TRIBE DYNAMICS

Mid-term presentation

IACS Capstone Project Course

DataShack 2018

POLITECNICO MILANO 1863

IACS INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE AT HARVARD UNIVERSITY

Team

Who are we?

- Data Scientist
- Computer Engineer
- Communication Designer

Andrea Porelli

Srivatsan Srinivasan

Harvard University

Alessandro Bianchi

Ginevra Terenghi

Politecnico di Milano

Outline

1. Who is Tribe Dynamics
2. Problem Statement
3. Motivation
4. EDA - Exploratory Data Analysis
5. Models
6. Future Implementation

Who is Tribe Dynamics?

Consulting San Francisco-based startup

Measure social media engagement

for beauty, fashion and lifestyle brands

Around 50 people in the company

mostly non-specialized

4horsemedia.com/product/professional-plan-social-media/

What do they do?

Clients want to know how much people are talking about them on social media, versus their competitors

Tribe provides a metric to value the companies' social media campaigns - EMV [Weighted average of likes, comments and shares of posts]

Two subtasks to compute it:

BINOMIAL TASK: "Is this post about Brand X or not?

MULTINOMIAL TASK: "What is this post about?" (Product Categories)

Tribe Dynamics

Who are their clients?

Leading brands in
beauty, fashion and lifestyle industries

Mostly related to cosmetics

Clients want to get a general understanding of digital trends within the beauty industry

today.mims.com/malaysia-s-moh-tightening-governance-over-online-cosmetics-businesses

