Analyzing Premier League Supporters Opinion on VAR

1st Axel Karlsson
Department of Computer and Information Science
Linköpings Universitet
Linköping, Sweden
axeka379@student.liu.com

***Abstract*—This work conducts a sentiment analysis on Premier League supporters opinion on VAR. Using data from twitter in the form of tweets, it is investigated if the overall opinion from the supporters is negative or positive. After the tweets have been cleaned and made available for analysis, a model is trained on tweets that have their sentiment manually annotated and are then used to classify the tweets from the Premier League supporters to find the sentiment on each tweet. It is also investigated if the league position of the team and the number of VAR decisions for each team has any impact. The results show that the overall opinion of the supporters regarding VAR is negative and that the combination of a lower league table position and VAR decisions given against a team have an impact on the supporters reaction, although it is not clear which factor has the largest impact. It is concluded that more data from a full season is needed in order to confirm the results.**

***Index Terms*—VAR, Sentiment Analysis, Machine Learning, Premier League**

I. Introduction

The need for video analysis in football has always been a debated topic. Refereeing decisions can at times decide the outcome of a game or even a whole season for a team. Errors in these decisions are therefore critical to keep at a minimum, and will be the source of causing great controversies when not being able to correctly carry them out. Notable examples of this are Sulley Muntaris wrongly denied goal against Juventus [3] and Thierry Henrys controversial handball goal against Ireland. [1].

In 2019, the english Premier League decided to implement VAR and use it for the matches being played for the season 2019/20. The decisions when VAR will be used and what it will look at primarily are:

- Goal/no goal
- Penalty/No penalty
- Direct red card
- Mistaken identity (where the wrong player could have been sent off)

During the season supporters have voiced their opinions on the usage of VAR and whether or not they actually believe it has improved the game.

This report will investigate the opinions of the supporters for the teams in Premmier League and see if they are in favor of the introduction of VAR or not. In order to achieve this, data from Twitter (tweets) will be used to get the opinions from supporters of different teams. Then a sentiment analysis will be performed to classify if the tweet is positive or negative, in order to determine if the overall opinion of VAR is positive or negative. Tweets linked to each team will also be analyzed to see if there are some teams supporters that are more positive or negative, compared to the others.

VAR statistics for each team will be collected as well, in order to see if there is any factors that are more contributing to supporters being negative or positive. For example, supporters could be expected to be more negative to VAR if their team has gotten a large amount of calls against them.

II. Theory

This section will briefly discuss theory relevant to the report.

A. TF-IDF

Since the characters constructing words in a text does not contain much information on their own, a machine learning model must make use of a algorithm to make sense of the data. Term Frequency Inverse Document Frequency (TF-IDF) is one method of doing so. TF-IDF determines the relative frequency of words in a specific document compared to the inverse ratio of that word over the entire document. [10]

B. Stop Words

The stop words, are words that occur frequently but do not contribute much of value in the text. Data processing can consist of cleaning texts of stop words with the goal of making the texts more compact with only the necessary information available. [8]

C. Models

There exists several machine learning models that can be used for working with classification tasks. Some models might be better in comparison to others for solving a certain task. In this report the focus will be on models that are used for classification and they are obtained from the Scikit Learn library. [1]

[1] scikit-learn.org/stable/

1) Naive Bayes: Naive Bayes classifier is based on Bayes theorem which describes the probability of an event. Bayes theorem can be seen in Equation 1. Where **A** and **B** are events and P(**B**) is seperated from zero. [5]

$$P(\mathbf{A}|\mathbf{B}) = P(\mathbf{A})\frac{P(\mathbf{B}|\mathbf{A})}{P(\mathbf{B})} \tag{1}$$

The multinomial naive Bayes implements bayes theorem for data which is multinomially distributed.

Say we predict the class of a document, then the same features are extracted. Lets say we have a class Y_k where k is all the possible classes and lets say we have a word $x \in X$ where X is the dependent feature vector. Then the probability of a certain class can be seen in Equation 2. [13]

$$P(y_k, \mathbf{X}) = P(y_k)\prod_{i=1}^{n} a_i P(x_i \mid y_k) \tag{2}$$

2) Logistic Regression: Logistic Regression is linear model which can be used when working with classification. It extends the multi-class linear model, which can be seen in Equation 3, where $\hat{y}$ is the output vector, x is the input vector, W is the weight matrix and b is the bias vector. [7]

$$\hat{y} = x * W + b \tag{3}$$

The logistic model extension can be seen by the adding of a pointwise logistic function f [9]

$$\begin{aligned} &\hat{y} = f(z) \\ &where \\ &z = x * W + b \end{aligned} \tag{4}$$

The logistic regression measures a relationship between a categorical dependent variable and one or more independent variable by estimating probability of an occurrence by using its logistic function.

