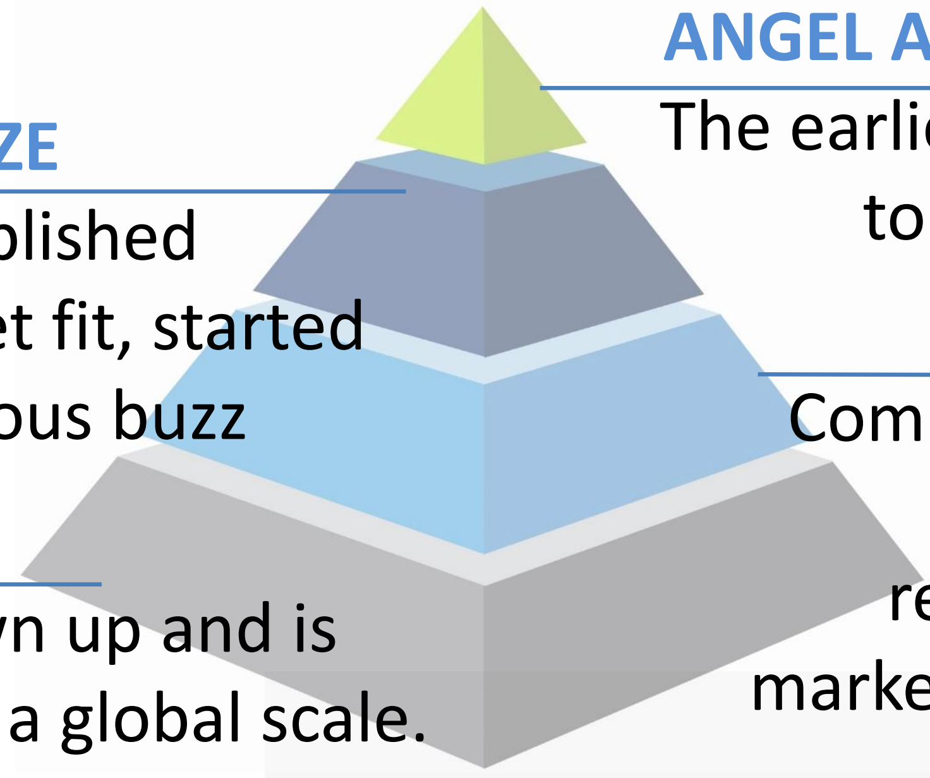
Micro-Ventures :
Predicting Potential Startups for Micro-Investments

Immad Imtiaz, Ravi Bisla, Shariful Islam

Motivation

- Venture Capital is the money provided by investors to startups
- Large VCs have their own data analysis team to make informed decision about investments
- Small investors lack comprehensive information to make informed decision before investment
- We try to approach this problem using **Machine Learning (ML)**



ANGEL AND SEED FUNDING
The earliest stage of funding to get the party going

SERIES A - OPTIMIZE
Company has established product and market fit, started to make some serious buzz

SERIES B - Build
Company has started to make considerable revenues in selected markets and is looking to expand operations

SERIES C - Scale
Company has grown up and is likely operating on a global scale. Ready for IPO or acquisition.

What Is Our Contribution

- Collect and unify data from multiple sources
- Use ML to **predict which start-up will reach series-C**
- Use online articles about companies for topic modeling
- Find important features and topics related to company success

