Football league win prediction based on online and league table data

IOP Conf. Series: Materials Science and Engineering **263** (2017) 042068 doi:10.1088/1757-899X/263/4/042068

Football league win prediction based on online and league table data

Prateek Par, Ankit Kumar Gupt, Samarth Singh, Neelu Khare and Sweta Bhattachrya

VIT University, Vellore, Tamilnadu-642014, India.

E-mail: neelu.khare@vit.ac.in

Abstract. As we are proceeding towards an internet driven world, the impact of internet is increasing in our day to lives. This not only gives impact on the virtual world but also leave a mark in the real world. The social media sites contains huge amount of information, the only thing is to collect the relevant data and analyse the data to form a real world prediction and it can do far more than that. In this paper we study the relationship between the twitter data and the normal data analysis to predict the winning team in the NFL (National Football League).The prediction is based on the data collected on the on-going league which includes performance of each player and their previous statistics. Alongside with the data available online we are combining the twitter data which we extracted by the tweets pertaining to specific teams and games in the NFL season and use them alongside statistical game data to build predictive models for future or the outcome of the game i.e. which team will lose or win depending upon the statistical data available.
Specifically the tweets within the 24 hours of match will be considered and the main focus of twitter data will be upon the last hours of tweets i.e. pre-match twitter data and post-match twitter data. We are experimenting on the data and using twitter data we are trying to increase the performance of the existing predictive models that uses only the game stats to predict the future.

1. Introduction

Information from social media has been utilized to anticipate and clarify miscellaneous collection of genuine wonder which includes, opinion polls, elections, spreading of transmissible viruses and stock markets. It gives us the proof that Twitter posts contains very important data which can be combined with Statistical approaches. Along these lines, Twitter may offer an approach to handle the Knowledge of group for improving expectations about genuine occasions. In this paper, we consider the connection between National Football League (NFL) recreations and the Twitter messages saying the groups required, all together to make forecasts about diversions. We concentrate on the NFL on the grounds that amusements are communicate broadly on TV all through the US and groups play at generally once every week, empowering many to remark on diversions by means of online networking. NFL football additionally has dynamic betting markets. One of the most famous and surely understood is the point spread line, which is an harm for the more grounded group picked by bookmakers to yield break even with wagers on both sides. Considering in the bookmaker's bonus, a wagering technique that predicts the victor with the spread in over 53% of amusements will be gainful. In this paper we prepare prototypes which will forecast the game and staking results, seeing various twitter properties which routine Twitter plus game numerical records.

2 Literature Survey

2.1Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis:-

In this paper we studies that the pre-processing methods affects the sentimental polarity classification from the twitter. They lead a progression of investigations utilizing four classifiers to confirm the viability of a few pre-preparing strategies on five Twitter datasets. The test comes about show that the evacuation of URLs, the expulsion of stop words and the expulsion of numbers negligible influence the execution of classifiers; besides, supplanting invalidation and growing acronyms can enhance the order exactness. In this manner, evacuating stop words, numbers, and URLs is fitting to diminish clamour, however, does not influence execution. Swapping refutation is viable for slant investigation. We select proper pre-handling strategies and highlight models for various classifiers for the Twitter conclusion characterization undertaking.

2.2 Rating Prediction based on Social Sentiment from Textual Reviews:-

In this paper, a suggestion model is proposed by mining assumption data from twitter clients' surveys. They utilize client feeling comparability, relational estimation impact, and thing notoriety closeness into a brought together lattice factorization system to accomplish the rating expectation errand. Specifically, they utilize social clients' opinion to indicate client inclinations. Their motivation of this approach is to discover viable hints from surveys and foresee social clients' evaluations. In this paper, they right off the bat separate item highlights from client audit, and after that present the strategy for distinguishing social clients' opinion and depicts the three nostalgic elements. Finally, we held every one of them into a slant based rating forecast strategy (RPS).this approach is partitioned in sub-area which are Extracting Product Features in which information pre-processing for LDA (a Bayesian model) is done after that generative procedure is a set up after that they separate item highlights. In User Sentimental Measurement. Investigated, tuned, adjusted and coordinated with a specific end goal to fabricate the general framework for movement occasion identification. Among the investigated classifiers, they have demonstrated the prevalence of the SVMs, which have accomplished exactness of 95.75%, for the 2-class issue, and of 88.89% for the 3-class issue, in which we have likewise considered the activity because of outer occasion class. This grouping model has been utilized for on-going observing of a few zones of the Italian street. We have likewise demonstrated the after effects of an observing effort, performed in September and early October 2014.