3) Random Forest Classifier: Random forests are a combination of tree predictors where each tree depends on the values of a randomised vector which is sampled independently and has the same distribution for all the trees in the forest.

What defines the random forest classifier is that it consists of a collection of tree-structured classifiers.

$$\{h(x, \Theta_k), k = 1, ...\} \tag{5}$$

$\{\Theta_k\}$ are independent identically distributed random vectors and each tree will cast a unit vote for the most popular class at input x. [4]

D. Accuracy

A common evaluation criteria used together with machine learning models, is accuracy. It is calculated using true positives **(TP)**, true negatives **(TN)**, false positives **(FP)**, false negatives **(FN)** as can be seen in Equation 6.

$$\frac{\mathbf{TP} + \mathbf{FN}}{\mathbf{TP} + \mathbf{TN} + \mathbf{FP} + \mathbf{FN}} = Accuracy \tag{6}$$

III. Related work

Go, Alec, Richa Bhayani, and Lei Huang [6] performed a sentiment classification of tweets as positive or negative by using machine learning models. The dataset used in the work contained noisy labels in the form of emoticons and is shown to be an effective way of performing distance supervised learning. Machine learning models that where used in the work managed to achieve an accuracy above 80%. The preprocessing of the tweets performed in the work served as inspiration for this report and similar techniques where used.

IV. Method

This section will discuss the method that was carried out during this work.

A. Prerequisites

At the beginning of this work, information of how the VAR decisions have gone is gathered in order to see which teams have gotten the most decisions for and against them. The idea is that supporters of a team that have received calls against them, will be more negative when tweeting about VAR. In this work we will primarily focus on the decisions and also the league position of the team, in order to find out if there is a trend in causing more negativity from these two factors.

TABLE I
The decisions by VAR for each of the teams

Team	For	Against
Crystal Palace	3	4
Tottenham Hotspur	1	3
Burnley	1	1
AFC Bournemouth	1	3
Manchester United	2	2
Manchester City	1	1
Arsenal	1	0
Leicester City	0	1
Liverpool	3	0
Everton	2	0
Southampton	5	0
Chelsea	0	1

Not enough data is present for all the teams that took part in Premier League 2019/20 and 2020/21, simply due to the fact that there did not exist tweets for these teams regarding VAR. Instead the teams that are available will be used in this work which will be seen in the Results in Section V.

B. Data

1) Tweets: The data that is used in this work consists of tweets from supporters of teams in the Premier League. These tweets are obtained from a dataset on Kaggle[2] [12] where the tweets have been collected in the period of 2020-07-09 - 2020-10-13. The tweets have been categorized to be linked to a specific premier league team by using hashtags (#). It can be noted that the period that the tweets were written in range from match days 34-38 in the 2019/2020 season and match

[2]www.kaggle.com

TABLE II
THE PREMIER LEAGUE TABLE 2019/20 AFTER MATCH DAY 38

Barclays Premier League Table		
Position	Club	# Points
1	Liverpool	99
2	Manchester City	81
3	Manchester United	66
4	Chelsea	66
5	Leicester	62
6	Tottenham Hotspur	59
7	Wolves	59
8	Arsenal	56
9	Sheffield United	54
10	Burnley	54
11	Southampton	52
12	Everton	49
13	Newcastle	44
14	Crytal Palace	43
15	Brighton	41
16	West Ham	39
17	Aston Villa	35
18	Bournemouth	34
19	Watford	34
20	Norwich	21

days 1-4 in the 2020/2021 season. In Fig 1 the total number of tweets distributed for each team, can be seen. This dataset will further on be referred to as the ”kaggle dataset”.

A second dataset is also obtained in order to train a classifier. This dataset also contains tweets and it has the purpose to be used to train a classifier for sentiment analysis. The tweets in this dataset are categorized as either positive, represented as 4, or negative, represented as 0. The dataset was proposed in a paper by Saif, Hassan, et al [11] as an attempt to improve the distinctive sentiment annotation that can occur in tweets and its entities. This dataset will further on be referred to as the ”training dataset”.