Our Data

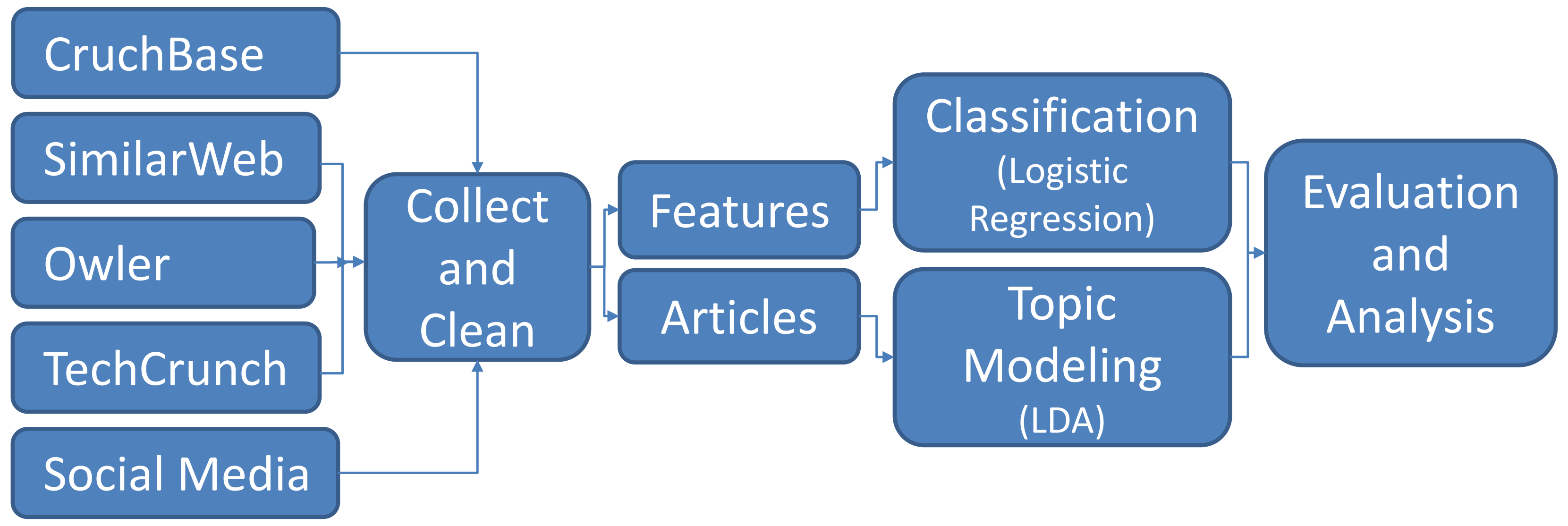
We collect data of **44927 companies** from the following sources -

- Crunchbase - https://www.crunchbase.com/
- Similarweb - https://www.similarweb.com/
- Owler - https://www.owler.com/
- Online articles from techcrunch (58,157 articles)

