

Article

# Innovative Approaches in Sports Science—Lexicon-Based Sentiment Analysis as a Tool to Analyze Sports-Related Twitter Communication

Fabian Wunderlich \* and Daniel Memmert 

Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Am Sportpark Müngersdorf 6, 50933 Cologne, Germany; d.memmert@dshs-koeln.de

\* Correspondence: f.wunderlich@dshs-koeln.de

Received: 12 November 2019; Accepted: 30 December 2019; Published: 7 January 2020



**Abstract:** Sentiment analysis refers to the algorithmic extraction of subjective information from textual data and—driven by the increasing amount of online communication—has become one of the fastest growing research areas in computer science with applications in several domains. Although sports events such as football matches are accompanied by a huge public interest and large amount of related online communication, social media analysis in general and sentiment analysis in particular are almost unused tools in sports science so far. The present study tests the feasibility of lexicon-based tools of sentiment analysis with regard to football-related textual data on the microblogging platform Twitter. The sentiment of a total of 10,000 tweets with reference to ten top-level football matches was analyzed both manually by human annotators and algorithmically by means of publicly available sentiment analysis tools. Results show that the general sentiment of realistic sets (1000 tweets with a proportion of 60% having the same polarity) can be classified correctly with more than 95% accuracy. The present paper demonstrates that sentiment analysis can be an effective and useful tool for sports-related content and is intended to stimulate the increased use of and discussion on sentiment analysis in sports science.

**Keywords:** sentiment analysis; opinion mining; social media; Twitter; computer science in sports; football

---

## 1. Introduction

Science, to some extent, has always been an image of society, as a changing world imposes changing research areas and questions. In sports science, the invention and development of new sports, changes in demographics and recreational behavior or—as outlined in our approach—digitization and increasing volume of sports-related data are just some aspects influencing opportunities and relevance of research. While the importance of traditional fields of sports science will remain, sports science also needs to be open to new and innovative fields of research.

Digitization has led to a massive increase of available data and data complexity in almost every aspect of life which is imposing new opportunities as well as new challenges and is often referred to as Big Data [1]. Sports science has started to adapt to the era of Big Data, for example, by using positional data or tracking data in match analysis of various sports such as football [2], tennis [3], and basketball [4]. The increased technical and computational effort going hand in hand with these data-driven research areas has given rise to the discipline of computer science in sports.

While the analysis of positional or tracking data has become possible through technical effort, other sources of data are becoming available through increasing human communication. One of these innovative fields is the analysis of online and social media data that have recently been investigated in

various domains. Examples are the use of Google searches for early detection of influence epidemics [5] or forecasting unemployment rates [6], the use of Twitter data to forecast elections [7,8] or stock market prices [9] and the use of Facebook in psychological experiments [10]. On the one hand, these examples demonstrate a hardly deniable potential of making use of online data in science. On the other hand, various methodological issues in this young field of research will need to be discussed making it a highly controversial topic. The present approach focuses on the possibility of analyzing textual online data as an innovative avenue in sports science.

Sentiment analysis (also referred to as opinion mining) refers to the extraction of subjective information from textual data by the use of algorithmic approaches [11,12]. Accurate techniques of sentiment analysis can be a mighty tool, given the huge quantity of available textual data and the time required to analyze text by human evaluation. While it is rather easy for a human being to understand opinions or attitudes expressed in written language, it is highly complex to imitate this human skill computationally in an algorithmic way. Consequently, sentiment analysis is assessed as “one of the fastest growing research areas in computer science” [12] (p. 16) that is driven by the “rapid growth of user generated content on the Web” [11] (p. 122). In a review on sentiment analysis [12], the authors evaluated the literature until 2016 and found that 99% of the papers on sentiment analysis have been published after 2004, making it a relatively young field of research. A good reflection of the state of the art in sentiment analysis is the international SemEval workshop annually announcing current research tasks related to semantic evaluation. It is evidence that research on semantic evaluation is a current and high developing research topic which—besides further areas—also focuses on research with regard to Twitter data [13,14].

In this context, it should be considered whether sports science is a field of research that could benefit from this development. Several application domains have been identified, namely society, security, travel, finance and corporate, medical, entertainment and other [12]. While entertainment also includes sports-related research, no individual sports-related category was identified. Likewise, the authors of another review [11] neither mention sports science as a subject in which sentiment analysis research has been done nor sports as an application area. We conducted a search for articles related to social media analysis by means of sentiment analysis for two well-known multidisciplinary sports journals: *The Journal of Sports Sciences* and the *European Journal of Sport Science*. For both journals, not a single article title included the terms “sentiment analysis” or “opinion mining.” Even for the term “social media,” only one survey on the use of social media by sports nutritionists in the *Journal of Sports Sciences* [15] and an editorial explaining the increased use of social media by the *European Journal of Sports Sciences* [16] were found. A search for the terms “sentiment analysis” or “opinion mining” in the Web of Science Core Collection resulted in 1114 articles, but not a single article resulted when filtering for the Web of Science Category “sport sciences.” Nevertheless, some literature linking social media analysis and sports exists, such as investigations on the use of Twitter by sports journalists, athletes or fans [17–19], but are not related to algorithmic sentiment analysis. Moreover, a few applications of sentiment analysis in sports forecasting exist [20–22], however, these studies are focused on applying sentiment analysis in predictive tasks instead of validating sentiment analysis methods.

