. Dit kwam slechts eenmaal voor in de volledige testverzameling en zorgde voor een vertraging van 99 seconden.

Tabel 6.26: Statistiek van de vertraging tussen het scoren van een doelpunt in de live-uitzending en het beschikbaar zijn van de geannoteerde herhaling. Alle waarden zijn in seconden uitgedrukt.

	Totaal
Gemiddelde	56
Mediaan	55
Standaard afwijking	12
[0%,75%]	62
[0%,90%]	68
Min	34
Max	99

We kunnen besluiten dat het hoogtepuntextratiealgoritme in staat is op accurate wijze herhalingen van een doelpunt uit een videosequentie te extraheren met behulp van doelpuntdetectie in Twitter. Zowel de precisie als recall liggen hoger dan 93%. In deze sectie stelden we daarnaast ook enkele verbeteringen voor die de recall en precisie nog zouden kunnen verhogen.

Hoofdstuk 7

Besluit en toekomstig werk

In deze masterproef zijn we erin geslaagd een systeem te bouwen dat in realtime hoogtepunten kan extraheren uit een videosequentie door gebruik te maken van een gebeurtenisdetectiesysteem dat belangrijke gebeurtenissen in Twitter detecteert. Bovendien kunnen we het videofragment van het hoogtepunt verrijken met semantische informatie door een semantische analyse uit te voeren op de Twitterberichten.

7.1 Gebeurtenisdetectie en -identificatie in Twitter

Voor de ontwikkeling van het realtime gebeurtenisdetectiesysteem hebben we ervoor gekozen om 6 basisgebeurtenissen te detecteren en te identificeren in Twitter: het doelpunt, de strafschop, de fout, de wissel, het einde van de eerste helft en het einde van de tweede helft. Voor het detecteren van deze gebeurtenissen hebben we ondermeer gebruik gemaakt van de sterke stijging in tweetvolume tijdens gebeurtenissen en van de algemene kenmerken van deze 6 gebeurtenissen. We denken hierbij aan sleutelwoorden, spelersnamen en sentimentkenmerken die gebruikt werden door een Support Vector Machine (SVM) om de tweets in te delen in 1 van de 6 basisgebeurtenissen en de klasse 'geen gebeurtenis'. Diezelfde SVM werd ook gebruikt om de gebeurtenissen te identificeren. Finaal werd er relevante informatie geëxtraheerd uit Twitter die eigen was aan de geïdentificeerde gebeurtenis. We denken hierbij aan de speler die het doelpunt gescoord heeft of welke type fout er gemaakt is.

Uit de evaluatie blijkt dat doelpunten het best kunnen gedetecteerd worden wanneer de berichtstromen van beide ploegen gecombineerd worden. In dat geval zijn we in staat doelpunten te detecteren met een precisie van 95.8% en een recall van 97.7%. Wanneer we de strafschopdetecties enkel gebruiken als aanvulling bij een doelpunt, kunnen we alle strafschoppen perfect detecteren. Voor de andere 4 gebeurtenissen stellen we vast dat we sterk afhankelijk zijn van het Twittergedrag van de supporters. Sommige gebeurtenissen

worden slechts door enkelingen gerapporteerd en andere gebeurtenissen zelfs helemaal niet. Dit is het sterkst merkbaar bij fouten en wissels. Bij de gemiste detecties van het einde van de eerste of tweede helft ligt het probleem eerder bij de beperkte verzameling sleutelwoorden die de SVM gebruikt en de hoge grenswaarde bij identificatie. Indien 1 van deze 4 gebeurtenissen gedetecteerd wordt, zal echter telkens met hoge precisie de correcte informatie geëxtraheerd worden.