Problem Statement

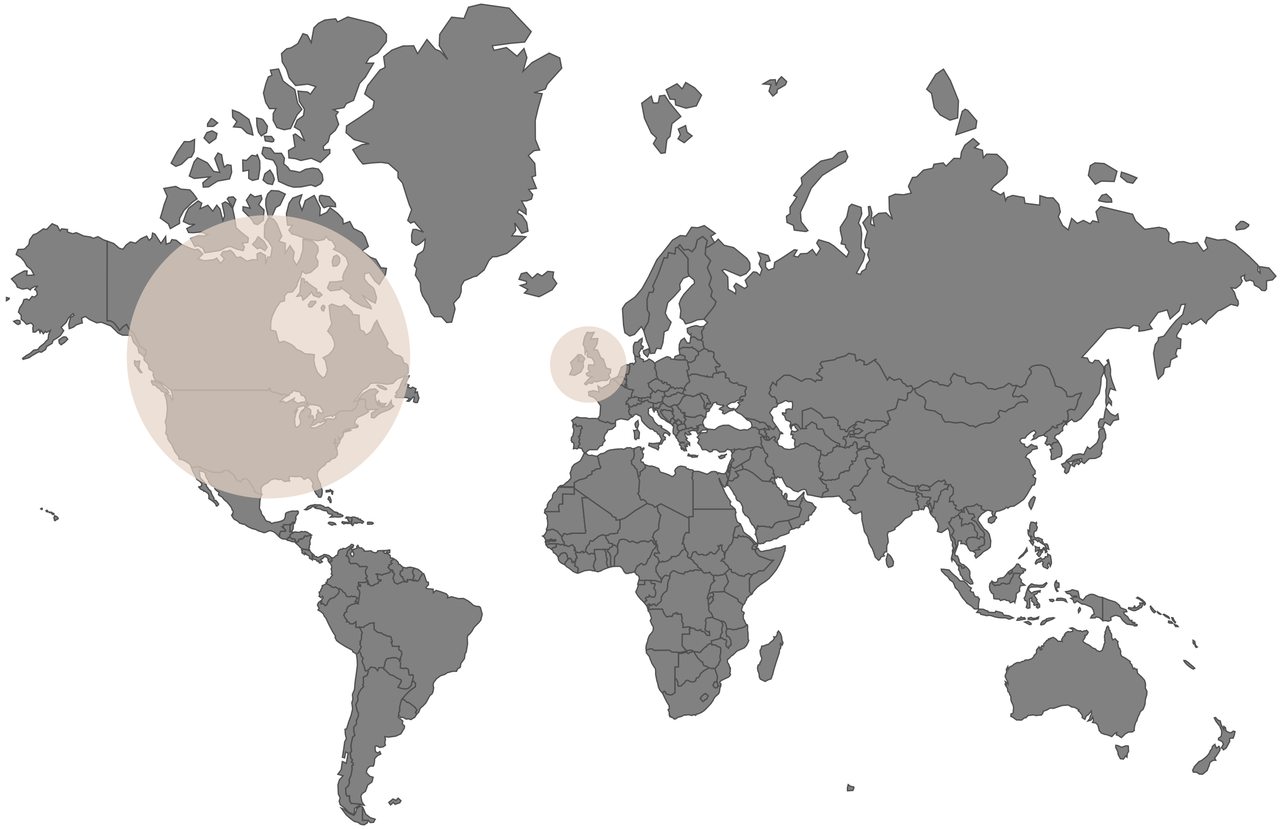
Where Tribe currently works

U.S. // English market

Where Tribe would like to expand its analysis

European market

Asian market



Problem Statement

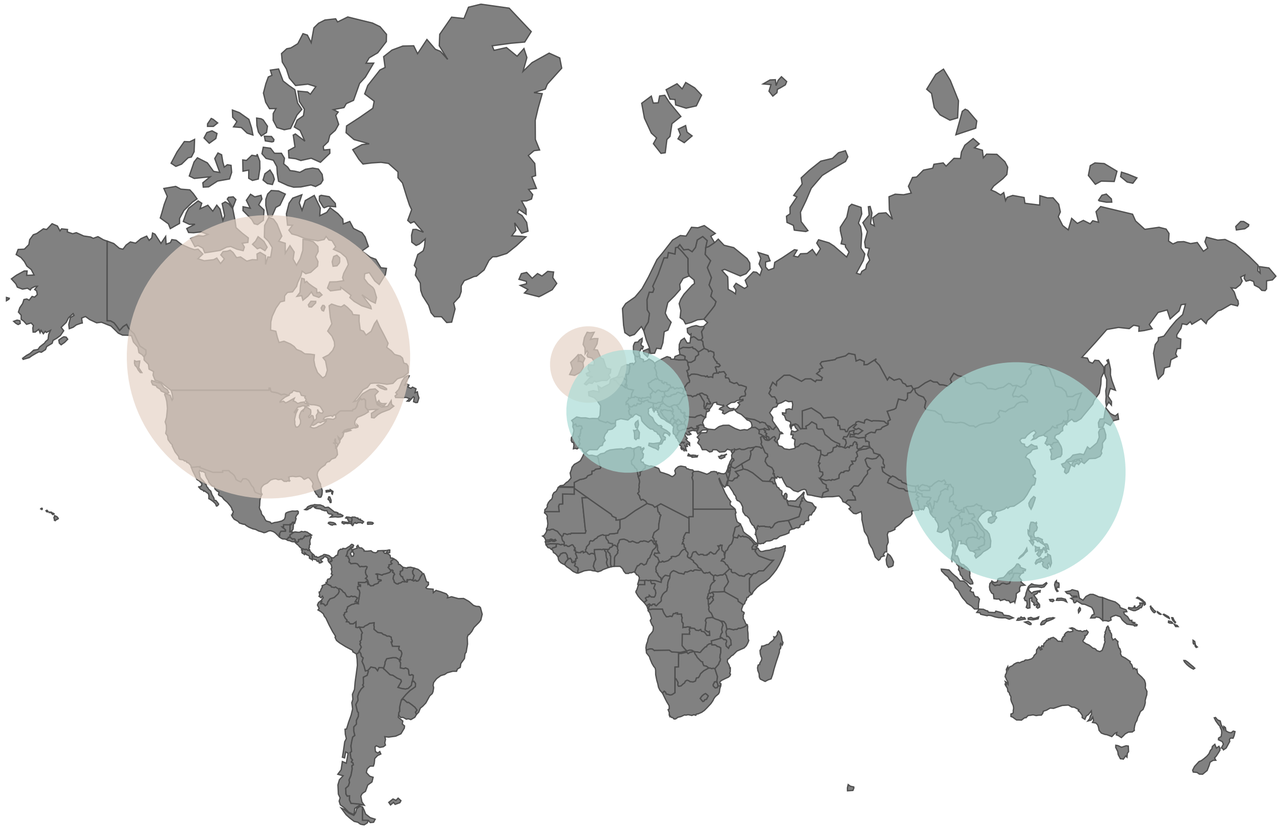
Where Tribe currently work

U.S. // English market

Where Tribe would like to expand its analysis

European market

Asian market



Problem

How can they develop a classification algorithm that scales well across different languages, while maintaining the same performance of the English model?

Goal

Provide a model that scales well across languages with limited amount of training data, by exploring the structural similarity across languages

Project

How do we plan on doing it?

Setting some baselines

- Linear models
- Pre-trained word-embeddings

Evaluating different tradeoffs

- Machine translation [costs vs. benefits]
- Model complexity [accuracy vs. scalability]

Moving on to more advanced approaches

- Cross-lingual embeddings
- Transfer learning

Project

How do we plan on doing it?

Setting some baselines

- Linear models
- Pre-trained word-embeddings

Evaluating different tradeoffs

- Machine translation [costs vs. benefits]
- Model complexity [accuracy vs. scalability]

Moving on to more advanced approaches

- Cross-lingual embeddings
- Transfer learning

How do we plan on doing it?

Setting some baselines

- Linear models
- Pre-trained word-embeddings

Evaluating different tradeoffs

- Machine translation [costs vs. benefits]
- Model complexity [accuracy vs. scalability]

Moving on to more advanced approaches

- Cross-lingual embeddings
- Transfer learning

What would we like to find?

Why Tribe Dynamics cares about this problem?

They are rapidly expanding in Europe and Asia

They want to increase accuracy on different languages other than English

What is the challenge for our team?

Explore social media contents across different linguistic structures

What we would like to find?

Higher performance of our models for contents written in non-English languages

texasgopvote.com/economy

EDA - Exploratory Data Analysis

What data do we have?

Dataset:

106 Cosmetic Brands - mostly based on European market

How does it look like:

	text	lang	model_decision	labels	mturker	link
0	Favorites Summer Crushes - $30This is another ...	en	None	[1, 1, 1, 1]	1	http://www.thebeautyisle.com/2015/01/sephora-u...
1	.@CaudalieUSA's new C15 line takes charge of o...	en	None	[1, 1, 1, 1]	1	https://www.twitter.com/Sephora/statuses/46565...
2	I love a simple, cheap and easy to find produc...	en	None	[1, 1]	1	http://www.la-fille-en-rose.com/2012/10/dry-sk...
3	Yay it's @CaudalieUSA turn #sephora #weareseph...	en	None	[1, 1, 1, 1]	1	https://www.twitter.com/CaitlinAPearl/statuses...
4	@LadyAndTheLook I started using a face oil fro...	en	None	[1, 1, 1, 1]	1	https://www.twitter.com/stylecontext/statuses/...