In light of the existing literature, algorithmic analysis of social media content can be considered a widely unused approach in sports science so far, although numerous possible applications exist. Table 1 shows a list of disciplines in sports science that could profit from its potential for several research questions. Driven by the large body of literature on sentiment analysis in general, but the absence of articles in sports science, the present study validates the appropriateness of sentiment analysis approaches in a sports-related context focusing on tweets originating from the microblogging platform Twitter. Despite all the opportunities, caution and preliminary work is required as most methods of sentiment analysis have been developed for, validated in and used in other domains such as economics or politics, and taking over a tool from other domains without proper validation seems unreasonable.

**Table 1.** Research questions related to social media analysis that could profit from algorithmic solutions of sentiment analysis.

Discipline	Research Question
Sports ethics	How do fans think about ethical questions related to sports such as corruption in sports, criticism on Olympic Games, use of performance-enhancing drugs? How does public opinion change over time or after important events?
Sports economics	How do fans think about clubs and athletes? How can clubs or athletes optimize their online image in order to improve sponsoring value? Can social media content be valuable in sports forecasting? (i.e., in-play match forecasting, forecasting of individual player success)
Sports psychology	Which kind of moods or emotions of fans come up during sport competitions? What focus of attention do fans have when they watch, discuss or reflect sport games?
Computer Science in Sports	How can sports-related online content be efficiently extracted? What is the validity of sentiment analysis methods when analyzing sports-related content? Are there improvements in sentiment analysis methods considering the special characteristics of sports-related content?
Sports journalism	How do consumers react to media coverage of sports? What are current and relevant topics discussed by sports fans?

The discrepancy between extensive research on sentiment analysis methods and rare, if any, research applying it to sports science is evident and it may be reasoned that sports scientists are either not aware of the possibilities of sentiment analysis, not convinced by its accuracy, or deterred by its complexity. In this sense, there exists a trade-off between a high accuracy on the one side and a high practicability on the other side, as sentiment analysis methods would be used as a tool in sports science and not as the main subject of investigation. Techniques applicable in sports science consequently need an acceptable accuracy, however, even the most accurate technique is useless if not accessible or not possessing a reasonable degree of complexity. The goal is, thus, not to find or develop a gold-standard sentiment analysis technique in sports, but to validate the current applicability of easily accessible sentiment analysis tools. Therefore, we chose to limit our study to lexicon-based methods of sentiment analysis, analyzing and comparing three relatively easily accessible tools. Machine learning approaches or the development of improved methods of sentiment analysis are not within the scope of this study.

The contributions of this article are the following: 10,000 tweets written by users on Twitter and related to 10 top-class football matches were evaluated both manually and algorithmically in order to validate the accuracy of algorithmic sentiment analysis in football-related textual data. To the best of our knowledge, no study so far has focused on the validation of sentiment analysis tools based on manually annotated tweets from the sports domain. Moreover, an accuracy measure designed for real-world applications is introduced and evaluates the accuracy with regard to realistic sets of tweets instead of single tweets. The transfer of sentiment analysis to sports science, as well as future avenues and limitations of sentiment analysis in sports are discussed.

## 2. Methods

### 2.1. Dataset

Twitter [23] is one of the best known social media websites and can be described as a microblogging platform. Users can interact with tweets, which are short textual messages that are restricted to 280 characters and usually contain information and opinions on the latest news, current events or personal and social topics. To simplify the searchability of tweets, users can include so-called hashtags such as #LIVTOT if the tweet is related to the football match Liverpool vs. Tottenham. Twitter provides the possibility to obtain tweets containing certain words or hashtags as well as additional metadata via programming interfaces [24]. Tweets related to 10 football matches in national and international competitions have been collected in the time period between February and May 2019 via the real-time streaming API using the R package rtweet [25]. Ignoring retweets (reposted messages from other users) and non-English tweets, a total of more than 80,000 tweets were collected. Only matches from

top-class competitions, having a high sporting value and generating huge public interest including a sufficient number of tweets in the English language were chosen. A summary of the dataset is given in Table 2. For each match 1000 tweets were chosen randomly to be part of the validation dataset, thus consisting of 10,000 tweets.

**Table 2.** Summary information on the dataset.

Date	Match	Competition	Hashtag	Number of Tweets
20 February 2019	Atletico Madrid vs. Juventus Turin	Champions League	#AtletiJuve	5121
24 February 2019	Manchester United vs. Liverpool	Premier League	#MUNLIV	30,044
7 March 2019	Chelsea vs. Dynamo Kiev	Europa League	#CHEDYN	3262
31 March 2019	Cardiff City vs. Chelsea	Premier League	#CARCHE	13,114
31 March 2019	FC Liverpool vs. Tottenham Hotspur	Premier League	#LIVTOT	17,503
10 April 2019	Ajax Amsterdam vs. Juventus Turin	Champions League	#AJAJUV	2213
16 April 2019	FC Barcelona vs. Manchester United	Champions League	#FCBMUN	3615
1 May 2019	FC Barcelona vs. FC Liverpool	Champions League	#FCBLIV	3603
25 May 2019	RB Leipzig vs. Bayern Munich	German Cup	#RBLFCB	1252
29 May 2019	Chelsea vs. Arsenal	Europa League	#UELfinal	5096

## 2.2. Manual Annotation

Tweets were manually annotated by one of two human annotators classifying each tweet into one of the following four categories: “positive,” “negative,” “neutral” or “nonsense” [26]. The category nonsense was clearly defined and includes those tweets that did not have any relation to the match, were not written entirely in the English language or were incomprehensible. For the other categories, however, a certain degree of subjectivity cannot be excluded and no detailed formal definition of each category could be given. Manual annotators were asked to carefully evaluate the sentiment of the tweet expressed by the writer and categorize the sentiment of the tweet accordingly based on their understanding of the tweet.