Machine Learning and Evaluation

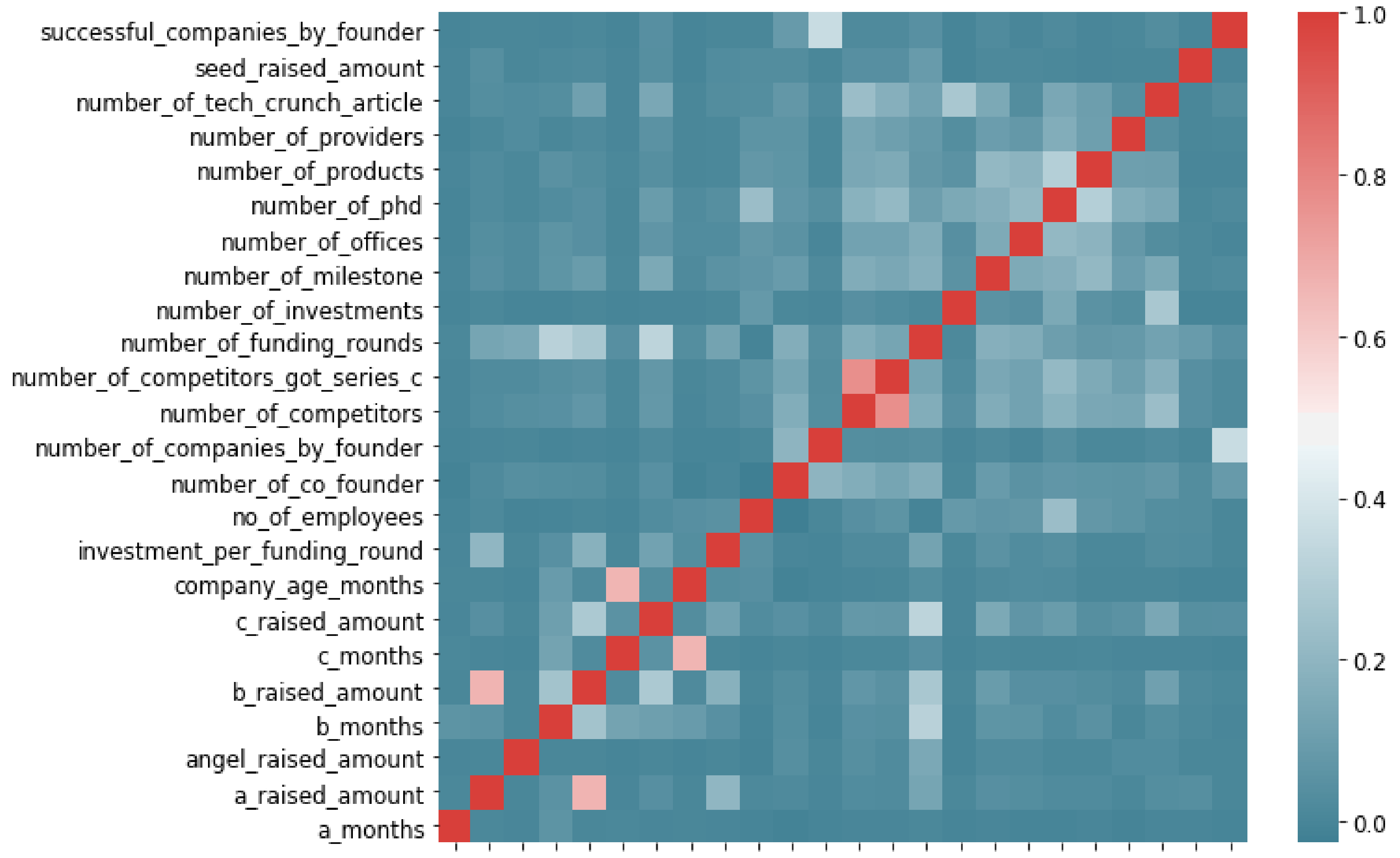
- After data collection, cleaning and integration, we generate the features for ML model
- We use LDA to find out the most relevant topics in the articles
- Logistic Regression is used for classification
- We split the data into train and test set and also use 10-fold cross validation to report the findings
- We observe TPR, FPR and area under ROC curve (Investors will care more about TP and FP compared to accuracy)
- We train the model for different categories of business
- We train ML model with and without the learned topics as features and observe their effect on prediction

Data Pipeline



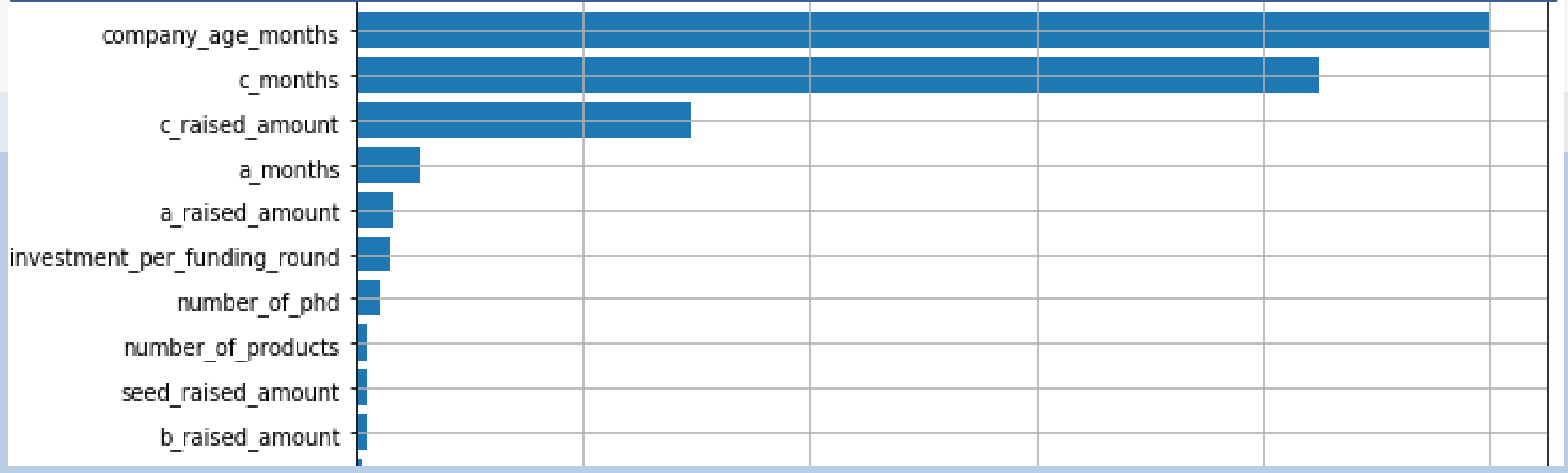
CruchBase
SimilarWeb
Owler
TechCrunch
Social Media
Collect and Clean
Features
Articles
Classification (Logistic Regression)
Topic Modeling (LDA)
Evaluation and Analysis