C. Hypothesis

A hypothesis of what the results are expected to look like is given based on the prerequisites of the teams. The idea is that supporters of teams that have a lower table position, as can be seen in Table III, will have a more negative attitude towards VAR. This is also applied to teams that have received VAR decisions against them as can be seen in Table IV-A. A combination of these two factors should naturally result in a negative response from supporters but here we will also investigate if one factor is more dominant than the other, based on the results.

As an example, Southampton has had 5 calls in favor and 0 calls against them while maintaining a mid-table position.

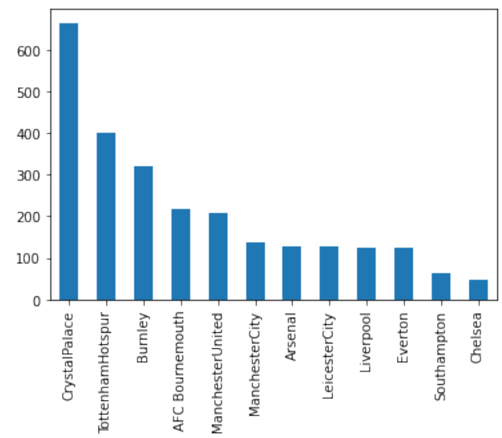

Fig. 1. The number of tweets made by each team

They should therefore have a positive attitude based on the preconditions, at least in comparison to the other teams.

Bournemouth has had 1 decision for them and 3 against and they ended the season in 18th place which meant they where relegated from the Premier League. According to the hypothesis given, this should result in them having a very negative attitude towards VAR.

For viewing the table position of the teams during all the match days, this paper refers the reader to the official Premier League website[3]

1) VAR decisions: The decisions that have affected the teams during the matches will be gathered manually, using the news source ESPN[4] who have gathered all the VAR decisions made for and against each team during the season. In this work we will use the decisions that have been made during the match days that are relevant to this report.

D. Preprocessing data

The data, which consists of tweets, must be ”cleaned” before it can be used for analysis. Tweets can contain special characters, emojis, HTML tags or other data which is not necessary or relevant for the sentiment analysis of the data. Therefore the tweets are cleaned of these characters since it could possibly cause error in the analysis.

The steps involved in cleaning the data were the following:

1) *Removing unwanted columns*: The kaggle dataset consisted of several columns including team names, tweet text, id, etc. The only relevant columns used in this work are the actual tweets and their associated team name.
2) *Removing noise*: Removing unwanted characters for example ’@’, ’#’, HTML-tags and also emojis.
3) *Filtering releant words*: Only the tweets containing the words ’VAR’, ’video assistant referee’ etc, where used in this analysis.
4) *Removing retweets* All of the retweets, characterized by having ’RT’ in the beginning, are removed.

[3]www.premierleague.com/
[4]www.espn.co.uk

5) *Tokenizing*: The tweet data was tokenised using NTLK tokeniser which was obtained from the NTLK-package. In the same package, a stop word list exists that could be used.
6) *Stemming*: Eliminates circumfixes, suffixes, prefixes and infixes from a word in order to obtain a word stem. Example: 'Working' becomes 'Work'. For this task, Porter Stemmer from NTLK was used.
7) *Vectorizing data*: Converts the tokens to numbers in order for the machine learning models to work with the data. Vectorization is done using tf-idf mentioned in Section II-A.

E. Training models

The models were available and used from the scikit-learn package.[5] Models for *Multinomial Naive Bayes*, *Logistic Regression* and *Random Forest Classifier* where implemented and evaluated against each other to see which one could obtain the highest accuracy. This task was performed with the aid of the training dataset which was mentioned in Section IV-B. The dataset was split into two smaller sets which where used to first train the classifier on one of them and then to test the classifier with the other part in order to see how accurately it could predict the sentiment compared to what it actually was.

The models had their primary parameters tuned using the class GridSearchCV which performs an exhaustive search of the best parameter for each estimator using cross validation. The model used for classification, was chosen based on the highest accuracy which it achieved when trying to classify the training data as positive or negative.

F. Evaluation

When the model which achieved the highest accuracy on the training dataset was chosen, the task would then be to use it to classify the tweets in the Kaggle dataset. The task of the model would then be to classify the tweets as positive or negative. The overall sentiment of VAR will then be measured to see what it has been classified as. A comparison for each of the teams measured sentiment is then performed in relation to the hypothesises that where given in the prerequisites to see if trends between league position/VAR decisions and sentiment, can be seen.