Het voorgestelde systeem is nieuw in zijn soort. Tot op heden bestaat er nog geen enkel systeem dat in realtime gebeurtenissen detecteert in voetbalwedstrijden. We zijn de eersten die zo een uitgebreid systeem voorstellen dat 6 verschillende gebeurtenissen kan detecteren en identificeren en bovendien voor elke gebeurtenis extra informatie kan afleiden. Het systeem kan ook op simpele wijze aangepast worden naar andere talen. In [18] werd voor het eerste een realtime gebeurtenisdetectiesysteem voorgesteld voor American Football-wedstrijden. Hierbij werd enkel gebruik gemaakt van het tweetvolume en sleutelwoorden en was de detectie en identificatie beperkt tot 4 gebeurtenissen. Gebaseerd op [19] hebben we deze aanpak uitgebreid met een SVM die tweets classificeert en rekening houdt met spelersnamen en sentimentkenmerken.

Zoals reeds aangehaald, is er nog ruimte voor verbetering. Uit de evaluatie blijkt dat het systeem soms moeite heeft met meervoudige gebeurtenissen. Een uitbreiding van de trainingsverzameling van de SVM en het herbekijken van de gebeurtenisidentificatieregels kan hier een oplossing bieden. Een tweede verbetering kan bereikt worden bij de minder belangrijke gebeurtenissen. De gekozen kenmerken zijn bijzonder goed in het detecteren van doelpunten en strafschoppen maar minder goed in het detecteren van fouten en het einde van de eerste en tweede helft. We stellen voor om gebruik te maken van een tweede SVM die gebruik maakt van unigrams of bigrams in plaats van algemene kenmerken. Hierdoor zullen gebeurtenissen met een meer gediversifieerde lexicon beter gedetecteerd worden en zal er rekening kunnen gehouden worden met negaties van gebeurtenissen. Het grote nadeel is dat het systeem dan sterk afhankelijk wordt van de taal.

7.2 Hoogtepuntextractie uit videosequenties op basis van Twitter

Als toepassing van het gebeurtenisdetectiesysteem hebben we een hoogtepuntextractie-algoritme ontwikkeld dat in realtime het bijhorende fragment uit de videosequentie kan extraheren. We hebben hier de focus gelegd op doelpunten omdat deze gebeurtenis het meest voorkomt, het best kan gedetecteerd worden door het gebeurtenisdetectiesysteem en herhaald wordt in de videosequentie.

Voor de extraheren van hoogtepunten maken we gebruik van shotgrensdetectie en logode-

tectie op basis van lokale histogrammen. Deze lichtgewichtaanpak haalt in beide gevallen een precisie en recall hoger dan 94% en blijkt te kunnen concurreren met bestaande algoritmes [24]. De hoogtepuntextractie op basis van Twitter haalt een precisie en recall van 93.75%.

Het voorgestelde algoritme is uniek omwille van 2 redenen. Voor het eerst wordt een algoritme voorgesteld dat op basis van sociale media op accurate wijze doelpuntfragmenten uit videosequenties kan extraheren en annoteren in realtime. In de literatuur [21, 22] worden enkel aanpakken voorgesteld waarbij na afloop van de wedstrijd de Twitterberichtstroom naast de videosequentie wordt gelegd en op basis van sleutelwoorden [21, 22] of tweetvolume [22], shots [22] of videofragmenten van 1 minuut [21] kunnen opgevraagd worden. Dit geeft meteen aanleiding tot een tweede reden. We maken als eersten gebruik van een geavanceerd gebeurtenisdetectiesysteem in Twitter om de verschillende gebeurtenissen samen te vatten en vervolgens het correcte fragment te selecteren en te annoteren.

De hoogtepuntextractie is echter nog niet perfect. Zoals blijkt uit de evaluatie worden sommige logo's niet goed gedetecteerd en was er 1 geval waarbij de gebeurtenisdetectie in Twitter pas plaatsvond na afloop van de herhaling. Er werd telkens een oplossing voorgesteld. Op basis van deze oplossing stellen we een uitbreiding voor die in mindere mate afhankelijk is van logo's. In de volledige testverzameling bestond elke doelpuntherhaling uit meerdere shots die in elkaar overvloeiden via graduele overgangen. Buiten de herhalingen werd er nooit gebruik gemaakt van graduele overgangen maar enkel van abrupte overgangen. Een hoogtepuntextractiesysteem die van deze observatie gebruik maakt is onafhankelijk van het gebruik van een logo en van het type logo. In Sectie 3.2.1 werd reeds heel wat literatuur verzameld rond shotclassificatie. Op basis van shotclassificatie en cinematografische kenmerken van doelpunten zouden we ook het shot kunnen extraheren waar het doelpunt effectief gescoord is. Hierdoor zou een doelpuntfragment nog sneller beschikbaar gesteld kunnen worden.