Features and Topic Visualization



Important Features

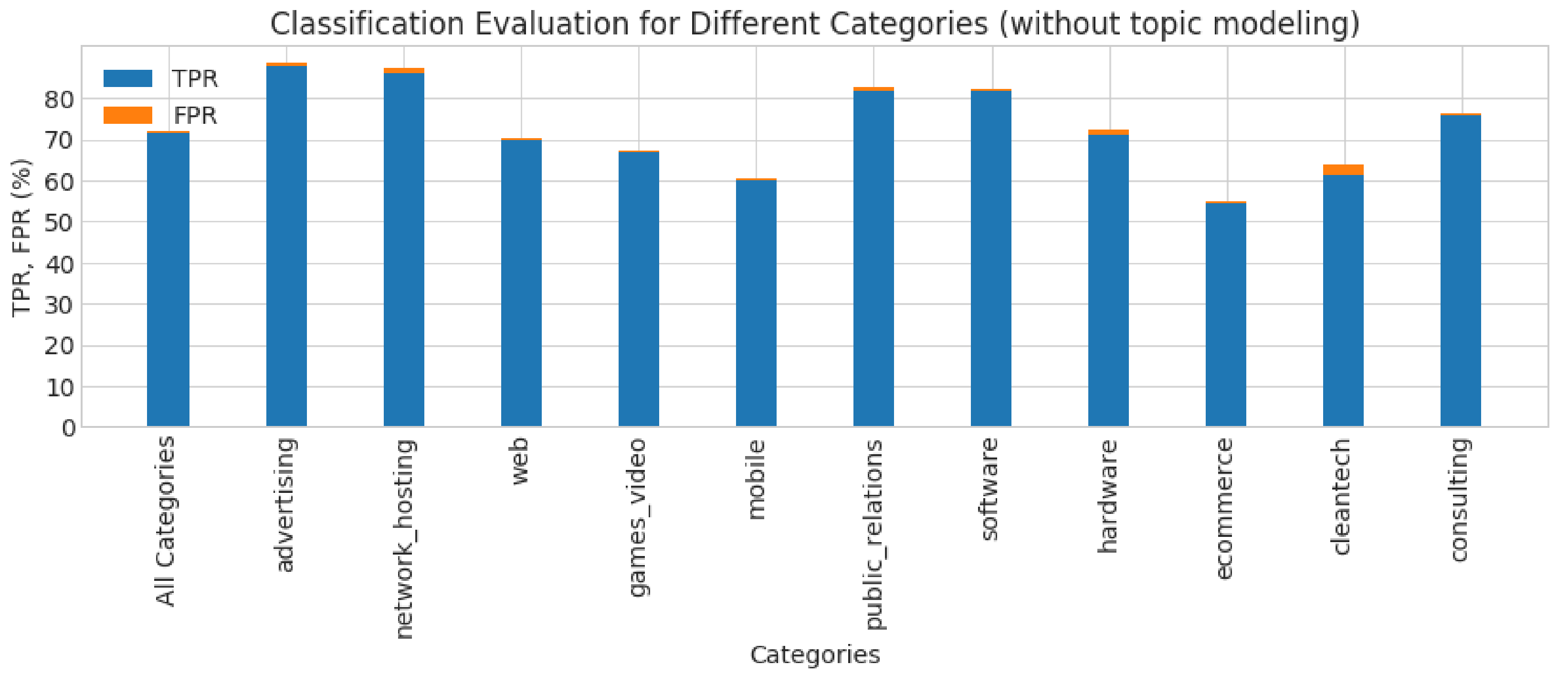
- For all the categories most important features are related to the amount of **raised funding** and **company age**
- Only for the category ‘public_relation’ the most important feature is number of employees