V. Results

This section will describe the results which were obtained when executing the method of the project. The code written for this work is available at a Github repository [2]

A. Preprocessed data

The total number of tweets used in the Kaggle dataset, after filtering had been applied as discussed in Section IV-D, was 2561. In comparison the training dataset consisted of 2034 tweets.

[5]scikit-learn.org/stable/

B. Model Evaluation

The results from evaluating the models accuracy on the training dataset can be be seen in Table III.

TABLE III
ACCURACY FOR EACH MODEL

Model	Accuracy
Naive Bayes	0.82
Logistic Regression	0.83
Random Forest	0.79

C. Classification Results

Logistic Regression was used for classifying the tweets and resulted in 2047 tweets classified as negative and 514 as positive. The distribution of the sentiments can be seen in Fig 2.

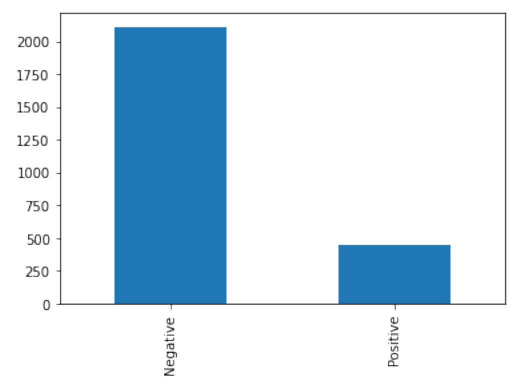

Fig. 2. The total distribution of positive and negative tweets regarding VAR

The results for each team can be seen in Table IV.

TABLE IV
THE NEGATIVE AMOUNT OF TWEETS FOR EACH TEAM

Team	Negative Tweet Amount
Crystal Palace	81.0%
Tottenham Hotspur	83.5%
Burnley	81.5%
AFC Bournemouth	86.2%
Manchester United	81.1%
Manchester City	74.4%
Arsenal	82.0%
Leicester City	70.0%
Liverpool	77.7%
Everton	86.4%
Southampton	73.0%
Chelsea	19.5%

Everton was the team with the most negative amount of tweets which can be seen in Table IV and the distribution of their tweets can be seen in Fig 3. Chelsea had the lowest amount of negative tweets and the distribution can be seen in Fig 4. The distribution for all of the different teams can be seen in Appendix A.

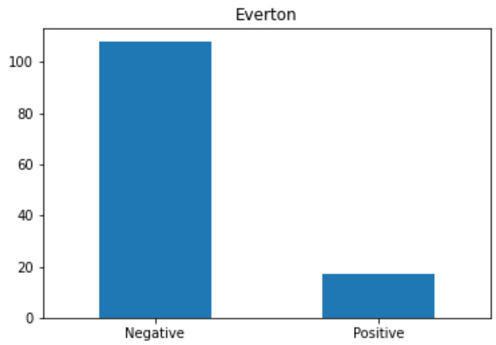

Fig. 3. The distribution of positive and negative tweets regarding VAR for Everton

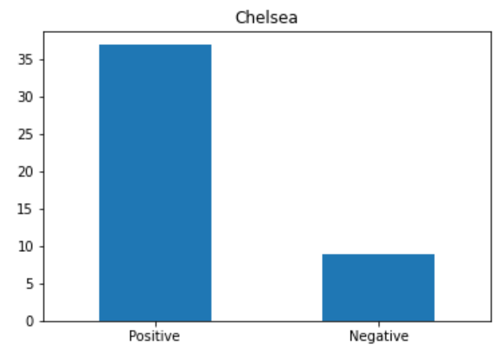

Fig. 4. The distribution of positive and negative tweets regarding VAR for Chelsea

VI. Discussion

In this section the data, results and method of the carried out work will be discussed. Interpretations of the results will be given and the method will be criticised. The limits of the datasets used will also be discussed.

A. Data

The dataset attained from kaggle contained a large amount of tweets but it was clear that after filtering for relevant tweets, mentioning VAR, that the amount of tweets decreased severely. More data containing VAR would be preferrable but would require an extra step in retrieving and there is also the fact that VAR has only been in use for one season so there is a limit to how much there exists as well.