2.3 Enhanced Sentiment Learning Using Twitter Hash tags and Smiley's:-

In this paper, the individual proposes a directed notion characterization structure which depends on information from Twitter, a prevalent micro blogging administration. By using 50 Twitter labels and 15 smiley's as notion marks, this system keeps away from the requirement for work escalated manual explanation, permitting recognizable proof and characterization of assorted notion sorts of short writings. We assess the commitment of various include sorts for notion order, what's more, demonstrate that our system effectively distinguishes notion sorts of untagged sentences. The nature of the notion recognizable proof was likewise affirmed by human judges. We likewise investigate conditions what's more, cover between various notions sorts spoke to by smiley's and Twitter hash tags. The pattern of results got by examination of shared component vectors is like those acquired by methods for name co-event, despite the fact that the quantities of the common elements are higher.

3. Related Work

Each NFL standard seasons have the durations of 17 weeks from September to January. Only a single game is played each week. In every match the home group plays at their own particular stadium and

hosts the away group. The most well-known bet in NFL football is to pick the group that will win given a specific incapacitate called the point spread. Point spread refers to the quantity set which is decided by the Bookmarker who scramble handicap with the Home group. It is added to the house group's score, and after that the group with the most focuses is known as the champ with the spread (WTS). For instance, if the NY Giants are facilitating the NY Jets and the point spread is −4, then the Giants should win by no less than 4 keeping in mind the end goal to win WTS.

Point spreads and over/under lines are set by sports betting agencies to reflect all publicly available information about upcoming games, including team performance and the perceived outlook of fans. Assuming market efficiency, one should not be able to devise a betting strategy that wins often enough to be profitable. In prior work, most have found the NFL point spread market to be efficient overall, or perhaps only slightly inefficient. Others pronounced more conclusively in favour of inefficiency, but were generally unable to show large biases in practice. Regardless of efficiency, several researchers have designed models to predict game outcomes.

3.1 Proposed Method

To predict the winner of the current NFL season we collect the data from the current season and from that we can know the current performance and the current form of the team. The data that is fetched from the social media site i.e. twitter has been categorized by using **Naïve Bayes classification** . The tweets collected were refined by removing the re-tweets and collecting the information in an organised way in an excel sheet so that we can run the data and differentiate between the positive and negative tweets of each team. Then the result set of each of the tweets is plotted in a bar chart and scatter plot using Weka Tool.

$$P(\text{positive}|\text{tweet}) = (P(tweet|positive)P(positive)/\,P(tweet))$$

4. Moulding and Training

We utilize a calculated relapse classifier to foresee diversion and wagering results, in request to gauge the execution of our capabilities. For this purpose, we have following set of features which can be used to give us the starting point for the predication. The features are:-

1. Statistical Game Features

2. Twitter Unigram Features

4.1 Data Gathering

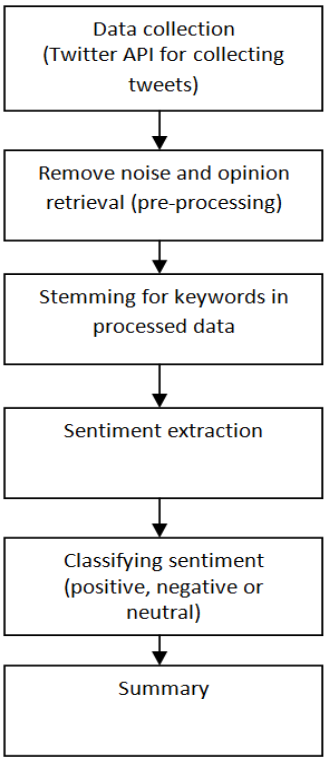
The data used in the prediction of football league is fetched from the NFL statistics obtained from www.nfl.com which consists of the league table data which consists of the current points scored by each team, total touch downs, total chances created, score of each player, and the points detail of the WTS. We then used twitter data collected from www.twitter.com as our other source of online data. An average of 60 million tweets is for each season is produced. This large amount of data can't be processed. So we are using R-Studio to fetch the relevant data. To fetch the data from twitter we tokenize the data i.e. we use the positive and negative dictionary to classify the positive and negative

tweets. We are using the data from the current season. The data collected is of the recent week. Each negative and positive tweets are used to predict the current mentality of the crowd.

