PREDICTING NEWS POPULARITY ON SOCIAL MEDIA

Junyi Meng, Lexi Li, Tianci Yang, Wenwen Liu

INTRODUCTION

How popularity levels differ when the **same piece of news** published on **different social media platforms**?

Example	# shares	
❑ Fact Check: Top 10 Lies in Obama's State of the Union"	40,836	42
❑ Hospital wait times costing national economy more than $1B	5	4,328
❑ Microsoft's 'teen girl' AI turns into a Hitler-loving sex robot within 24 …	22,346	1,009

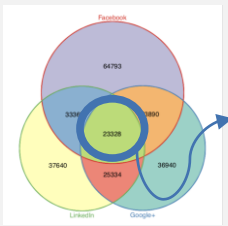
The same news piece can be widely popular on one platform but barely noticed on another one. We want to understand how popularity differs across social media platforms.

PROBLEM STATEMENT

- Which factors influence popularity across platforms?
- Can we predict the popularity based on the platform, source and titles/headlines of a given news?

DATASET

Our analysis and prediction are based on News Popularity in Multiple Social Media Platforms Data Set[1], which includes 100,000 news items of four categories and their respective social feedback on multiple platforms including Facebook, Google+ and LinkedIn.



Topic	#News
Obama	8,419
Economy	7,089
Microsoft	7,094
Palestine	736

we only look at the data of Facebook and LinkedIn given the different popularity definition of Google+ API.

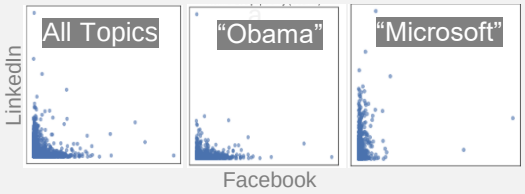
Platform	Definition of Popularity
Facebook	#shares
LinkedIn	#shares
Google+	#likes

METHOLOGY

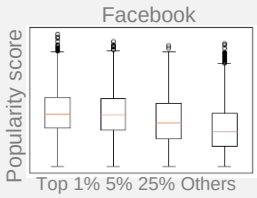
1. Exploratory Data Analysis to get usable features
2. Classify top 25% popular news respectively for Facebook and LinkedIn using (1) **concluded feature only**, (2) **text only** (3) **combined features**.

EXPLORATORY DATA ANLYSIS

Topic: Normalized News Popularity



Publisher: The more, the more popular



The dataset do not contain #followers for publishers. By assuming publishers with more news are more likely to generate popular news, sources are ranked by total number of news published, which are then used as a predictor.

Weekdays and sentiment score

Apart from topics and news sources, we also explored the **sentiment score** and **publish weekday** and add them as predictors.

PREDICITON RESULT

Target Since the distributions of news popularity are highly skewed and vary across platforms, we consider a binary classification model to predict if a piece of news is top 25% popularity.

Baseline For Lack of existing experiments on the same datasets, we set a baseline of assuming **"Everything from the Top 1% news sources will be a top 25% popularity".**

Model For interpretation purpose, association between features and news popularity are examined mainly by logistic regression (LR).

(1) Feature Based Prediction Four features: (1) news topic (Categorical) (2) source rank (Categorical) (3) sentiment score (numerical) (4) publish weekday (Categorical).

Interpretation We found news from **Top5% Sources**, with **Obama** topic, published in **weekdays** are more likely to be the TOP25% popular news on Facebook.

The news of Economy topic and Microsoft topic are more likely to be the TOP25% popular news on LinkedIn.

Input	Method	F		L	
		Acc.	F1	Acc.	F1
Baseline	/	62.5	31.3	65.4	36.9
Feature	LR	76.5	50.4	**74.9**	**45.4**
TF-IDF	LR	76.0	50.1	73.3	29.3
TF-IDF	LSTM	77.5	**56.4**	73.9	27.2
GloVe	LR	71.0	53.5	70.0	26.2
GloVe	LSTM	**78.5**	52.7	73.9	31.4
Combine	LR	**79.4**	53.5	**75.9**	**48.9**

Result of Feature/Text-based and Combined models

(2) Text Based Prediction Can text alone predict news popularity? We also try to incorporate the information from the text.

For text representation, we use the TF-IDF approach and Global Vectors (GloVe). Logistic regression and Long short-term memory (LSTM) network are implemented for popularity prediction. Not surprisingly, LSTM performs better since it considers a contextual representation of each token vector with respect to the entire news content.

Interpretation We mainly use TF-IDF features for model interpretation. Since each word is mostly 0/1 variable, using Bag of Words + TF-IDF and logistic regression, we can simply consider coefficients as an approximate word importance. We can observe using political words increases the probability of being popular on Facebook.

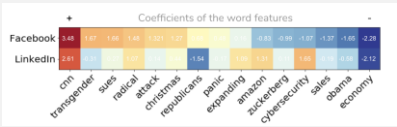
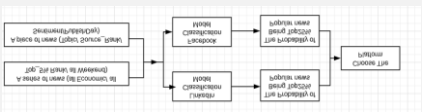


Figure: How the word usage influences the probability of being popular. Red means using the word increase the probability of being popular; blue means a decreasing.

(3) Combine features and word embeddings Finally we concatenate the numerical and lexical features and use both as our predictors in the mode. We find the classification accuracy has improved in both platforms compared with only text classification or feature-based classification.

CONCLUSION

In this project, we not only build a classification model to predict news popularity, but also provide the news publishers with an analytical tool to help them analyze how their news can get popular by choosing platform, topic, sentiment, word usage, and publish day. The implication of our classification model might work as follows:



Reference: [1] Moniz, Nuno, and Luís Torgo. "Multi-source social feedback of online news feeds." arXiv preprint arXiv:1801.07055 (2018).