The training dataset which was used to train classifiers, was considered a good match based on the litterature [11] mentioned previously. It also consisted of tweets which contributed in the choice of using it for the model training. The dataset contained more negative tweets than positive and this could of course have an effect when training the models. It also only contained positive and negative tweets while for this work it could be of interest to also classify tweets as "neutral" as some tweets could just be mentioning VAR without taking a stance on the subject.

B. Results

The results showed, as was expected, that the overall impression of VAR among supporters is negative. This does not seem surprising as supporters and media have had a negative stance towards VAR since its introduction.

What seems surprising are the results of Chelsea-supporters mentioning VAR in mostly positive light. With only about 20% of the data being negative. At the same time Everton was the club where most tweets had been classified as negative although they during this period had 2 VAR calls in favor of them and none against them. One should note however, that both of these clubs had fewer tweets available compared to the rest of the teams. More data should be available to properly evaluate the opinions of the teams supporters. It should also be taken into consideration that the model achieved a classification score of 83% when training, so there will most likely be some erroneous classification.

Bournemouth is the second most negative towards VAR and they have more data available, compared to Everton. They have received 3 calls against them and were relegated from the Premier League. We can compare this to Tottenham whom also received 3 calls against them but maintained a higher league position and we can see here that Bournemouth have a higher percentage of negative tweets in comparison.

Overall it is hard to draw a clear conclusion from whether which factor actually contributes the most but there does seem to be a pattern where the combination of calls against the team and lower league table position will often result in more negative opinions. Although the results for Everton might show otherwise, it can be argued that the amount of data is not enough to draw a clear conclusion from that team.

More data for all the teams would have been preferred in order to properly draw a good conclusion on these results.

C. Method

Training the classifiers on a different dataset could result in better results when classifying tweets and achieving higher accuracy for the classifiers could also be possible.

In this work the classifier with the highest accuracy was chosen to do the classification. An alternative method would be to use each of the classifiers to classify the tweets and see how the results from each one corresponded to the given hypothesis.

The usage of a hypothesis was conducted in order to see if there where clear factors that contributed to negative opinions using VAR and can also work as a point of comparison for evaluating how well the classifiers performed. In this work the hypothesises where used as a starting point and the work tried to see if they actually withheld. Issues may be to come up with a clear enough hypothesis that holds true so for this work the idea was to keep it simple but still something manageable.

It should also be noted that in this work, not only the opinions of supporters regarding a certain team will be regarded. No clear distinction is made between who supports a team and who does not. It may be the fact that journalists, commentators or other people are tweeting about a team which they do not have any affection to. It could still show, however, in which light VAR is portrayed for a certain team.

VII. Future work

Redoing this work it would be interesting to gather data over a whole season, that is from matchday 1 to 38. This would possibly result in enough data for all the teams in and could result in different results than where achieved in this work. It would also open up the possibility to examine if the teams ”size” is a contributing factor to the sentiments. Bigger clubs might have a more toxic fan base than the smaller clubs and this could possibly have the effect that negative reactions towards VAR are more apparent.

It could also be interesting to follow a specific team over a season and see if the sentiment changes over time when results are going in favor of the team and when they are going against the team. Then it could be interesting try and investigate if some certain factors have a larger impact on the mood of the supporters.

VIII. Conclusion

In this paper the objective has been to try and classify the sentiment of tweets mentioning VAR used in the english Premier League, in order to try and see the general opinion among supporters. Using a dataset containing tweets with their polarity manually annotated, machine learning models where trained and evaluated to be used in classifying tweets. A Logistic Regression classifier was chosen to perform the classification and was used to classify tweets regarding the Premier League and VAR. The work finds that the league position and VAR calls against the team, when combined, has an effect on how negative the teams supporters are against VAR. The work cannot confirm which one of the factors has the largest impact, although it is concluded that the combination of lower league position and more decisions against the team leads to a more negative opinion from supporters.

References

[1] Lizzy. Addley, Esther. McDonald Henry. Davies. Irish fury at thierry henry’s handball in world cup qualifier. *The Guardian*.

[2] Karlsson Axel. text-mining. https://github.com/Axeka379/text-mining, 2021.

[3] Nicky Bandini. Milan’s ’ghost goal’ against juventus leaves serie a on a knife-edge. *The Guardian*.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Bradley Efron. Bayes’ theorem in the 21st century. *Science*, 340(6137):1177–1178, 2013.

[6] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[7] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[8] Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008.

[9] Kuhlmann Marco. Slides: Text classification, November 2020.

