Computational linguistics and data analysis of the football specific tweets using LDA

C.H. Bhargava Phanindra, Palash Chaturvedi, Nirnai Rai, S.R. Rajeswari
Department of Computer Science,
SRM Institute of Science and Technology, Chennai, India.

Abstract: Soccer fans generate a large range of tweets to replicate their views and emotions on what's happening throughout numerous soccer matches round the globe. The tweets replicate the reactions of the fans throughout the commencement of the sport. Grouping the information of the sentiment expressed throughout can facilitate to grant complete image that expresses fan interaction throughout a specific soccer event. LDA is a generative applied mathematics model that permits sets of observations to be explained by unobserved teams that designate why some elements of the information square measure similar. LDA was accustomed to confirm what quite topics typically trend on twitter. Conducting in depth experiments on dataset to match the performance of various learning algorithms in distinguishing the sentiment expressed in soccer connected tweets. Topic modeling has been accustomed to confirm the subject of the tweets regarding soccer news.

IndexTerms - **Latent Dirchilet Allocation; topic modeling; twitter data; soccer.**

I. INTRODUCTION

Social media has become the necessary source and a source of gathering data regarding something that's happening or is going on within the world. Twitter being one amongst the foremost wellliked websites for varied events happening around provides vast varied sorts attention-grabbing topics and necessary data. The topics that area unit retrieved from tweets represents the trends, hot topics and varied alternative events that area unit happening round the world. Football is that the most well-liked sport within the world these days. soccer could be a most well-liked game of the planet even within the fashionable time. it's a most enjoyable and difficult game usually contend by two groups for the recreation and pleasure of the youths. the game has advanced to become the fashionable soccer as. This sport is also the foremost well-liked round the world. For the of analyzing the contexts within which soccer is documented, topic modeling victimization LDA are often used. Latent Dirichlet Allocation refers to the assumed likelihood distribution over the x topics assumed that area unit gift within the dataset's texts. This technique is acceptable once the text knowledge retrieved from twitter is analyzed, presumptuous every body of text to solely embody many topics. whereas running, the algorithmic program then adjusts the distribution of topics gift within the knowledge by learning from the gathering of text that's being obtained from varied tweets. Topic modeling helps in exploring massive amounts of text knowledge, finding clusters of words, similarity between documents, and discovering abstract topics. . The "topics" created by topic modeling techniques area unit clusters of comparable words. a subject model captures this intuition in an exceedingly mathematical framework, that permits examining a group of documents and discovering, supported the statistics of the words in every, what the topics could be and what every document's balance of topics is.

II. IDENTIFY, RESEARCH AND COLLECT IDEA

Topic modelling has been extensively used for research purposes in various fields such as making business models [1] on the basis of the topic modelling sometimes sentiments are being analyzed [2]. Topic modeling exploitation Twitter knowledge has conjointly been conducted by some researchers before [3][4][5]. Topic modeling of tweet knowledge has its own challenges compared with different text knowledge because of their unstructured language type and non-standard style of language. LDA methodology has been applied to search out topics on Twitter and there have been some new approaches to boost the performance of LDA [3][6].Mohammad F. A. Bashri and Retno Kusumaningrum [7] analyzed sentiments by using LDA and topic polarity wordcloudvisualization. Victoria Ikoro, Maria Sharmina, Khaleel Malik and Riza Batista Navarro [8] analyzed sentiments expressed on twitter by UK energy company consumers using two sentiment lexica. Kai Yang, Yi Ca Ho-fung Leung and Raymond LAU [9] analyzed topic-indiscriminate words in discovered topics using TWLDA.

III. ANALYSIS OF TOPICS AND SENTIMENTS

1. Proposed Model

The proposed system uses LDA to effectively model topics so that sentiment analysis can be performed on the data. But there are many other processes associated before we can effectively apply the LDA algorithm. The first step is collecting data of the sentiments expressed during a certain event. The second step is to pre-process and clean the noisy data so that it can be processed properly. The data is separated into batches. LDA topic modelling is applied to the pre-processed data. The LDA topic modelling is again applied to get higher topic accuracy. The results obtained from the LDA algorithm can be visualized in the next step. This visualization allows us to infer important information regarding the various topics from the data and also allows us to analyze it.

2. Implementation

2.1 Data Retrieval

Online data repositories like 'Kaggle.com' and 'Dataquest.io' provides users with many datasets. Thus, we can retrieve the necessary data sets from these websites. These sites allow the users to download datasets in different formats such as .csv , .txt , .xlsx etc. We have used .csv format for our implementation.

2.2 Data Preprocessing

The text which is retrieved is firstly split into sentences and then again split into words. These words are changed into lowercase and the punctuations removed in a process called as tokenization with the help of regular expressions in Python. In the next step all stopwords such as “the”, “at”, “is”, “which” etc. are removed from the set of words. The words are also lemmatized where the words in third person are changed to first person as well as verbs in past and future tenses are changed into present. Then the words are stemmed where they are reduced to their root form.

