Breaking News on Social Media

Instructor: Prof. Daniel Acuna

Team Members: Aditi Chawla, Aditya Chauhan, Niti Saluja, Zaid Khan

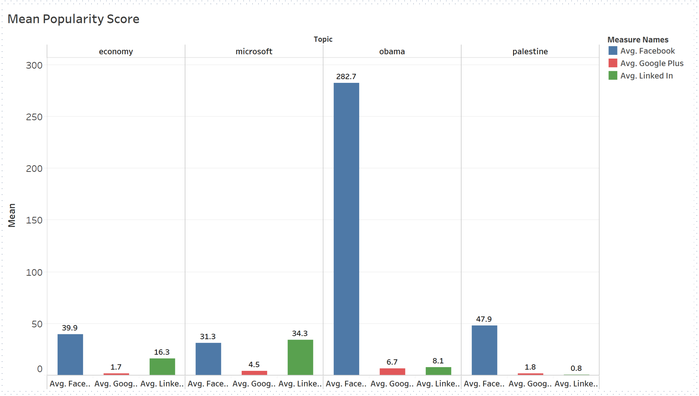
Problem & Objectives

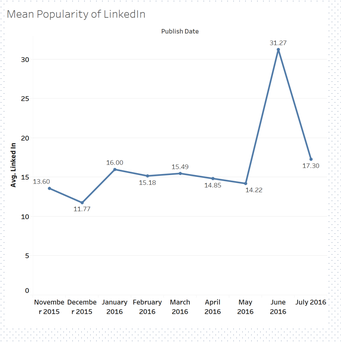
In today's world most of the people get the current information of current happenings from online news and articles. The current market of online news is large and is growing faster than expected .This leads to tough competition between online publishers and social media platforms in order to reach the largest possible audience.

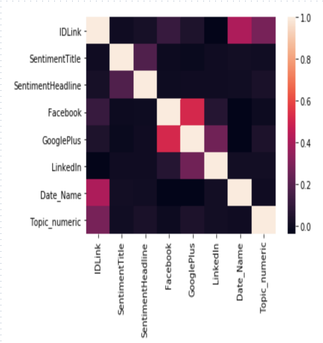
One of the important pre-requisites in this industry is to have a highly efficient aligned online strategy based on the trustworthy predictions of how popular these contents are on different platforms. Given the analysis of the current data we aim at predicting the future popularity of online news given recently published contents.

Data Description

The dataset consists of 11 features and has about 93239 records. It consists of data ranging from November 2015 to July 2016 and across various social media platforms - Google plus , Facebook and LinkedIn .It also contains the data about Sentiment Title , Topic it was published on before populating on these social media platforms along with publish date. Other columns are sparsely filled and roam around source it was published and Title of article. LinkedIn has a popularity score range of 0 to 20341