In deze masterproef werden 2 uitgebreide systemen voorgesteld om enerzijds gebeurtenissen te identificeren in Twitter en om anderzijds hoogtepunten te extraheren uit videosequenties op basis van Twitter. De voorgestelde algoritmen en technieken kunnen ook gebruikt worden om nog meer verschillende gebeurtenissen te identificeren in Twitter en te extraheren uit videosequenties, maar kunnen ook, mits enkele aanpassingen, toegepast worden op andere sporten.

Bijlage A

Uitgebreid overzicht van de gebruikte wedstrijden

Tabel A.1: Wedstrijden gebruikt voor de ontwikkeling van een detectiealgoritme.

Thuis (T)	Uit (U)	Datum	Hashtag	Goal		Gele kaart		Rode kaart		Fouten		Wissel	
				T	U	T	U	T	U	T	U	T	U
Chelsea	Arsenal	29/10/2011	#CFC, #AFC	3	5	2	3	0	0	7	18	3	3
Sunderland	Aston Villa	29/10/2011	#AVFC	2	2	1	3	0	0	7	14	3	1
Stoke City	Newcastle United	31/10/2011	#SCFC, #NUFC	1	3	0	1	0	0	8	12	3	3
Manchester United	Newcastle United	26/11/2011	#MUFC	1	1	1	5	0	1	11	16	2	3
West Bromwich Albion	Tottenham Hotspurs	26/11/2011	#THFC	1	3	2	1	0	0	11	13	3	1
Bolton Wanderers	Liverpool	03/11/2011	#LFC	3	1	0	2	0	0	10	13	2	2
Chelsea	Bolton Wanderers	25/02/2011	#CFC	3	0	0	1	0	0	7	13	3	2
Newcastle United	Wolverhampton	25/02/2011	#NUFC	2	2	1	1	0	0	8	8	3	3
Manchester City	Blackburn Rovers	25/02/2011	#MUFC, #BRFC	3	0	0	2	0	0	6	10	3	2
Arsenal	Tottenham Hotspurs	26/02/2011	#AFC	5	2	3	3	0	1	15	12	3	3
Norwich City	Manchester United	26/02/2011	#MUFC	1	2	1	2	0	0	9	10	3	2

Tabel A.2: Wedstrijden gebruikt voor de ontwikkeling van de SVM. De wedstrijd met een * is een League Cup wedstrijd.

Thuis (T)	Uit (U)	Datum	Hashtag	Goal		Gele kaart		Rode kaart		Fouten		Wissel	
				T	U	T	U	T	U	T	U	T	U
Chelsea	Wolverhampton	26/11/2011	#CFC	3	0	1	2	0	0	10	11	3	3
Liverpool	Manchester City	27/11/2011	#MCFC	1	1	1	5	0	1	11	15	1	3
Chelsea	Liverpool	29/11/2011	#LFC	0	2	5	1	0	0	/	/	3	3
Manchester City	Norwich City	03/12/2011	#MCFC	5	1	0	1	0	0	10	9	3	3
Wigan Athletic	Arsenal	03/12/2011	#AFC	0	4	3	2	0	0	8	12	3	3

Tabel A.3: Overzicht van de wedstrijden gebruikt voor de ontwikkeling van een algoritme om Twitter en videosequenties te combineren. Wanneer de resolutie 720x576 pixels is, is het videoformaat MPEG-2. Wanneer de resolutie 640x360 pixels is het videoformaat MPEG-4.