To test for inter-rater agreement, 200 tweets were randomly chosen prior to the manual annotation, evaluated by both annotators independently and Cohen’s kappa was calculated.

## 2.3. Preprocessing of Data

Tweets contain a lot of information that cannot be usefully processed by lexicon-based sentiment analysis tools. Therefore, the following steps to clean the data—being in line with common approaches in the literature [27,28]—were executed automatically prior to the algorithmic evaluation:

- Removing all URLs (“<https://t.co/hSizQPxVFy>”) and mentions (“@CardiffCityFC”)
- Removing emoticons not analyzable as provided by the Twitter API in R (“<U+009F>”)
- Removing punctuation characters and digits
- Removing tabs, line breaks and unnecessary spaces
- Removing any non-ascii character
- Changing text to lower case
- Replacing contractions with full forms (“wouldn’t” → “would not”)
- Replacing acronyms with full forms (“omg” → “oh my god”)
- Correcting and doubling intentionally misspelled words that express some sort of intensification (“niiiiiiiiiiiiiiiiice” → “nice nice”)

Although it is not uncommon to completely remove hashtags, we kept them included to consider those hashtags that express a sentiment such as #boring (i.e., removing the #-sign, but not the word) at the cost of also including specific hashtags that will not be included in any lexicon such as #livetot. Intentionally misspelled words repeating the same character various times are interpreted as a sign of intensification. As lexicon-based methods are used, the word will not be identified if kept in the given state. Therefore, the word is corrected in order to be identifiable in the lexicon, but doubled to account for the underlying intensification.

As a result of preprocessing, an original uncleaned tweet

“It’s just 38” mins into the match and @ChelseaFCare @ChelseaFC like this I’m sorry to say this Chelsea are really loosing their standards as a big team Sarri lick <f0><U+009F><U+0091><U+0085>my ass like I care y’all trash <f0><U+009F><U+0097><U+0091> #CARCHE”

will be changed to this cleaned tweet after preprocessing

“it is just mins into the match and like this i am sorry to say this chelsea are really loosing their standards as a big team sarri lick my ass like i care you all trash carche”.

#### 2.4. Algorithmic Evaluation

The algorithmic evaluation of tweets can be seen as a classification problem in which the tweets are assigned to two (positive vs. negative) or more categories. In the literature, the methods used are commonly divided into two groups: lexicon-based approaches and machine learning approaches [11,28,29]. Lexicon-based methods rely on a predefined list of words (lexicon) defining the semantic orientation of each word, such as positive and negative words or words with a specific positivity and negativity score. This can be considered rather simplifying as the sentiment of a piece of text is directly derived from the set of contained words by a rule-based approach. In machine learning approaches, representation of text can contain several features including single words or longer sequences of words (also referred to as n-grams) and the sentiment of a piece of text is derived by using this representation as an input for machine learning techniques such as Support Vector Machines. We focus on lexicon-based approaches within this study, as our primary aim is to validate the transferability and feasibility of sentiment analysis as a tool in the domain of sports science and currently, lexicon-based tools are still easier accessible.

In its simplest form, lexicon-based approaches [28,29] are based on the idea of having an extensive list of words that are either considered to have a positive or a negative meaning. Given the textual data of a tweet, the total number of words as well as the number of positive and negative words can be counted. By dividing the number of positive (negative) words by the number of total words, a positivity (negativity) score of the tweet is determined. The total sentiment of the tweet is calculated by subtracting positivity and negativity. By performing a median split, the tweets having different decimal sentiment scores can then be classified into the two categories of positive and negative.

Three easily accessible tools of lexicon-based sentiment analysis are used within this study. The commercial LIWC 2015 software [30] (referred to as LIWC throughout the remaining paper) and the SentimentAnalysis package [31] available for the open source software R [32] and using the QDAP dictionary [33] (referred to as QDAP throughout the remaining paper), as well as a lexicon from the R *lexicon* package [34] based on the SenticNet4 lexicon going back to the work of Cambria et al. [35] (referred to as SN throughout the remaining paper). Moreover, a combined evaluation considering the averaged results of all three tools is calculated and referred to as COMB throughout the remaining paper. Please note that only the capability of these tools in identifying positive and negative words and deriving the total sentiment of a tweet using a rule-based approach is tested. In particular, no conclusions should be drawn on the general usefulness of the LIWC software having more functions than the ones considered here or on the SenticNet4 approach possessing a far more complex methodological procedure than just the lexicon itself used here. The following example tweet (after preprocessing) is used to demonstrate how the lexicon-based and rule-based methods work.

“should have lost but won champion material livtot”

The algorithm underlying the QDAP\_tool identifies all words categorized as positive in the lexicon (“won”, “champion”), all words categorized as negative in the lexicon (“lost”) and all words

being in none of these categories (“should,” “have,” “but,” “material,” “livtot”). The tweet consists of eight words including two positive and one negative word, thus yielding a positivity of  $2/8 = 0.25$ , a negativity of  $1/8 = 0.125$  and a total sentiment of  $0.25 - 0.125 = 0.125$ .