[10] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. New Jersey, USA, 2003.

[11] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold. 2013.

[12] William Jiang . Epl teams - twitter sentiment dataset, 2014. Twitter Sentiment about Teams in the English Premier League, https://www.kaggle.com/wjia26/epl-teams-twitter-sentiment-dataset.

[13] Yaohui Zhang, Guofeng Yan, and Sailing He. Optical fiber spectrometer based on smartphone platform for refractive index sensing application. In *Asia Communications and Photonics Conference*, pages AF3C–3. Optical Society of America, 2016.

Appendix

Appendix A: Team Sentiment Data

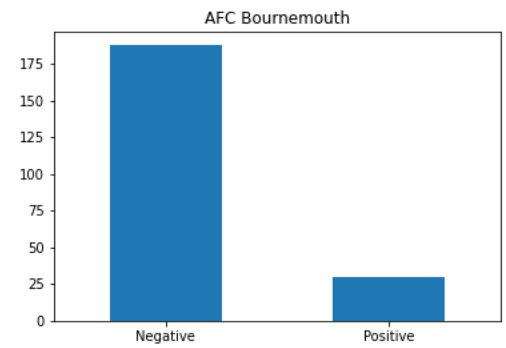

Fig. 5. The distribution of positive and negative tweets regarding VAR for Bournemouth

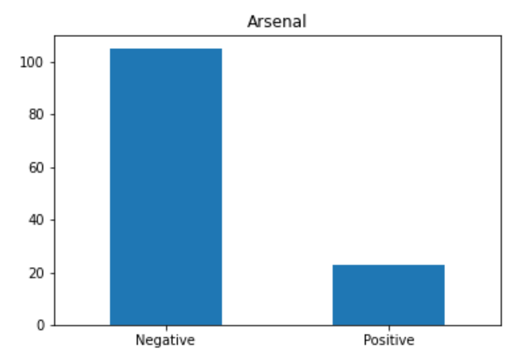

Fig. 6. The distribution of positive and negative tweets regarding VAR for Arsenal

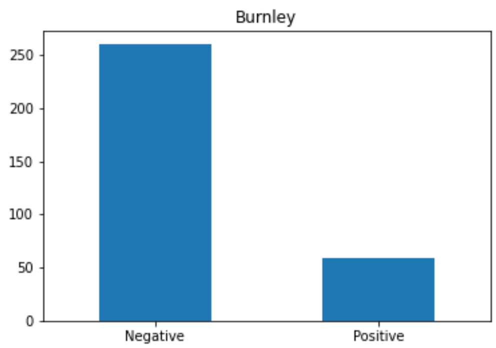

Fig. 7. The distribution of positive and negative tweets regarding VAR for Burnley

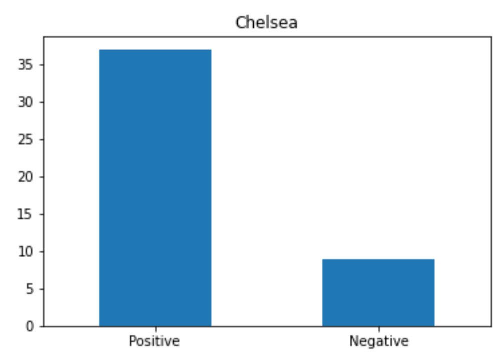

Fig. 8. The distribution of positive and negative tweets regarding VAR for Chelsea

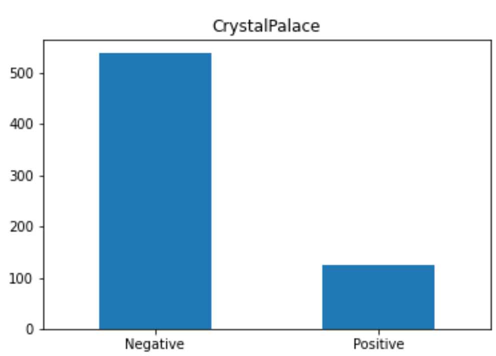

Fig. 9. The distribution of positive and negative tweets regarding VAR for Crystal Palace

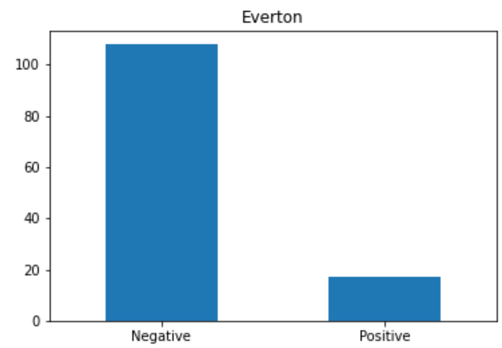

Fig. 10. The distribution of positive and negative tweets regarding VAR for Everton