				Goal		Gele kaart		Rode kaart		Fouten		Wissel	
Thuis (T) - Uit (U)	Datum	Hashtag	Resolutie	T	U	T	U	T	U	T	U	T	U
Arsenal - Newcastle United	12/03/2012	#AFC, #NUFC	640x360	1	1	0	0	0	0			0	0
Liverpool - Arsenal	03/03/2012	#LFC, #AFC	640x360	1	2	0	1	0	0	11	9	2	3
Norwich City - Manchester City	14/04/2012	#NCFC, #MCFC	720x576	1	6	1	2	0	0	10	5	3	3

Tabel A.4: Overzicht van de wedstrijden gebruikt voor de evaluatie van zowel het gebeurtenisdetectiesysteem als de finale videotoepassing. Wanneer de resolutie 720x576 pixels is, is het videoformaat MPEG-2. Wanneer de resolutie 640x360 pixels is het videoformaat MPEG-4.

				Goal		Gele kaart		Rode kaart		Fouten		Wissel	
Thuis (T) - Uit (U)	Datum	Hashtag	Resolutie	T	U	T	U	T	U	T	U	T	U
Fulham - Wolverhampton Wanderers	04/03/2012	#FFC, #WWFC	640x360	5	0	1	2	0	0	7	10	3	3
Tottenham Hotspurs - Manchester United	04/03/2012	#THFC, #MUFC	640x360	1	3	1	2	0	0	9	10	3	2
Arsenal - Chelsea	21/04/2012	#AFC, #CFC	720x576	0	0	3	4	0	0	11	11	3	3
Arsenal - Wigan Athletic	16/03/2012	#AFC, #WAFC	720x576	1	2	2	3	0	0	15	13	2	2
Blackburn Rovers - Manchester United	02/04/2012	#BRFC, #MUFC	720x576	0	2	1	1	0	0	8	5	0	3
Chelsea - Queens Park Rangers	29/04/2012	#CFC, #QPR	720x576	6	1	0	1	0	0	10	7	3	2
Chelsea - Wigan Athletic	07/04/2012	#CFC, #WAFC	720x576	2	1	2	2	0	0	14	11	3	3
Everton - Sunderland	09/04/2012	#EFC, #SAFC	720x576	4	0	0	2	0	0	7	12	3	3
Fulham - Chelsea	09/04/2012	#FFC, #CFC	720x576	1	1	1	3	0	0	5	16	3	2
Manchester City - Manchester United	30/04/2012	#MCFC, #MUFC	720x576	1	0	3	2	0	0	11	14	3	3
Manchester City - Sunderland	31/03/2012	#MCFC, #SAFC	720x576	3	3	2	2	0	0	11	13	3	2
Manchester United - Aston Villa	15/04/2012	#MUFC, #AVFC	720x576	4	0	3	3	0	0	9	7	3	3
Manchester United - Queens Park Rangers	08/04/2012	#MUFC, #QPR	720x576	2	0	1	0	0	1	13	10	3	3
Stoke City - Arsenal	28/04/2012	#SCFC, #AFC	720x576	1	1	1	2	0	0	8	10	3	3
Tottenham Hotspurs - Blackburn Rovers	29/04/2012	#THFC, #BRFC	720x576	2	0	0	0	0	0	5	14	3	2

Bibliografie

[1] Laura Gainor. Sports and social media, February 2012. http://blog.gmrmarketing.com/sports/sports-social-media-infographic-gmr-marketing/.

[2] Twitter. About twitter, May 2012. http://twitter.com/about.

[3] Twitter. Twitter support, May 2012. http://support.twitter.com/.

[4] Twitter. Year in review, December 2011. http://yearinreview.twitter.com/en/tps.html.

[5] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[6] Ana-Maria Popescu and Marco Pennacchiotti. Detecting controversial events from twitter. In *Conference, International Conference on Information and Knowledge Management (CIKM 2010), Toronto , ON, Canada - October 26 - 30, 2010*, pages 1873–1876, New York, NY, USA, 2010. ACM.