### 2.5. Negation Handling

Negation handling is a critical point in lexicon-based sentiment analysis as expressions like “this is not a good match” will lead to positive sentiment just by the presence of the word “good.” However, it is a complex topic as negation can apply to some parts of a statement while not applying to others. In line with our reasoning on the trade-off between accuracy and practicability, a very basic rule of negation flipping is used. If the tweet contains the negating word “not,” then the full sentiment of the tweet is reversed, i.e., positivity and negativity are flipped.

### 2.6. Accuracy Measures

A straightforward and common measure in sentiment analysis is the *accuracy* of sentiment classification with regard to a dataset of positive and negative tweets, which is—roughly speaking—the proportion of correct classifications. It is defined as “Accuracy =  $(TP + TN)/(TP + TN + FP + FN)$ ” [28] (p. 11) where TP (true positives) and TN (true negatives) refer to the number of tweets correctly classified (i.e., in line with the manual evaluation) as positive or negative by the algorithmic evaluation. FP (false positives) and FN (false negatives) refer to the number of tweets classified as positive or negative in contradiction to the manual evaluation. We excluded neutral and nonsense tweets from the test dataset and balanced the dataset in order to have the same number of positive and negative tweets [26]. The advantage of a balanced dataset is that the lower benchmark of accuracy is exactly 50% with reference to a random classification.

This approach of creating a validation dataset [26] and assessing the accuracy [28,29,36] is common in the literature and the natural choice when sentiment analysis techniques are the main subject of investigation as its accuracy can be expressed in a single straightforward value. However, it cannot be considered sufficient with regard to testing the applicability of sentiment analysis in the sports domain for two reasons: First, the exclusion of neutral and nonsense tweets creates an artificial validation dataset that does not correspond exactly to real-world datasets. Second, in real-world applications, there is no need to correctly classify the sentiment of a single tweet, but to classify the general sentiment of a set of hundreds or potentially thousands of tweets with a reasonable degree of accuracy. With regard to the aims of this study, more insights into the accuracy of classifying realistic sets of tweets as well as knowledge on how the accuracy depends on the structure of the tweet sets is needed.

In line with the objectives of this study, an alternative and more suitable accuracy measure denoted as *set accuracy* is introduced and calculated. Test sets of  $n$  tweets with a predefined number of tweets from each category are created. The proportion of neutral and nonsense tweets in the test sets equals the proportion in the validation dataset. For the remaining tweets in the test set, a predefined proportion  $p$  has the same polarity (either positive or negative). The tweets in the test dataset are randomly chosen from the validation dataset, taking into account the predefined distribution of tweets from the different categories, resulting in test sets with a predominantly positive or predominantly negative sentiment. Then, these test sets are classified as either positive or negative, based on the average sentiment of all tweets and this procedure is repeated  $m$  times. The calculation of the set accuracy follows the same formula as for the accuracy; however, TP and TN now refer to the number of test sets correctly classified by the algorithmic evaluation, and FP and FN refer to the number of test sets of tweets classified incorrectly.

### 3. Results

#### 3.1. Manual Annotation

In terms of accuracy, the agreement between both annotators was 82.5% when considering all four categories and 94.5% when considering only those tweets judged as either positive or negative by both annotators. The inter-rater reliability based on 200 pre-annotated tweets was found to be Kappa = 0.76 (95% CI (0.69, 0.83);  $p < 0.001$ ) which is considered to be a substantial agreement with reference to Landis and Koch [37] and even an excellent agreement with reference to Fleiss et al. [38]. A further improved agreement might be achievable when providing detailed formal definitions for the categories “positive,” “negative” and “neutral.” However, strict rules were deliberately not introduced, as the intention was to reflect human understanding and a higher agreement might thus come at the cost of artificially introducing biases to the understanding of the annotators.

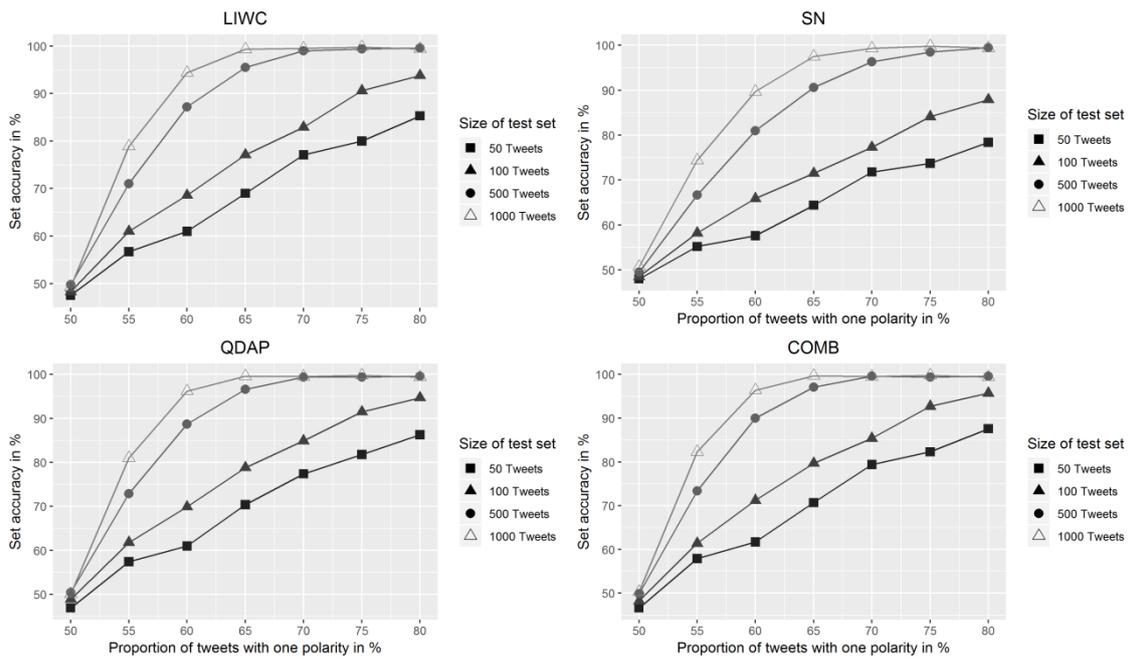
Manual annotation of the validation dataset resulted in 3184 positive, 2283 neutral and 3288 negative tweets. The remaining 1245 tweets were categorized as nonsense. Table 3 gives examples for tweets that were categorized in each of the categories.