Post Text

from different social networks or websites

Language

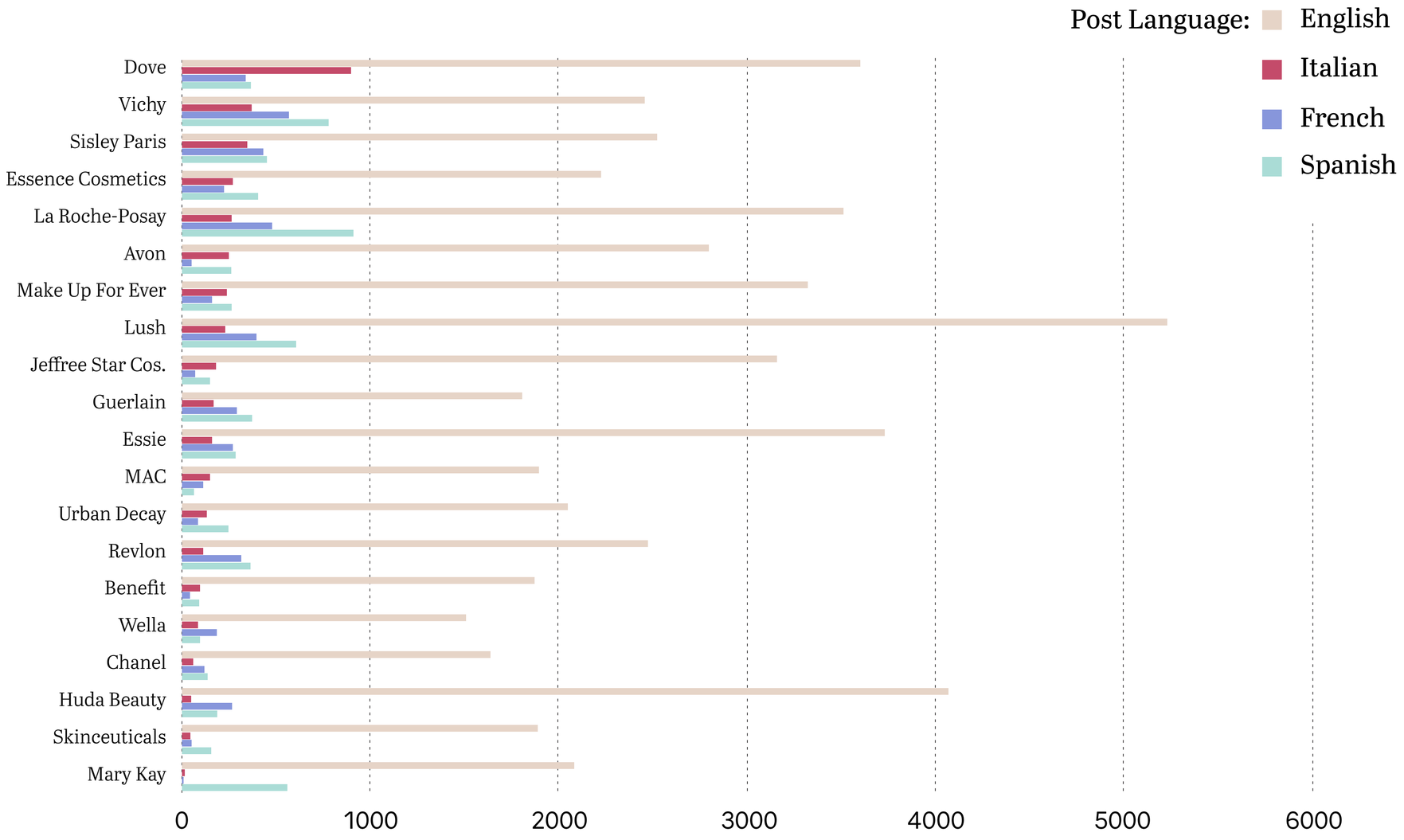
Human Labels

Labeller

Post link

Posts by language

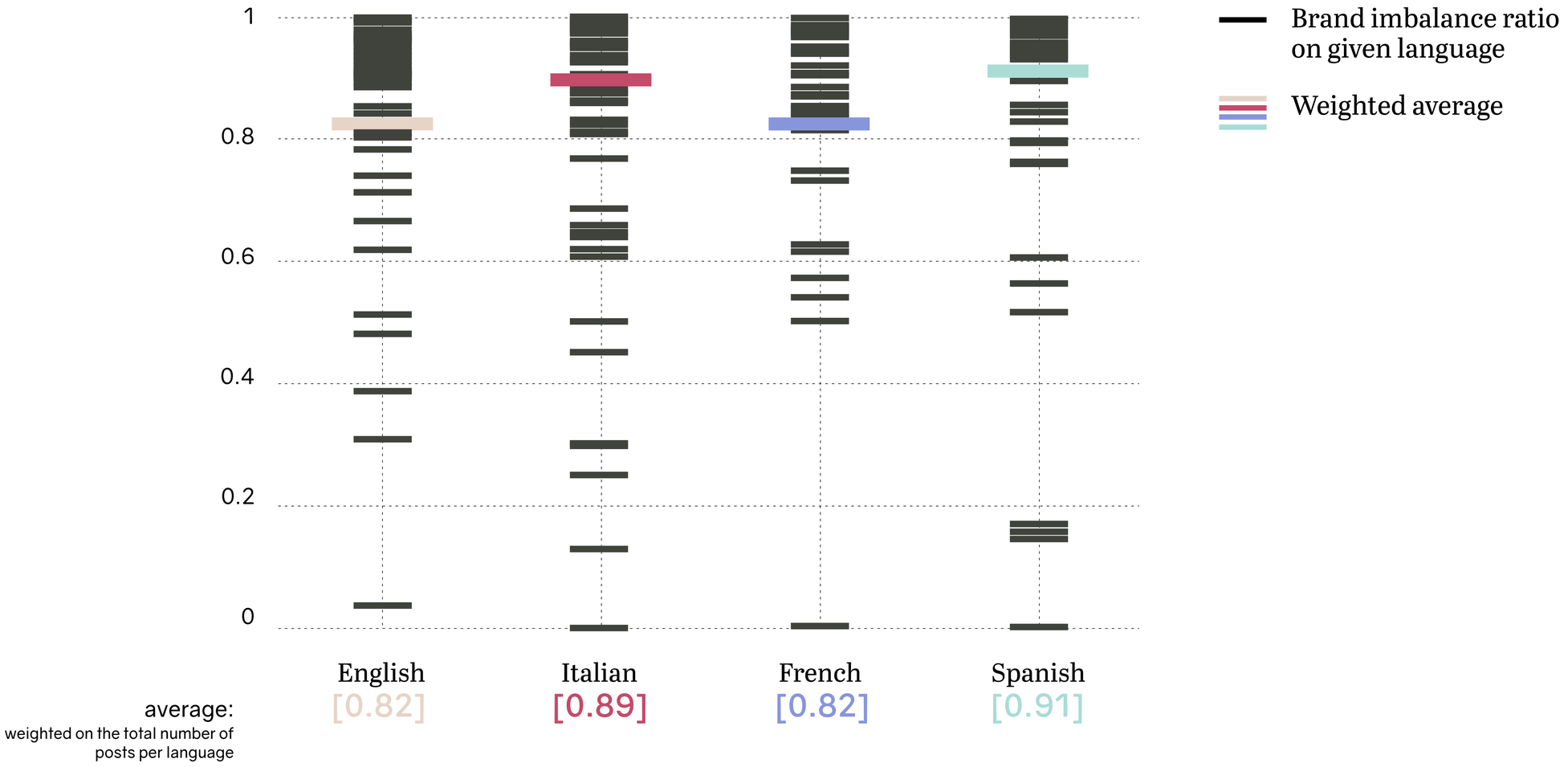
First 20 brands per total number of posts