[7] Ana-Maria Popescu and Marco Pennacchiotti. "dancing with the stars, "nba games, politics: An exploration of twitter users' response to events. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press, 2011.

[8] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6:248–260, 2009.

[9] David Shamma, Lyndon Kennedy, and Elizabeth Churchill. Conversational shadows: Describing live media events using short messages, 2010.

[10] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[11] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6, 2009.

[12] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[13] Hassan Saif, Yulan He, and Harith Alani. Alleviating data sparsity for twitter sentiment analysis. In *Making Sense of Microposts (#MSM2012)*, pages 2–9, 2012.

[14] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 841–842, New York, NY, USA, 2010. ACM.

[15] Christopher Horn. Analysis and Classication of Twitter messages. Master's thesis, Graz University of Technology, Oostenrijk, 2010.

[16] David Shamma, Lyndon Kennedy, and Elizabeth Churchill. Tweetgeist: Can the twitter timeline reveal the structure of broadcast events? *Horizon, In CSCW 2010*, 2010.

[17] Xingtian Shi, Zhenglu Yang, Masashi Toyoda, and Masaru Kitsuregawa. Harnessing the wisdom of crowds: video event detection based on synchronous comments. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 123–124, New York, NY, USA, 2011. ACM.

[18] Siqi Zhao, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. Human as real-time sensors of social and physical events: A case study of twitter and sports games. *CoRR*, abs/1106.4300, 2011.

[19] Smitashree Choudhury and John Breslin. Extracting semantic entities and events from sports tweets. In *Making Sense of Microposts (#MSM2011)*, pages 22–32, 2011.

[20] D. Chakrabarti and K. Punera. Event summarization using tweets. In *Proc. 6th AAAI Int. Conf. on Weblogs and Social Media*, 2011.

[21] John Hannon, Kevin McCarthy, James Lynch, and Barry Smyth. Personalized and automatic social summarization of events in video. In *Proceedings of the 16th international conference on Intelligent user interfaces*, IUI '11, pages 335–338, New York, NY, USA, 2011. ACM.

[22] James Lanagan and Alan F Smeaton. Using twitter to detect and tag important events in live sports. pages 542–545. AAAI, 2011.

[23] A. Ekin, A.M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *Image Processing, IEEE Transactions on*, 12(7):796 – 807, july 2003.

[24] Mohamed Y. Eldib, Bassam S. Abou Zaid, Hossam M. Zawbaa, Mohamed El-Zahar, and Motaz El-Saban. Soccer video summarization using enhanced logo detection. In *Proceedings of the 16th IEEE international conference on Image processing*, ICIP'09, pages 4289–4292, Piscataway, NJ, USA, 2009. IEEE Press.

[25] Youness Tabii and Rachid OuladHaj Thami. A new method for soccer video summarizing based on shot detection, classification and finite state machine. In *Proceedings of The 5th international conference SETIT 2009: Sciences of Electronic, Technologies of Information and Telecommunications, IEEE 2009*, 2009.

[26] Xiaofeng Tong, Qingshan Liu, and Hanqing Lu. Shot classification in broadcast soccer video. *Electronic Letters on Computer Vision and Image Analysis*, 7(1), 2008.

[27] Li Li, Xiaoqing Zhang, Weiming Hu, Wanqing Li, and Pengfei Zhu. Soccer video shot classification based on color characterization using dominant sets clustering. In *Proceedings of the 10th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, PCM '09, pages 923–929, Berlin, Heidelberg, 2009. Springer-Verlag.

[28] M. Pavan and M. Pelillo. Dominant sets and pairwise clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(1):167 –172, jan. 2007.

[29] Shaojie Cai, Shuqiang Jiang, and Qingming Huang. A two-stage approach to highlight extraction in sports video by using adaboost and multi-modal. In *Proceedings of the 9th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, PCM '08, pages 867–870, Berlin, Heidelberg, 2008. Springer-Verlag.