2.3 Topic Modelling using LDA

Topic modeling is the one of the most powerful methods in text mining that aims to identify patterns and find relationship among data from a collection of text documents [10]. The most popular method in topic modeling is LDA. LDA has been proven to be an effective unsupervised learning methodology for finding different topics in text documents [11]. LDA topic modeling is an unsupervised technique in machine learning which first introduced by Blei, et al [12] as a generative probabilistic model for text corpus.

Latent Dirichlet Allocation is a generative probabilistic model. It represents documents as a mixture of topics. Topic Modelling refers to the task of identifying topics that best describe a set of documents. Topics emerge during the topic modelling process. Each document can be described by a distribution of topics and each topic can be described by a distribution of words.

2.4 Data Visualization

The distribution of topics which is gained after LDA modelling can be visualized by using matplot and pyLDAvis libraries. The visualization allows for the comparison of topics on two reduced dimensions and observe the distribution of words in topics. From this visualization results can be accessed.

IV. RESULTS AND DISCUSSION

The results which were obtained after performing topic model are discussed in this section. We had used ‘FIFA World Cup 2018 Tweets’ dataset from Kaggle for the implementation of LDA. We get the following inferences from our results.

```
Topics found via LDA:

Topic #0:
saves penalty gn gl gg knocked shootout penalties schmeichel kasper

Topic #1:
japan reach want past argentina portugal spain germany coming win

Topic #2:
family incondicionales thank welcome support member newest penalty upsets goalkeepers

Topic #3:
play russia finals spain quarter world final https switzerland sweden

Topic #4:
power world cup exo ambassador got music turn russia play
```

Figure 3. Topics found after LDA implementation

We can visualize these topics further using WordCloud and seaborn histogram graphs. These are libraries available in python which can be used for visualization.

Figure 4 Output Visualization in WordCloud

As we can see ‘newest’, ’member’ and ‘welcome’ are the words which were used the maximum number of times. This can also be seen from the histogram.

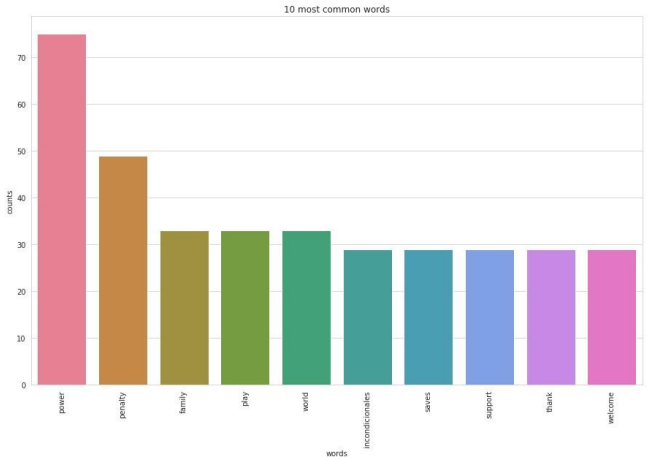

Figure 5 Histogram visualization

REFERENCES

[1] "Business reviews classification using sentiment analysis" by AndreeaSalinca published in 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing.
[2] "Sentiment Analysis Using Latent Dirichlet Allocation and Topic Polarity WordcloudVisualization" by Mohammad F. A. Bashri and Retno Kusumaningrum in 2017 Fifth International Conference on Information and Communication Technology (ICoICT).
[3] "Latent Dirichlet Allocation" by David M. Blei, Andrew Y. Ng and
[4] "Part of Speech Features for Sentiment Classification based on Latent DirichletAllocation
[5] Information Tech., Computer, and Electrical Engineering (ICITACEE), Oct 18-19, 2017, Semarang, Indonesia.
[6] S. Aloufi, F. Alzamzami, M. Hoda, and A. E. Saddik, ''Soccer fans sentiment through the eye of big data: The UEFA champions league as a case study'' in year 2018.
[7] "Sentiment Analysis Using Latent Dirichlet Allocation and Topic Polarity WordcloudVisualization" by Mohammad F. A. Bashri and Retno Kusumaningrum in 2017 Fifth International Conference on Information and Communication Technology (ICoICT).
[8] "Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers" by Victoria Ikoro, Maria Sharmina, KhaleelMalik and RizaBatistaNavarro in 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS).
[9] "Exploring Topic Discriminating Power of Words in Latent Dirichlet Allocation" by Kai Yang, Yi Ca Ho-fungLeung and Raymond LAU in COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2238– 2247, Osaka, Japan, December 11-17 2016.
[10] H. Jelodar, Y. Wang, C. Yuan, and X. Feng, "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey," 2017.
[11] L. Bolelli, Ş. Ertekin, and C. L. Giles, "Topic and trend detection in text collections using latent dirichlet allocation," Lect. Notes Comput.Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5478 LNCS, pp. 776–780, 2009.
[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.
[13] D. M. Blei, "Probabilistic topic models," Commun. ACM, vol. 55, no.4, pp. 77–84, 2012.