**Table 3.** Example tweet in each category of the manual annotation.

Category	Tweet
Positive	“Messi is anything but human. #FCBLIV”
	“What a save @Alissonbecker #MUNLIV #pl #lfc”
	“C’mon United. #GGMU #MUNLIV”
Negative	“Stupidity from Ole for putting Lingard instead of Sanchez #MUNLIV”
	“Never a pen, minimal contact outside box on Costa, Atletico lucky to get a free kick #AtletiJuve #ucl”
	“Liverpool BATTERED #lfc #FCBLIV @bt sportfootball”
Neutral	“VAR overrules ref. Freekick rather a pen #AtletiJuve”
	“Bit surprised Klopp has broken up the Milner/Henderson/Wijnaldum midfield that’s played in most of the big games this season if it is Milner at right back today #MUNLIV”
	“Barcelona vs. Liverpool: Messi scores two goals in seven minutes #FCBLIV <a href="https://t.co/sZWqueaNIL">https://t.co/sZWqueaNIL</a> ”
Nonsense	“Looking for professional business flyer designer? Please contact following link. Messi #MUNLIV #AtikulsWinning HOLD THE DATE #SundayMorning #FelizDomingo <a href="https://t.co/jvMrrxjN4D">https://t.co/jvMrrxjN4D</a> ”
	“INVIOLEBLE LE BUT DE TER STEGEN #FCBLIV”
	“@TheSportsman The Savior has emerged... Imam #Ahmedalhasan (as). The Messenger from Imam Mahdi (as). The Mahdi (as) that is born during the end of time... The Messenger from Jesus (as). The Messenger from Elijah (as) #AtletiJuve #skamfrance <a href="https://t.co/Qq8eA54ESE">https://t.co/Qq8eA54ESE</a> <a href="https://t.co/6bGIO6DNEw">https://t.co/6bGIO6DNEw</a> ”

#### 3.2. Accuracy Measures

Accuracy was found to be 61.0% for LIWC, 63.6% for QDAP, 62.6% for SN and 67.4% for COMB. Accuracies are significantly higher than 50% according to one-sided binomial tests ( $p < 0.001$  for all four methods). A total of 43.4% of tweets for LIWC (25.0% for QDAP, 7.8% for SN) included neither an evaluable positive nor an evaluable negative word, which is further evidence for the limited informative value of analyzing single tweets. Figure 1 illustrates the set accuracy for various proportions  $p$ , various test set sizes  $n$  and  $m = 10,000$  test sets each. Results show that for sets of 50 or 100 tweets, even tweet sets with a clear imbalance in sentiments (75% of one polarity) are not sufficient to enable set accuracies to exceed 95% for any method. However, for the best-performing methods and sets of 500 tweets, a proportion of 65% from the same polarity is sufficient to correctly classify more than 95% of sets for both methods. For sets of 1000 tweets, even a proportion of 60% of one polarity is sufficient.



**Figure 1.** Set accuracies for various numbers of tweets (n) and proportions of polarity (p) based on 10,000 repetitions.

### 3.3. Qualitative Observations

Algorithmic sentiment analysis is a challenging task and several general issues have been mentioned in the literature, such as sarcasm detection and order dependence [28], handling of spelling mistakes and slang [39] or intensification and negation [29].

We will not repeat these known issues, but will demonstrate some qualitative observations and resulting challenges in sports-related Twitter data in line with the argument of Kharde and Sonawane [28] (p. 13) on the issue of domain dependence in sentiment analysis: “The same sentence or phrase can have different meanings in different domains. For example, the word ‘unpredictable’ is positive in the domain of movies, dramas, etc., but if the same word is used in the context of a vehicle’s steering, then it has a negative opinion.” Just as every other domain, sports in general and football in particular have their own vocabulary. We use the QDAP method to demonstrate some issues of football-related textual data being a consequence of not using a football-specific sentiment lexicon. Mentions of popular players such as Eden Hazard are categorized as negative due to the meaning of his last name. The terms “free kick” and “champions league” are categorized as positive and the term “penalty shootout” as negative, although in football-related communication none of these terms expresses any sentiment by itself. At the same time, terms with a clear connotation such as the negative term “diving” or the positive term “Mexican wave” will be categorized as neutral. Sports-specific lexica would help to overcome these domain-specific problems, yet not the general limitations in sentiment analysis. So far, the applicability has been investigated in various domains including banks, cars, computers, cookware, hotels/travel destinations, movies, music and phones [29,36], but we are not aware of any prior study focusing on the validation of manually annotated textual data from the sports domain.