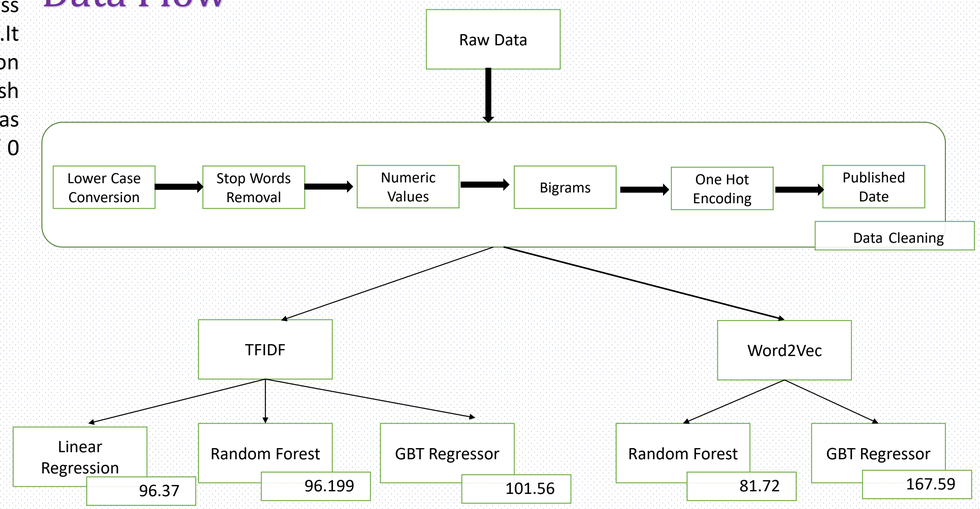




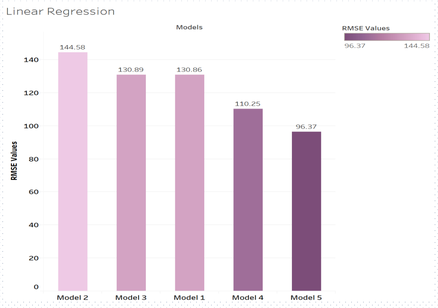
Model Description

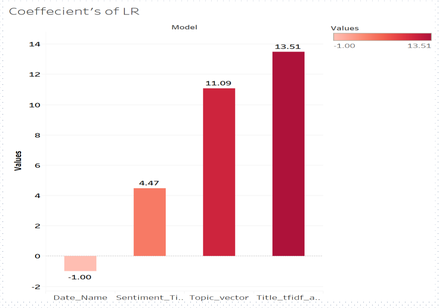
Model	Features	Techniques	Evaluation
Linear Regression	Title TFIDF ,Topic Vector, Date, Sentiment Title	Spark pipelines ,One Hot Encoding, Dummy Variables, Regularization , Cross Validation, TFIDF conversion	RMSE
Random Forest Model	Title TFIDF ,Topic Vector, Date, Sentiment Title	Spark pipelines, One Hot Encoding, Word2Vec, Vector Indexer ,Cross Validation, TFIDF conversion	RMSE
GBT Regressor	Title TFIDF ,Topic Vector, Date, Sentiment Title	Spark pipelines, One Hot Encoding, Feature Scaling, TFIDF conversion	RMSE

Data Flow



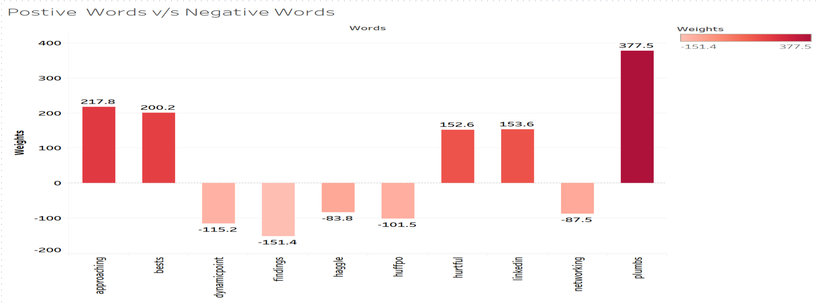
Model - Linear Regression



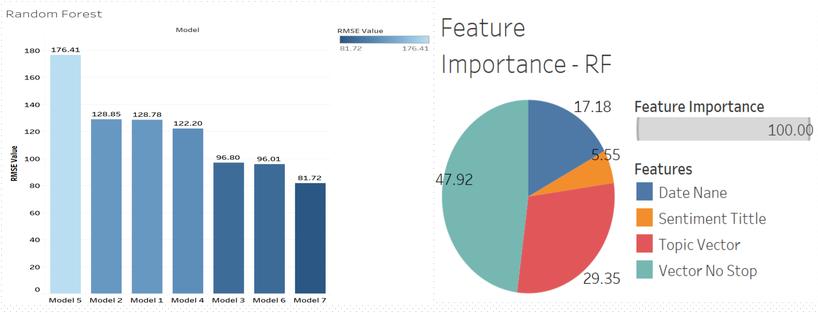


Model 1-TitleTFIDF + Topic Vector(One hot) | **Model 2** - TitleTFIDF + Topic Vector(Dummy)
Model 3- TitleTFIDF + Topic Vector(One Hot) + Publish Date | **Model 4** –Regularization (TitleTFIDF + Topic Vector(One Hot) + Publish Date)
Model 5 – Regularization(TitleTFIDF + Topic Vector(One Hot)+ Publish Date + Sentiment Title)

Pos v/s Neg words by Linear Regression

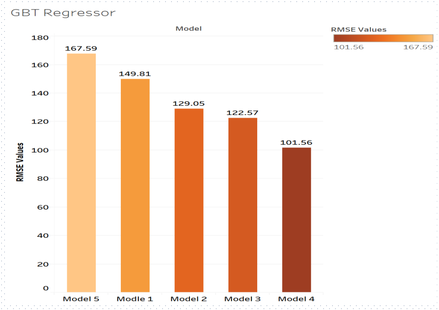


Model – Random Forest



Model 1 - Sentiment Title | **Model 2** – Sentiment Title + Sentiment Headline
Model 3- Title(TFIDF) | **Model 4** –TFIDF + Sentiment Title | **Model 5** – Title(TFIDF) + Sentiment Title + Topic + Publish Date
Model 6 – (Word2Vec) Title + Sentiment Title +Publish Date + Topic
Model 7 – Cross Validation(depth=4, bins=5) (Word2Vec) Title + Sentiment Title +Publish Date + Topic

Model – GBT Regressor



Model 1-Title(TFIDF) + Sentiment Title
Model 2 - Title (TFIDF) + Sentiment Title + Topic(One Hot)
Model 3- Title (TFIDF)
Model 4 – Title (TFIDF) +Sentiment Title + Topic (One Hot) + Publish date
Model 5 – Title(Word2Vect) +Sentiment Title + Topic (One Hot) + Publish date

Conclusion

A person could use our platform to input a Title and in return it will analysis and forecast/predict the popularity of that particular Title based on inputs with an best of RMSE of 81.72. Further the model could be improvised by providing more data i.e. data collected over longer period of time and data spread over more topics.

Reference

Data Source - UCI machine learning repository
Paper - Moniz, N., Torgo, L (2018). Multi-Source Social Feedback of Online News Feeds. Retrieved from https://archive.ics.uci.edu/ml/datasets/News+Popularity+in+Multiple+Social+Media+Platforms