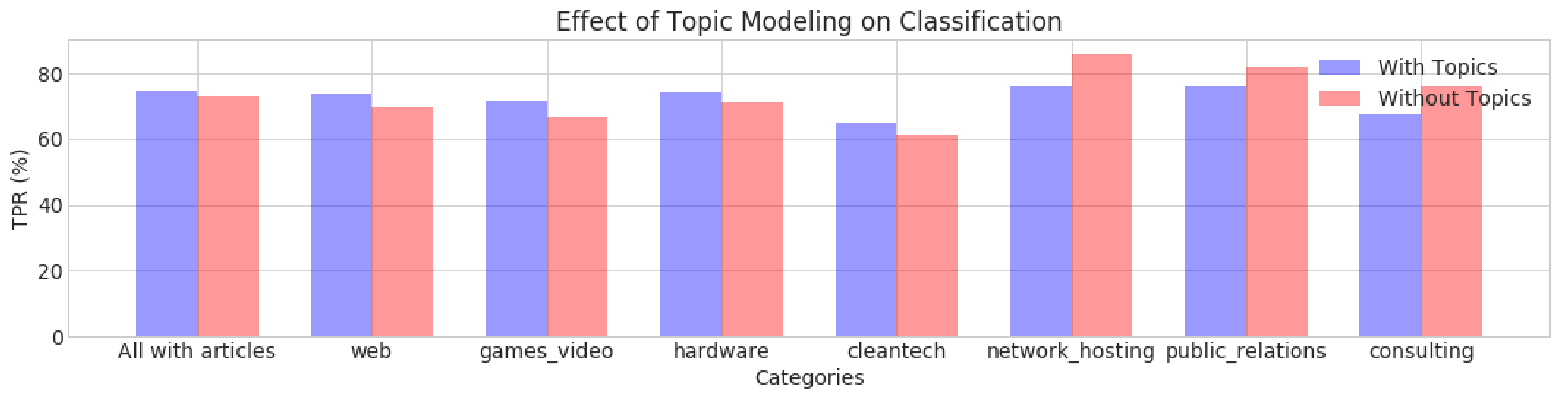


Topic Modeling

Topics	Words
Social Media	User, Facebook, Social, Twitter, Friend, Photo
Business	Business, Start-up, Technology, Platform, Founder
Mobile Device	App, Mobile, Apple, Android, Nokia, Kindle, iPad
Stocks	Million, Billion, Revenue, Share, Stock, IPO
Advertising	Google, Ad, Search, Advertising, Web, Yahoo
Funding	Venture, Investor, Round, Capital, CEO, Raised

Results





Findings

- True positive rates are between 60% to 80%
- False positive rates are very low (in most cases < 1%)
- Area under ROC curves are greater than 0.8
- For most of the technology companies Topic Modeling enhances the performance of the classifier (2% to 8%)
- For categories like ‘public_relation’ and ‘consulting’ the classifier works worse with features from topic modeling

Learning and Future Plans

- Data collection, cleaning and integration is a tedious process
- Collected real life data can be very sparse
- Future work would be to explore more about how to handle data sparsity efficiently (find more source, how to fill them up)
- To predict **when** the star-up will reach series-C
- And to build an interactive web interface for the user