Class imbalance by language

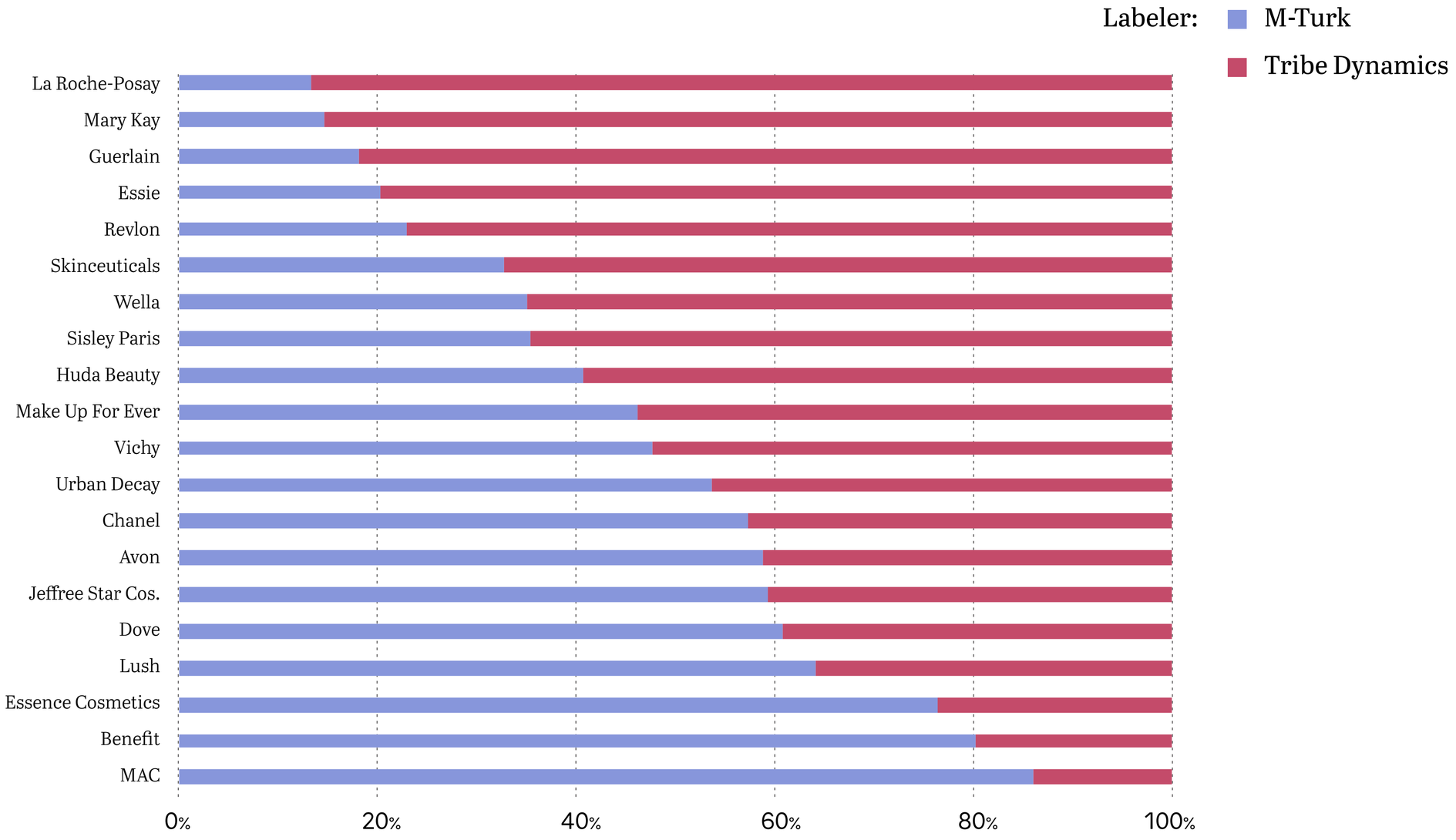
Imbalance ratio of True // False labels for each brand

$$\text{Imbalance ratio} = 1 - \frac{\text{number of instances of the } \textbf{minority class}}{\text{number of instances of the } \textbf{majority class}}$$