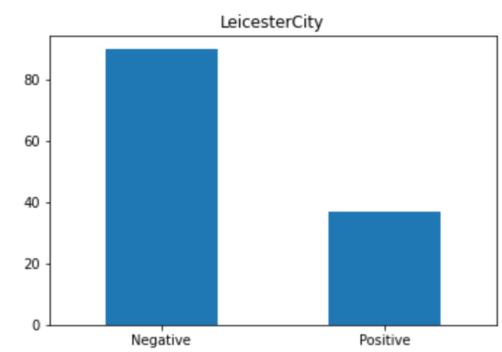

Fig. 11. The distribution of positive and negative tweets regarding VAR for Leicester City

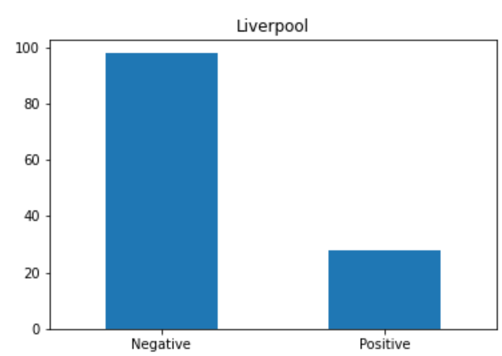

Fig. 12. The distribution of positive and negative tweets regarding VAR for Liverpool

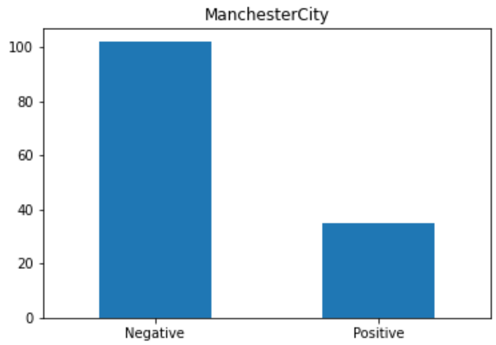

Fig. 13. The distribution of positive and negative tweets regarding VAR for Manchester City

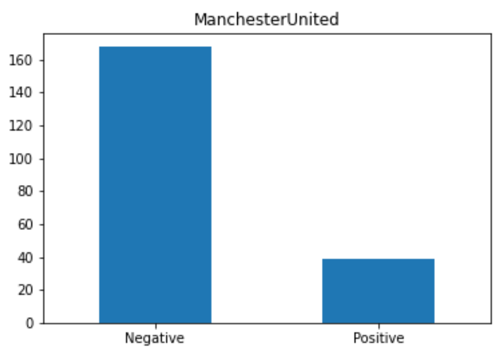

Fig. 14. The distribution of positive and negative tweets regarding VAR for Manchester United

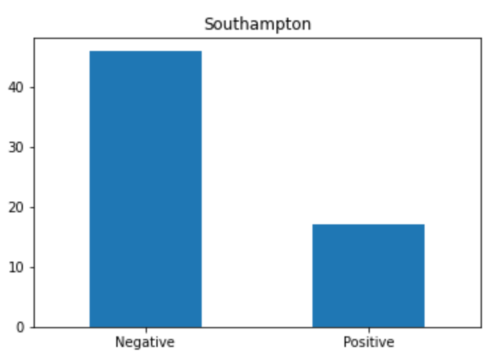

Fig. 15. The distribution of positive and negative tweets regarding VAR for Southampton

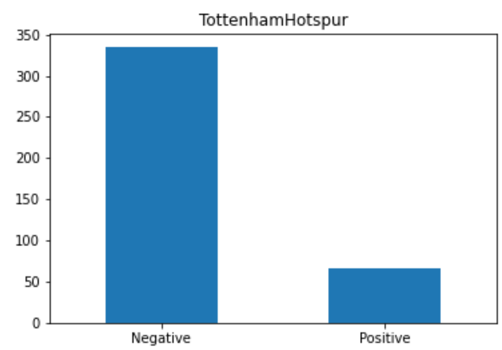

Fig. 16. The distribution of positive and negative tweets regarding VAR for Tottenham Hotspur