Another sports-specific issue is the length of tweets. The database contains a significant quantity of tweets consisting of only a few words, such as the following examples:

“Epic #LIVTOT”  
 “Gol please. #MUNLIV”  
 “Lionel!!! #FCBLIV”

It may be presumed that in other domains, such as politics or finance, the average tweet length is higher than in the sports domain. To test this empirically, two comparative datasets were collected using a selection of politics-related hashtags (#brexit, #generalelection2019, #vote) and a selection of finance-related hashtags (#dowjones, #wallstreet, #stocks, #bitcoin). The average tweet length for cleaned tweets was 13.3 words for the football-related tweets, 20.8 words for politics-related tweets and 18.0 words for finance-related tweets. For the football-related data, a proportion of 49.3% of tweets contained ten or less words (politics: 28.1%, finance: 31.6%), and 21.6% of all cleaned tweets contained even five or less words (politics: 11.8%, finance: 8.8%). This is evidence that the tweet length can vary widely and—at least with regard to the investigated selection of hashtags—football-related tweets are rather short. The low number of words, and even lower number of evaluable positive or negative words, makes a correct classification difficult. The rather low accuracy of classifying single tweets thus might, to some extent, be an artefact of the low number of words in each tweet.

#### 4. Discussion

Based on the set accuracy, it can be said that the tools are accurate enough to correctly classify sets of a few hundreds or thousands of tweets. As an example, sets of 1000 tweets, where a proportion of 60% have the same polarity, are correctly classified in more than 95% of all cases. As a high number of tweets is by no means a problem in usual real-world applications, this result proves the validity of sentiment analysis methods in such applications.

As argued in the Methods section, the accuracy of classifying sets of tweets does have a higher relevance in real-world applications than the classification of single tweets, and a high set accuracy is achieved, although the accuracy with regard to single tweets is rather limited. The accuracy of the sentiment analysis methods (63.6% for the best performing method and 67.4% when combining all methods) were significantly higher than 50%. This shows the general capability to identify sentiments correctly in single tweets as well, but still appears to be a rather weak result, given that the lower benchmark of 50% would be achieved by guessing. To make a fair assessment of this result, it needs to be discussed in light of the experimental set-up and the existing literature. As the algorithm is intended to have the highest possible agreement to human evaluation, an upper benchmark of 100% is unreachable, given the fact that even manual annotators will not reach full agreement due to the influence of subjectivity. As demonstrated in the Results section, the inter-rater-agreement in terms of accuracy was 94.5%, if considering only those tweets judged as either positive or negative by both annotators. Comparability to results stated in the literature is highly limited, as the accuracy does not only depend on the method, but also on the domain and the type of textual data [29]. Longer pieces of text, for example, imply an easier task for correct sentiment classification, while tweets are very short pieces of text and thus only comparison to literature reporting sentiment classification tasks of Twitter data seems reasonable. Accuracies around 75% have been reported for such tasks [26,28], showing the general difficulty and relativizing the result found. Moreover, the aforementioned studies use machine learning techniques such as Support Vector Machines while the present study focused on techniques and tools that are purely lexicon-based and thus, easily available and feasible for non-computer scientists. The short length of football-related tweets, as described in the Results section, and further influence of domain dependence, as explained in the Methods section, could be other reasons for falling short of the accuracies reported in other domains. One area of future research could, therefore, be the improvement of methods and lexica for sports-related content.

With regard to the different tools used, QDAP was able to slightly outperform LIWC and SN in terms of accuracy. To make a fair assessment of this result, it needs to be said that the SentimentAnalysis package in R and the QDAP lexicon are specifically provided for sentiment analysis as performed within this study. LIWC is a software for text analysis that has a broad scope of functions and only one aspect (positive and negative emotions) was used within the study. Moreover, the difference in accuracy seems to be partly attributable to the rather high number of tweets not containing any evaluable positive or evaluable negative word identifiable by LIWC. Regarding SN, we did not test the

more complex SenticNet4 method as proposed by Cambria et al. [35], but just made use of a binary polarity lexicon (i.e., list of positive and negative words) deduced from their work.

With reference to the set accuracy, it can be said that all tools perform comparably in real-world applications and that the number of tweets, as well as the polarity in the test data set, had a higher influence on the set accuracy than the choice of the tool.

Sentiment analysis could have a significant impact in sports. Applications, as outlined in Table 1, are possible right now and—as shown within the present study—can be approached with easily accessible tools possessing a reasonable degree of complexity. In the future, opportunities of sentiment analysis might become even broader. Cambria, Schuller, Xia and Havasi [40] point out that sentiment analysis does not need to be bound to textual data and mention further applications of automatically extracting opinions from data such as “facial expression, body movement, or a video blogger’s choice of music or color filters” (p. 19). This is yet another link to sports science where body language plays a large role, however, not analyzed algorithmically so far. Other visions of sentiment analysis refer to the use of questionnaires: The necessity to find participants to fill out questionnaires and ask for their opinion might be avoidable as a virtually endless number of people give away their opinion on the internet voluntarily. The necessity to ask for recreational behavior in questionnaires might be avoidable if the data can be extracted from social media or related online data without having the problem of social desirability in questionnaires [41].