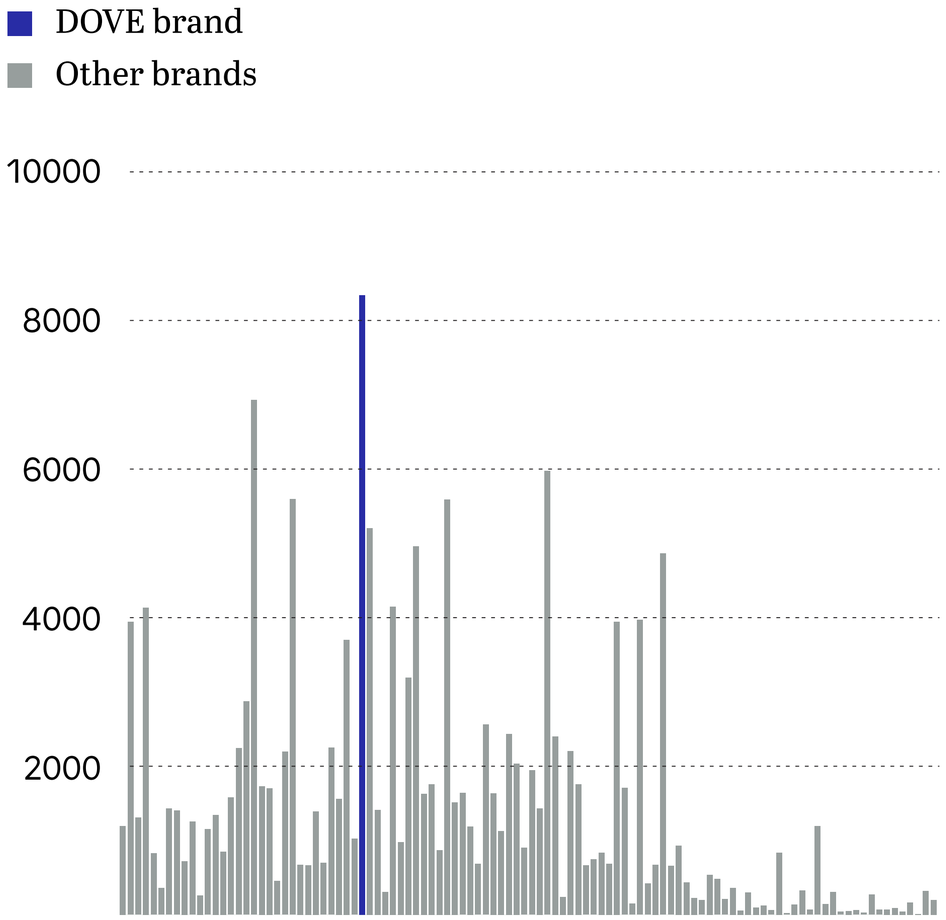
Labelers

Post labeler comparison on the first 20 brands per total number of posts



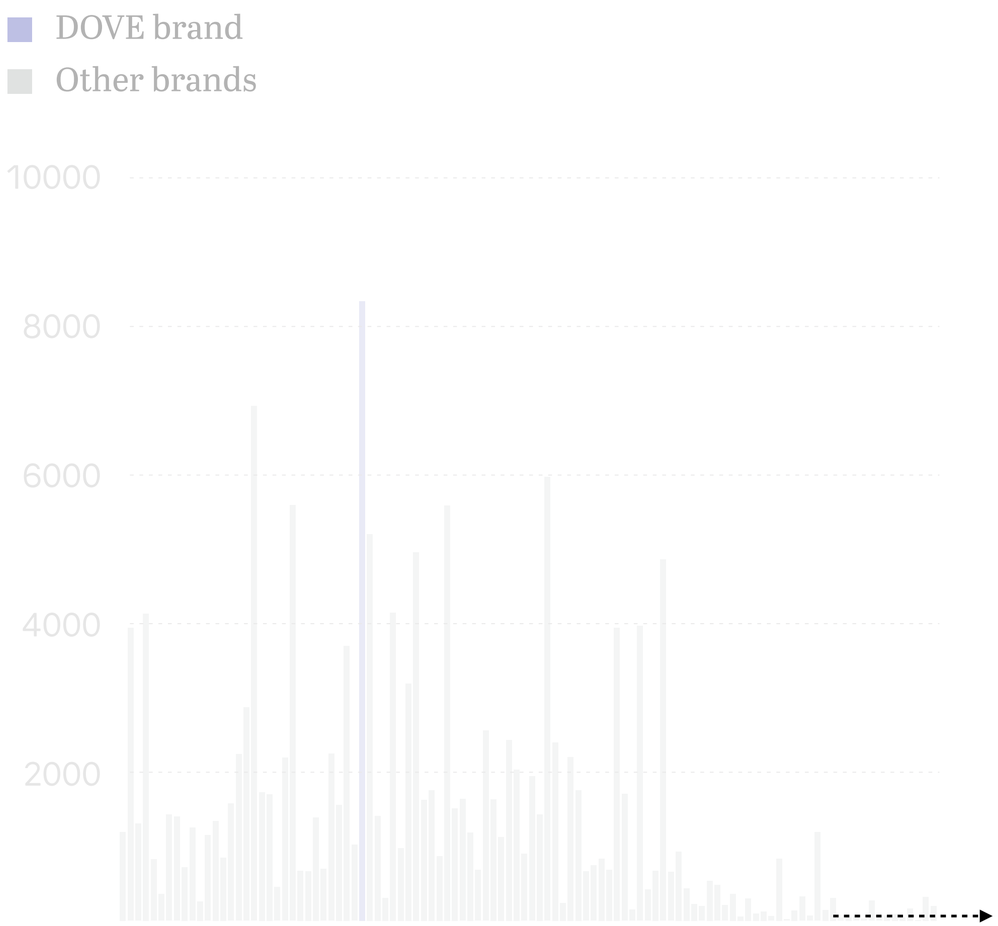
Why did we choose “DOVE” dataset?