[30] Z. Xiong, Regunathan Radhakrishnan, Ajay Divakaran, and T.S. Huang. Highlights extraction from sports video based on an audio-visual marker detection framework. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, page 4 pp., july 2005.

[31] Cunxun Zang, Qingshan Liu, Xiaofeng Tong, and Hanqing Lu. A framework for providing adaptive sports video to mobile devices. In *Proceedings of the 2nd international conference on Mobile multimedia communications*, MobiMedia '06, pages 37:1–37:6, New York, NY, USA, 2006. ACM.

[32] Qingshan Liu, Zhigang Hua, Cunxun Zang, Xiaofeng Tong, and Hanqing Lu. Providing on-demand sports video to mobile devices. In *Proceedings of the 13th annual ACM international conference on Multimedia*, MULTIMEDIA '05, pages 347–350, New York, NY, USA, 2005. ACM.

[33] Changsheng Xu, Jinjun Wang, Kongwah Wan, Yiqun Li, and Lingyu Duan. Live sports event detection based on broadcast video and web-casting text. In *Proceedings of the 14th annual ACM international conference on Multimedia*, MULTIMEDIA '06, pages 221–230, New York, NY, USA, 2006. ACM.

[34] Yifan Zhang, Changsheng Xu, YongRui, Jinqiao Wang, and Hanqing Lu. Semantic event extraction from basketball games using multi-modal analysis. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 2190 –2193, july 2007.

[35] Changsheng Xu, Jinjun Wang, Hanqing Lu, and Yifan Zhang. A novel framework for semantic annotation and personalized retrieval of sports video. *Multimedia, IEEE Transactions on*, 10(3):421 –436, april 2008.

[36] Minh-Son Dao and N. Babaguchi. Mining temporal information and web-casting text for automatic sports event detection. In *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, pages 616 –621, oct. 2008.

[37] James F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, November 1983.

[38] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[39] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *COMPUTATIONAL LINGUISTICS*, 22:39–71, 1996.

[40] Asa Ben-Hur and Jason Weston. A users guide to support vector machines. 609, November 2008.

[41] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

[42] Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38:367–378, 1999.

[43] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[44] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, CIKM '98, pages 148–155, New York, NY, USA, 1998. ACM.

[45] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features, 1998.

[46] Jerome H. Friedman. RECENT ADVANCES in PREDICTIVE (MACHINE) LEARNING.

[47] Soccerway. Soccerway players, November 2011. `http://www.soccerway.com`.

[48] Twitter. Twitter streaming api, May 2012. `https://dev.twitter.com/docs/streaming-apis`.

[49] Soccer corner. Premier league twitter hashtags, May 2012. `http://www.soccer-corner.com/en-gb/football-clubs`.

[50] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. 2011.

[51] Ted Pedersen. A wordnet stop list, February 2012. `http://www.d.umn.edu/~tpederse/Group01/WordNet/wordnet-stoplist.html`.

[52] Machine Learning Group at University of Waikato. Weka, December 2011. `http://www.cs.waikato.ac.nz/ml/weka/`.

[53] Jordi Mas and Gabriel Fernandez. Video shot boundary detection based on color histogram. *Processing*, 2003.

[54] Xinguo Yu, Kong Wah Wan, Hwee Keong Lam, and Chong Yee Lee. Time recognition of video clock for live event alert system. In *Proceedings of the Second International Conference on Internet Multimedia Computing and Service*, ICIMCS '10, pages 33–36, New York, NY, USA, 2010. ACM.

[55] Huiqing Liu, Jinyan Li, and Limsoon Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome informatics International Conference on Genome Informatics*, 13(0919-9454 LA - eng PT - Journal Article RN - 0 (Proteome) SB - IM):51–60, 2002.

[56] Changjing Shang and Dave Barnes. Combining support vector machines and information gain ranking for classification of mars mcmurdo panorama images. In *Proceedings of the International Conference on Image Processing, ICIP 2010, September 26-29, Hong Kong, China*, pages 1061–1064. IEEE, 2010.

[57] Bing Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.