However, technological, methodological and ethical questions will need to be overcome in order to establish social media analysis in general and sentiment analysis in particular as a tool in sports science. Technical feasibility for non-computer scientists is crucial as tools need to be manageable. We addressed this issue by testing easily accessible tools possessing a reasonable degree of complexity. The possibly most serious methodological issues concern representativeness, as users of social media cannot be assumed to be a representative sample of the population, e.g., in terms of demographic groups or age, as has been criticized in the domain of election forecasting [7]. While studies like this one, analyzing existing textual data for sentiments excluding any personal data are unproblematic in terms of ethical questions, such questions would become relevant if metadata like personal characteristics or geographical information would be analyzed. Carrying out experiments manipulating the behavior of users in social media deliberately, would impose even further issues with regard to informed consent.

## 5. Conclusions

Being a rapidly growing research area in computer science, sentiment analysis is still pioneer work in sports science. The present approach has shown the potential impact of algorithmically analyzing textual online data in general and social media data such as Twitter in particular on sports science. The accuracy of classifying single tweets or small sets of tweets is not satisfactory and could profit from sports-specific lexica and the presence of more sophisticated sentiment analysis tools. However, despite domain-specific characteristics, easily accessible tools of lexicon-based sentiment analysis are capable of classifying the polarity of realistic sets of tweets (e.g., 1000 tweets, with a proportion of 60% having the same polarity) with an accuracy of more than 95%, thus being accurate enough in real-world applications when analyzing hundreds or thousands of tweets. We hope that our approach will stimulate the increased use of sentiment analysis as a tool in sports science.

**Author Contributions:** Conceptualization, F.W. and D.M.; data curation, F.W.; formal analysis, F.W.; methodology, F.W.; project administration, D.M.; resources, D.M.; writing—original draft, F.W.; writing—review and editing, D.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by an internal research funding scheme of the German Sport University Cologne (HIFF).

**Acknowledgments:** We would like to thank Marius Schürmann and Alessandro Seck for the manual evaluation of Twitter tweets.

**Conflicts of Interest:** No potential conflict of interest was reported by any of the authors.

## References

1. Mauro, A.D.; Greco, M.; Grimaldi, M. *What Is Big Data? A Consensual Definition and a Review of Key Research Topics*; AIP Publishing LLC: Melville, NY, USA, 2015; pp. 97–104.
2. Memmert, D.; Lemmink, K.; Sampaio, J. Current Approaches to Tactical Performance Analyses in Soccer Using Position Data. *Sports Med.* **2017**, *47*, 1–10. [[CrossRef](#)] [[PubMed](#)]
3. Wei, X.; Lucey, P.; Morgan, S.; Sridharan, S. Sweet-spot: Using spatiotemporal data to discover and predict shots in tennis. In Proceedings of the 7th Annual MIT Sloan Sports Analytics Conference, Boston, MA, USA, 6–7 March 2013.
4. Lucey, P.; Bialkowski, A.; Carr, P.; Yue, Y.; Matthews, I. How to get an open shot: Analyzing team movement in basketball using tracking data. In Proceedings of the 8th Annual MIT SLOAN Sports Analytics Conference, Boston, MA, USA, 30 April 2014.
5. Ginsberg, J.; Mohebbi, M.H.; Patel, R.S.; Brammer, L.; Smolinski, M.S.; Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature* **2009**, *457*, 102–104. [[CrossRef](#)] [[PubMed](#)]
6. D’Amuri, F.; Marcucci, J. The predictive power of Google searches in forecasting US unemployment. *Int. J. Forecast.* **2017**, *33*, 801–816. [[CrossRef](#)]
7. Gayo-Avello, D.A. Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data. *Soc. Sci. Comput. Rev.* **2013**, *31*, 649–679. [[CrossRef](#)]
8. Huberty, M. Can we vote with our tweet. On the perennial difficulty of election forecasting with social media. *Int. J. Forecast.* **2015**, *31*, 992–1007. [[CrossRef](#)]
9. Bollen, J.; Mao, H.; Zeng, X.J. Twitter mood predicts the stock market. *J. Comput. Sci.* **2011**, *2*, 1–8. [[CrossRef](#)]
10. Kramer, A.D.I.; Guillory, J.E.; Hancock, J.T. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 8788–8790. [[CrossRef](#)]
11. Piryani, R.; Madhavi, D.; Singh, V.K. Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Inf. Process. Manag.* **2017**, *53*, 122–150. [[CrossRef](#)]
12. Mäntylä, M.V.; Graziotin, D.; Kuuttila, M. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Comput. Sci. Rev.* **2018**, *27*, 16–32. [[CrossRef](#)]
13. Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; Kiritchenko, S. Semeval-2018 task 1: Affect in tweets. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 1–17.
14. Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel, F.; Rosso, P.; Sanguinetti, M. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 54–63.
15. Dunne, D.M.; Lefevre, C.; Cunniffe, B.; Tod, D.; Close, G.L.; Morton, J.P.; Murphy, R. Performance Nutrition in the digital era—An exploratory study into the use of social media by sports nutritionists. *J. Sports Sci.* **2019**, *30*, 54–63. [[CrossRef](#)]
16. Hendricks, S.; Jones, A. European Journal of Sport Science gears up its social media. *Eur. J. Sports Sci.* **2014**, *14*, 519–520. [[CrossRef](#)] [[PubMed](#)]
17. Sheffer, M.L.; Schultz, B. Paradigm Shift or Passing Fad. Twitter and Sports Journalism. *Int. J. Sport Commun.* **2010**, *3*, 472–484. [[CrossRef](#)]
18. Hambrick, M.E.; Simmons, J.M.; Greenhalgh, G.P.; Greenwell, T.C. Understanding Professional Athletes’ Use of Twitter. A Content Analysis of Athlete Tweets. *Int. J. Sport Commun.* **2010**, *3*, 454–471. [[CrossRef](#)]
19. Witkemper, C.; Lim, C.H.; Waldburger, A. Social Media and Sports Marketing. Examining the Motivations and Constraints of Twitter Users. *Sport Mark. Q.* **2012**, *21*, 170–183.
20. Schumaker, R.P.; Jarmoszko, A.T.; Labeledz, C.S. Predicting wins and spread in the Premier League using a sentiment analysis of twitter. *Decis. Support Syst.* **2016**, *88*, 76–84. [[CrossRef](#)]
21. Brown, A.; Rambaccussing, D.; Reade, J.J.; Rossi, G. Forecasting with social media. Evidence from tweets on soccer matches. *Econ. Inq.* **2017**, *20*, 1363. [[CrossRef](#)]
22. Godin, F.; Zuallaert, J.; Vandersmissen, B.; De Neve, W.; van de Walle, R. Beating the bookmakers. leveraging statistics and Twitter microposts for predicting soccer results. In Proceedings of the KDD Workshop on Large-Scale Sports Analytics, New York, NY, USA, 6 June 2014.
23. Twitter Inc. 2019. Available online: <https://twitter.com/> (accessed on 30 August 2019).
24. Twitter API. 2019. Available online: <https://developer.twitter.com/> (accessed on 30 August 2019).