1. Dataset with highest number of posts

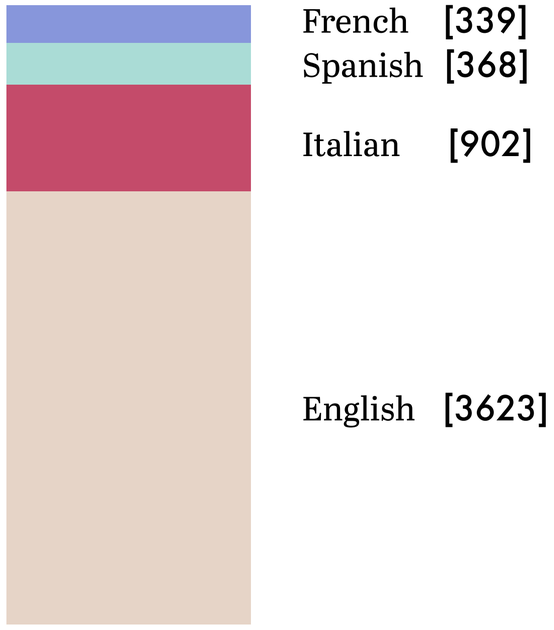


Why did we choose "DOVE" dataset?

1. Dataset with highest number of posts

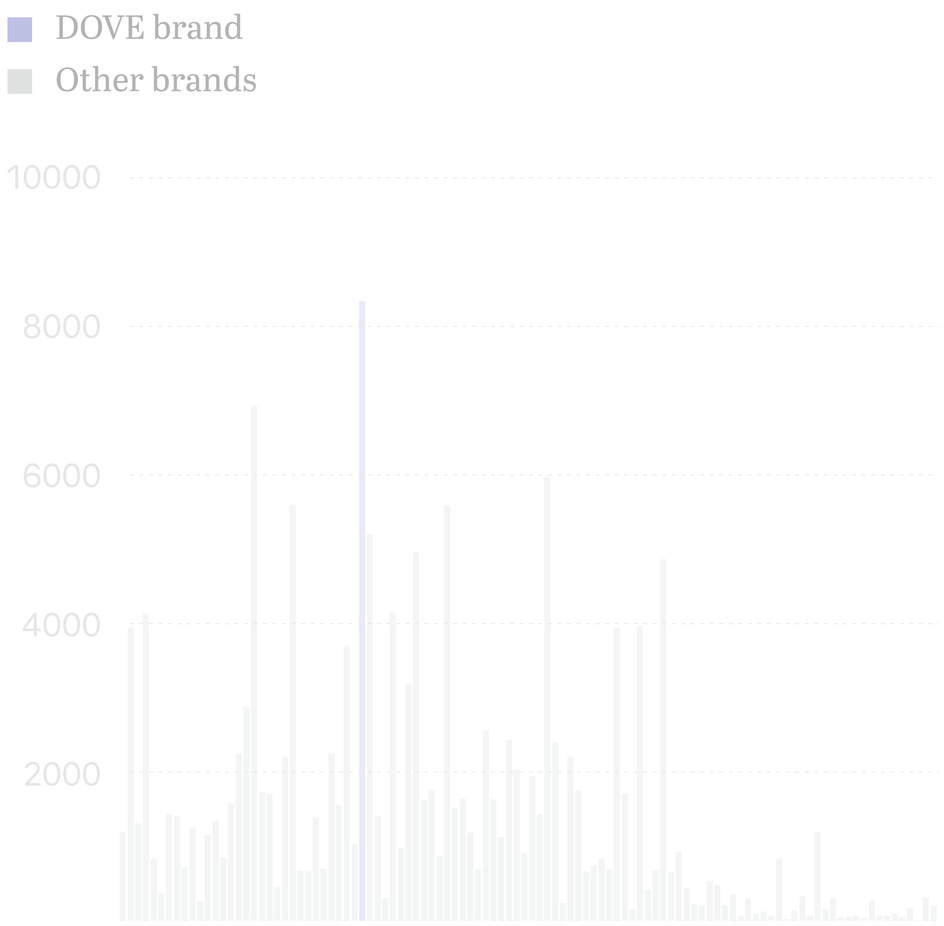


2. Dataset with enough posts per language

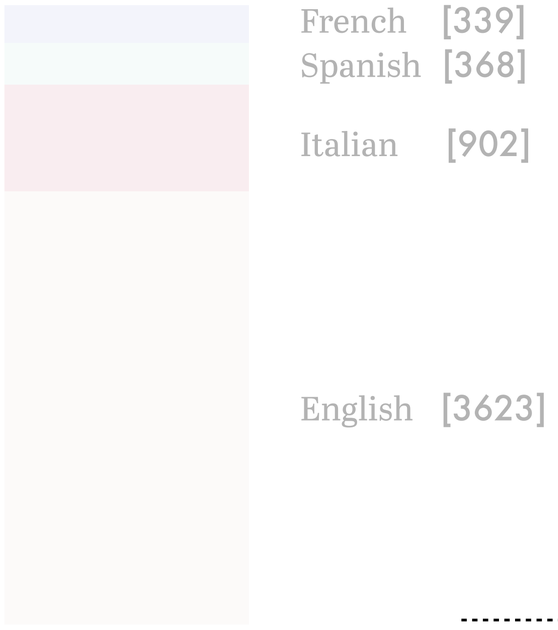


Why did we choose "DOVE" dataset?