For data collection hash tags are used to classify the data for each team. The tweets are arranged for each team based on the data dictionary containing positive and negative words and using **Naïve Bayes** classification using R-Studio. For example New York giant's tweets will be like #giants #new york giants #NY giants and negative tweets be like #lose giants #beat giants Using this we can exploit the prediction system and prepare a more accurate set of results by combining the current prediction plus the sentimental analysis of the people obtained by their tweets.

4.2 Finding Relevant Tweets

Our current classification is based on the positive and negative tweets obtained by twitter assigned for the particular team. Time stamp is used to fetch the most recent trends in twitter and link each team with the twitter obtained which is the hard part. We created a list of hash tags used by the each or against the team and ran through the R-studio and validated the search queries on twitter.
We focused our tweets for each game to predict the outcome of each game rather than calculating the outcome of the final result of the season. We connected the twitter feeds obtained for each team and calculated the number of tweets for and against the team. Using this data we can predict the outcome of the game.
For each game we classified the tweets obtained in the last 12 hours before the game and compared the statistics with the outcome of the result based on the tweet analysis. The most used dataset is of the pregame data which was gathered before an hour of starting the game. We verified our prediction with the results and post-game tweets. When we are making certain prediction on a certain team we use only the tweets that are assigned to the current team or relevant to the current match.