25. Kearney, M.W. rtweet: Collecting Twitter Data. 2019. Available online: <https://CRAN.R-project.org/package=rtweet> (accessed on 30 August 2019).
26. Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; Passonneau, R. Sentiment Analysis of Twitter Data. In Proceedings of the Workshop on Languages in Social Media, Oregon, Portland, 23 June 2011.
27. Angiani, G.; Ferrari, L.; Fontanini, T.; Fornacciari, P.; Iotti, E.; Magliani, F.; Manicardi, S. A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. *KDWeb*. 2016. Available online: [https://pdfs.semanticscholar.org/09c0/136d4e3d9defc50a72253a967180e86be244.pdf?\\_ga=2.120249992.1431766548.1578324939-1266920968.1578324939](https://pdfs.semanticscholar.org/09c0/136d4e3d9defc50a72253a967180e86be244.pdf?_ga=2.120249992.1431766548.1578324939-1266920968.1578324939) (accessed on 16 August 2019).
28. Kharde, V.A.; Sonawane, S.S. Sentiment Analysis of Twitter Data: A Survey of Techniques. *Int. J. Comput. Appl.* **2016**, *139*, 5–15. [[CrossRef](#)]
29. Taboada, M.; Brooke, J.M.; Voll, K.M. Lexicon-Based Methods for Sentiment Analysis. *Comput. Linguist.* **2011**, *37*, 267–307. [[CrossRef](#)]
30. Pennebaker, J.W.; Boyd, R.L.; Jordan, K.K. *The Development and Psychometric Properties of LIWC2015*; University of Texas at Austin: Austin, TX, USA, 2015.
31. Feuerriegel, S.; Proelochs, N. SentimentAnalysis: Dictionary-Based Sentiment Analysis. 2019. Available online: <https://CRAN.R-project.org/package=SentimentAnalysis> (accessed on 30 August 2019).
32. R Core Team. R: A Language and Environment for Statistical Computing. 2017. Available online: <https://www.R-project.org/> (accessed on 30 August 2019).
33. Rinker, T.W. *qdapDictionaries: Dictionaries to Accompany the qdap Package*; University at Buffalo: Buffalo, NY, USA, 2013.
34. Rinker, T.W. *Lexicon Data*; Buffalo: Buffalo, NY, USA, 2018.
35. Cambria, E.; Poria, S.; Bajpai, R.; Schuller, B. SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives. In Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, COLING 2016, Osaka, Japan, 11–16 December 2016; pp. 2666–2677.
36. Turney, P.D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Pennsylvania, PA, USA, 7–12 July 2002; pp. 417–424.
37. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [[CrossRef](#)]
38. Fleiss, J.L.; Levin, B.; Paik, M.C. *Statistical Methods for Rates and Proportions*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2003.
39. Feldman, R. Techniques and applications for sentiment analysis. *Commun. ACM* **2013**, *56*, 82. [[CrossRef](#)]
40. Cambria, E.; Schuller, B.; Xia, Y.; Havasi, C. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intell. Syst.* **2013**, *28*, 15–21. [[CrossRef](#)]
41. Furnham, A. Response bias, social desirability and dissimulation. *Personal. Individ. Differ.* **1986**, *7*, 385–400. [[CrossRef](#)]