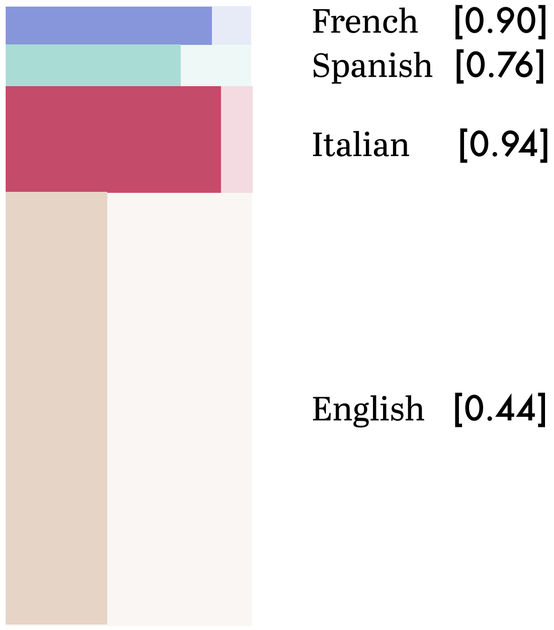
1. Dataset with highest number of posts



2. Dataset with enough posts per language



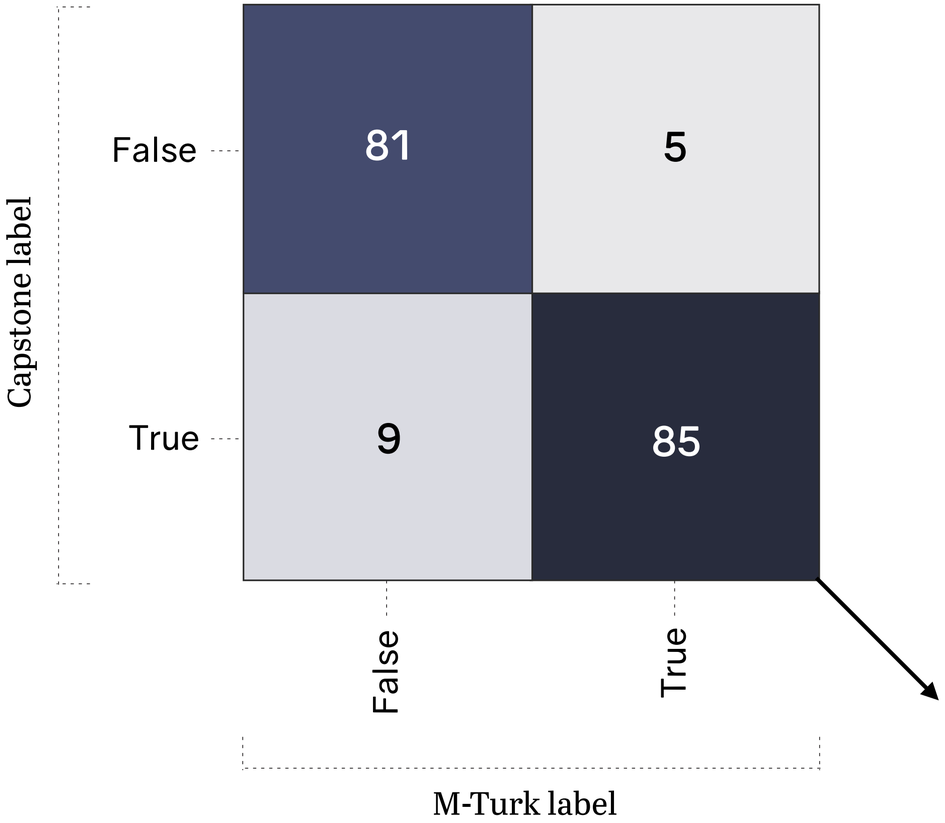
3. Dataset with acceptable imbalance



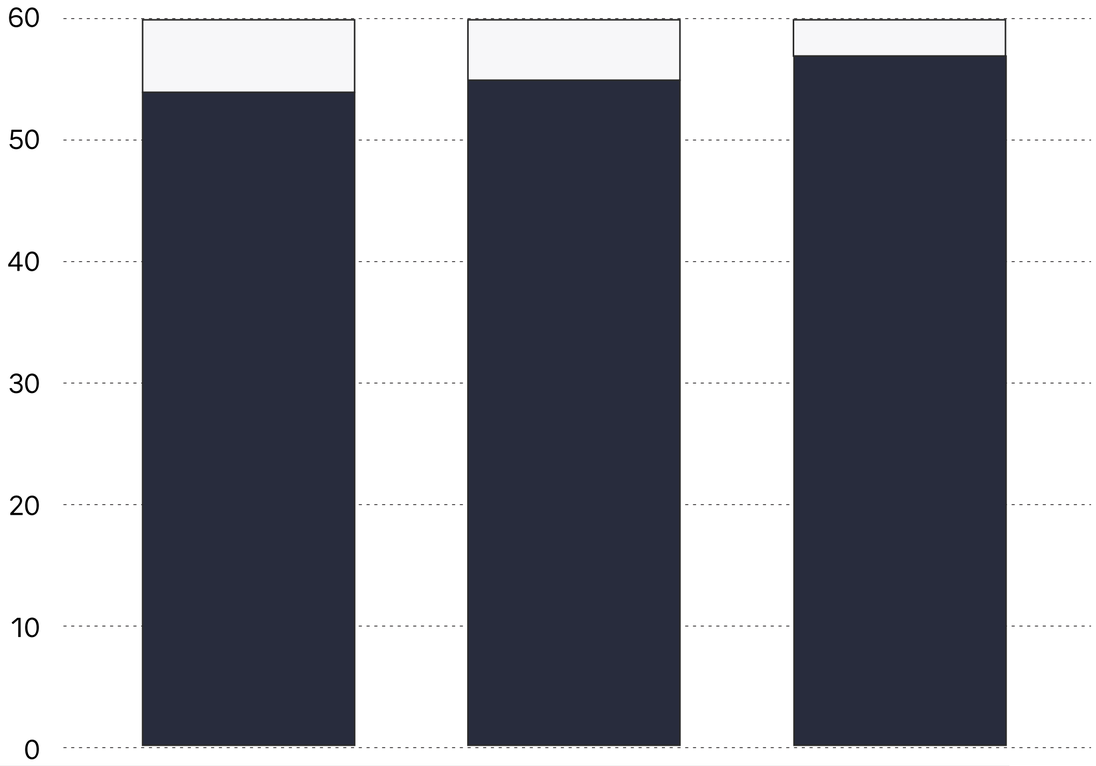
From DOVE dataset

Ground True reliability

Labeling:



Labelling accuracy per language



TOTAL	Italian	French	Spanish
[92.2%]	[90%]	[92%]	[95%]

Which is Tribe's current approach?