5. Results and Discussion

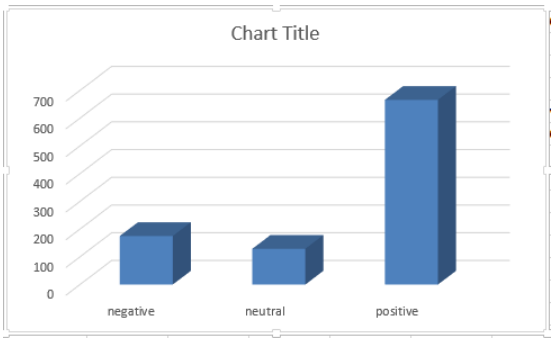


Fig. 1. Raider's Positive, neutral and negative tweets

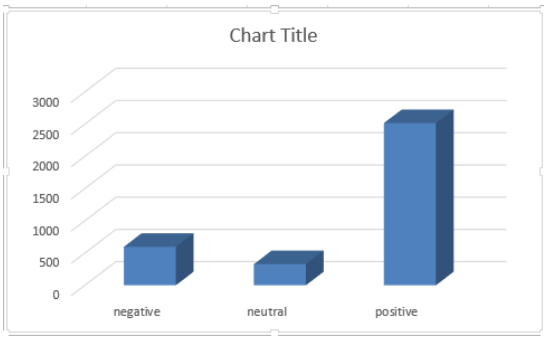


Fig. 2. Patriot's Positive, neutral and negative tweets

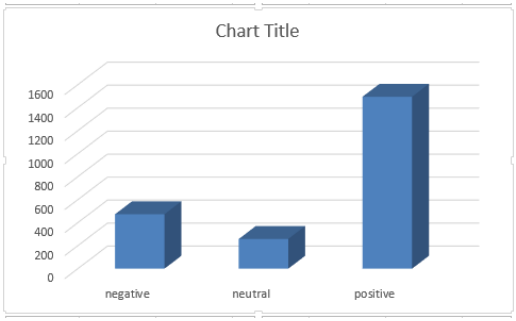


Fig. 3. Chief's Positive, neutral and negative tweets

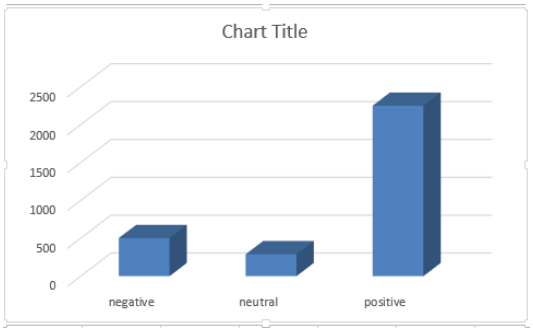


Fig. 4. Steeler's Positive, neutral and negative tweets

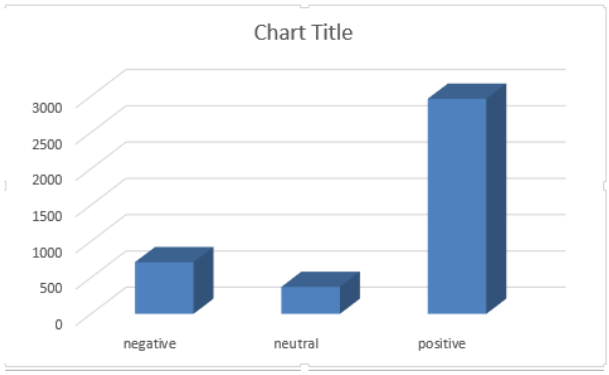


Fig. 5. Texan's Positive, neutral and negative tweets

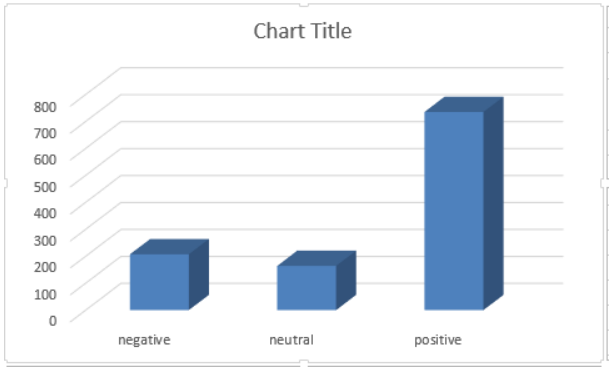


Fig. 6. Dolphin's Positive, neutral and negative tweets

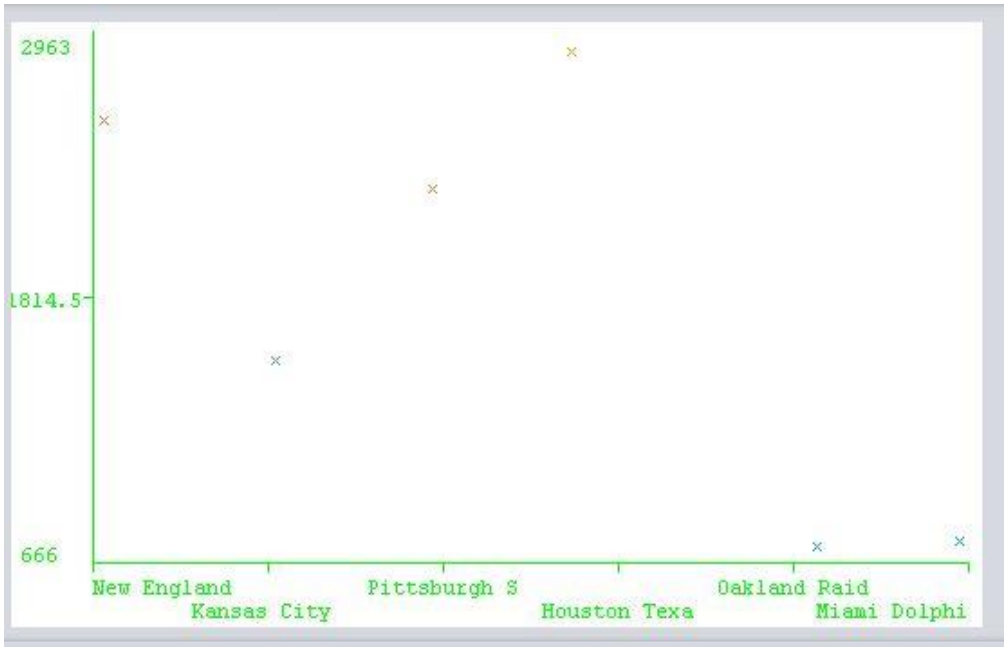


Fig. 7. Scatter Plot Of all the teams for positive tweets

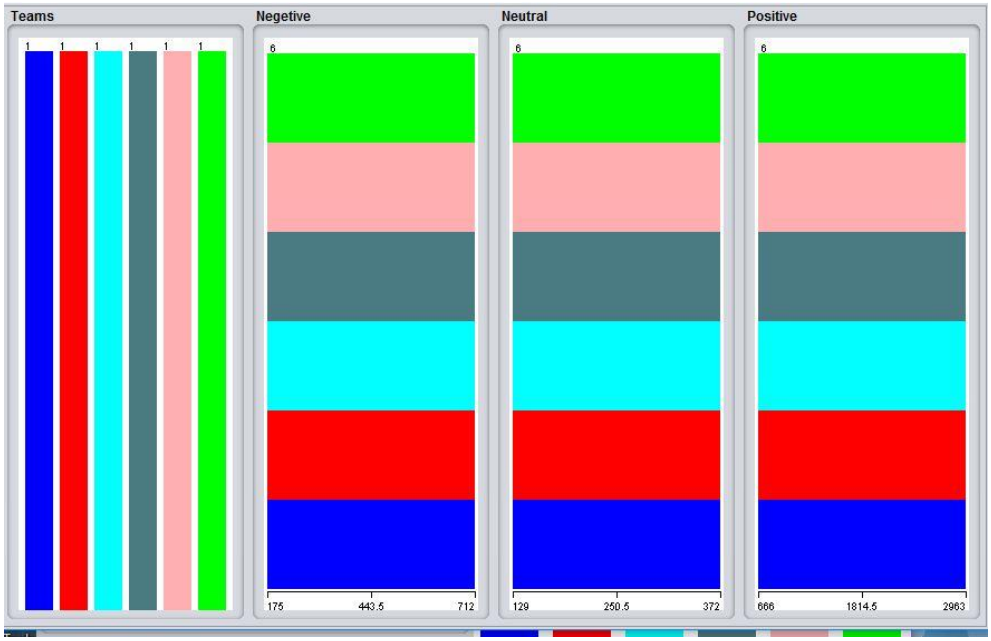


Fig. 8 Visualization of data using Weka Tool

6. Conclusion

The analysis confirmed our assumption on how effective an approach twitter sentiment analysis is. The Naive Bayes classifier used in the algorithm, along with two software for better results depict clearly the sentiment of the mass crowd and thus the airlines could easily interpret the data and benefit from it by trying to improve on the aspects that seem negative or is disliked by the targeted audience. There is still scope for improvement in this analysis since it is very new and yet has not been tested on many other classifying models. And the major setback is the limit in the number of tweets to be analysed using AYLIEN in Rapid Miner being 1000 tweets a day for a free user otherwise one has to opt for plans. So in the future we are planning to further expand our research and analysis by gather a huge number of data and expanding the process of data mining involved in this analytical approach.

7. References

[1] Z. Zhao, C. Wang, Y. Wan, Z. Huang, J. Lai, "Pipeline item-based collaborative filtering based on Map Reduce" 2015, IEEE Fifth International Conference on Big Data and Cloud Computing, 2015, vol **18,** pp. 1-12

[2] S. Gaol, Z. Yu, L. Shi, X. Yan, H. Song, "Review expert collaborative recommendation algorithm based on topic relationship," IEEE/CAA Journal of Automatic Sonica, 2015,vol **2**(4): pp. 403-411.

[3] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, M. Stayers, "Learning author-topic models from text corpora," ACM Transactions on Information Systems (TOIS). vol.**28**, no.1. 2010, pp. 1-30.

[4] K. K. Fletcher, X.F Liu, "A collaborative filtering method for personalized preference-based service recommendation," 2015 IEEE International Conference on Web Services (ICWS), 2015 vol **26**, pp. 400-407.

[5] Y. Chen, A. Cheng, and W. H. Hsu, "Travel recommendation by mining people attributes and travel group types from community-contributed photos," IEEE Trans. Multimedia 2013, vol. **15**, pp. 1-13

[6] R. Salakhutdinov, and A. Minch, "Probabilistic matrix factorization," in NIPS, 2008.pp. 1-8

[7] X. Yang, H. Stock, and Y. Liu, "Circle-based recommendation in online social networks, " in Proc. 18th ACM SIGKDD Int. Conf. KDD, New York, NY, USA, Aug. 2012,vol **173**, pp. 1267–1275
.
[8] M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu, and S. Yang, "Social contextual recommendation," in proc. 21st ACM Int. CIKM, 2012, vol **19**,pp. 45-54.

[9] M. Jamal and M. Ester, "A matrix factorization technique with trust propagation for recommendation in social networks," in Proc. ACM conf. Recasts, Barcelona, Spain. 2010, vol **170**, pp. 135-142.

[10] Z. Fu, X. Sun, Q. Liu, et al., "Achieving Efficient Cloud Search Services: Multi-Keyword Ranked Search over Encrypted Cloud Data Supporting Parallel Computing," IEICE Transactions on Communications, 2015, vol **98**(1):pp. 190-200.

[11] G. Gnu, N. Elhadad, A Marian, "Beyond the stars: Improving rating predictions using Review text content," in 12th International Workshop on the Web and Databases (Webb 2009),vol **9**, pp. 1-6.

[12] J. Xu, X. Zhen, W. Ding, "Personalized recommendation based on reviews and ratings alleviating the sparsely problem of collaborative filtering," IEEE International Conference on e-business Engineering. 2012, vol. **18**,pp. 9-16.