Evaluated on "**Kate Tokyo**" **dataset**
(only one with Tribe's model predictions)

N-grams Logistic Regression
on a multilingual-vocabulary

ENGLISH

Number of observations: **2610**
Imbalance ratio: **0.87** [True=**299** ; False=**2311**]

Performance:
Accuracy = **0.92**
ROC curve AUC = **0.70**
Average precision score = **0.38**

ENGLISH + ITALIAN

Number of observations: **2746**
Imbalance ratio: **0.88** [True=**305** ; False=**2441**]

Performance:
Accuracy = **0.92**
ROC curve AUC = **0.70**
Average precision score = **0.38**

NON - ENGLISH

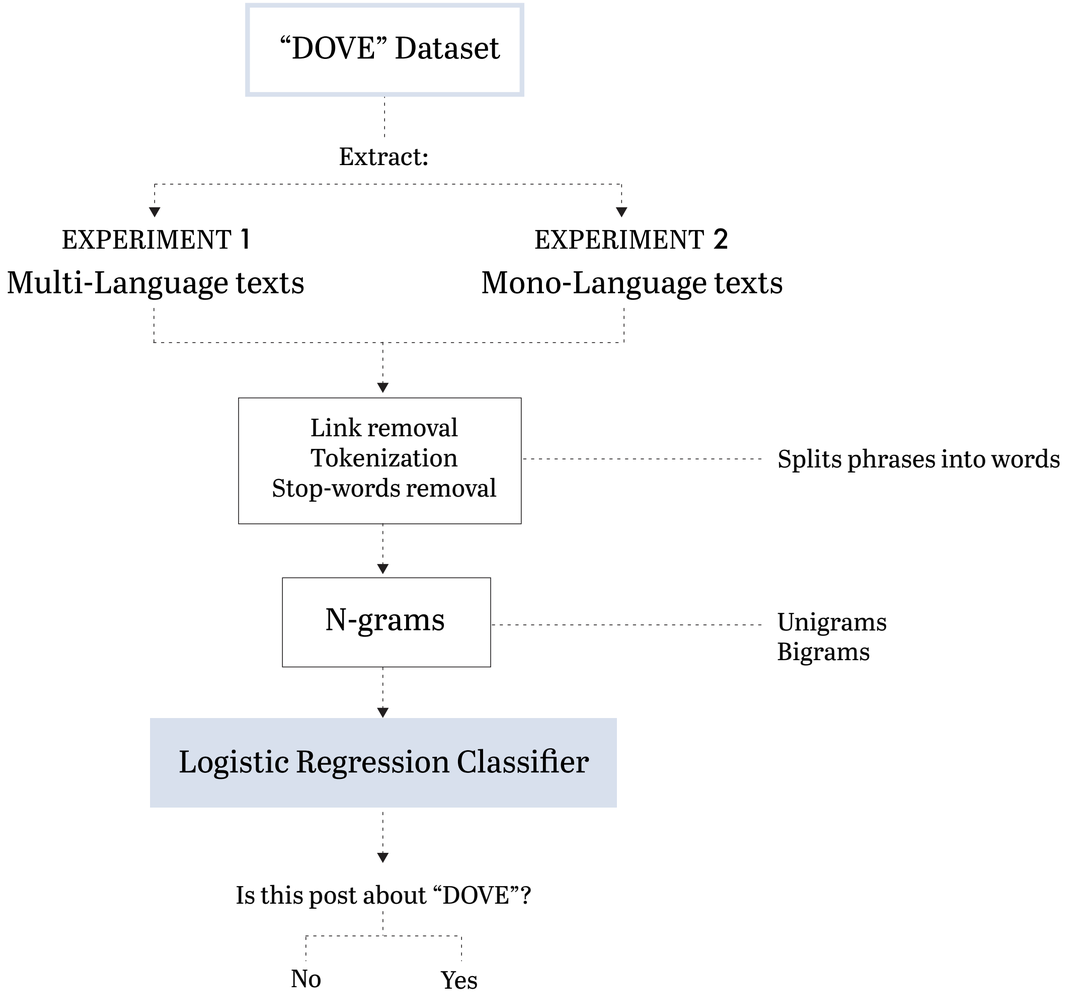
Number of observations: **2087**
Imbalance ratio: **0.50** [True=**699** ; False=**1388**]

Performance:
Accuracy = **0.84**
ROC curve AUC = **0.84**
Average precision score = **0.67**

Models - First Approach

N-grams Logistic Regression

- Starting Dataset
- Pre-Processing
- ML Model



N-grams Logistic Regression

Most relevant features for the binomial classification task

Ranking based on the coefficients of the Logistic Regression model

- Feature
- Product
- Target
- Market

Voyant-tools

N-grams Logistic Regression

Performance

Model: LogisticRegression()

N-grams: (1,2)

Preprocessing: Removed links + Encoded hashtags + Encoded mentions + Removed stopwords

Cross-Validation: Stratified 10-Fold CV

Evaluated on "**DOVE" dataset** (lacking Tribe's model predictions)

ENGLISH

Number of observations: **3623**
Imbalance ratio: **0.45** [True=**2336** ; False=**1287**]

Performance:

Accuracy = **0.91** +/- **0.01**

ROC curve AUC = **0.96** +/- **0.01**

Average precision score = **0.98** +/- **0.01**

ENGLISH + ITALIAN

Number of observations: **4525**
Imbalance ratio: **0.10** [True=**2383** ; False=**2142**]

Performance:

Accuracy = **0.92** +/- **0.02**

ROC curve AUC = **0.97** +/- **0.01**

Average precision score = **0.97** +/- **0.01**

ITALIAN

Number of observations: **902**
Imbalance ratio: **0.95** [True=**47** ; False=**855**]

Performance:

Accuracy = **0.96** +/- **0.01**

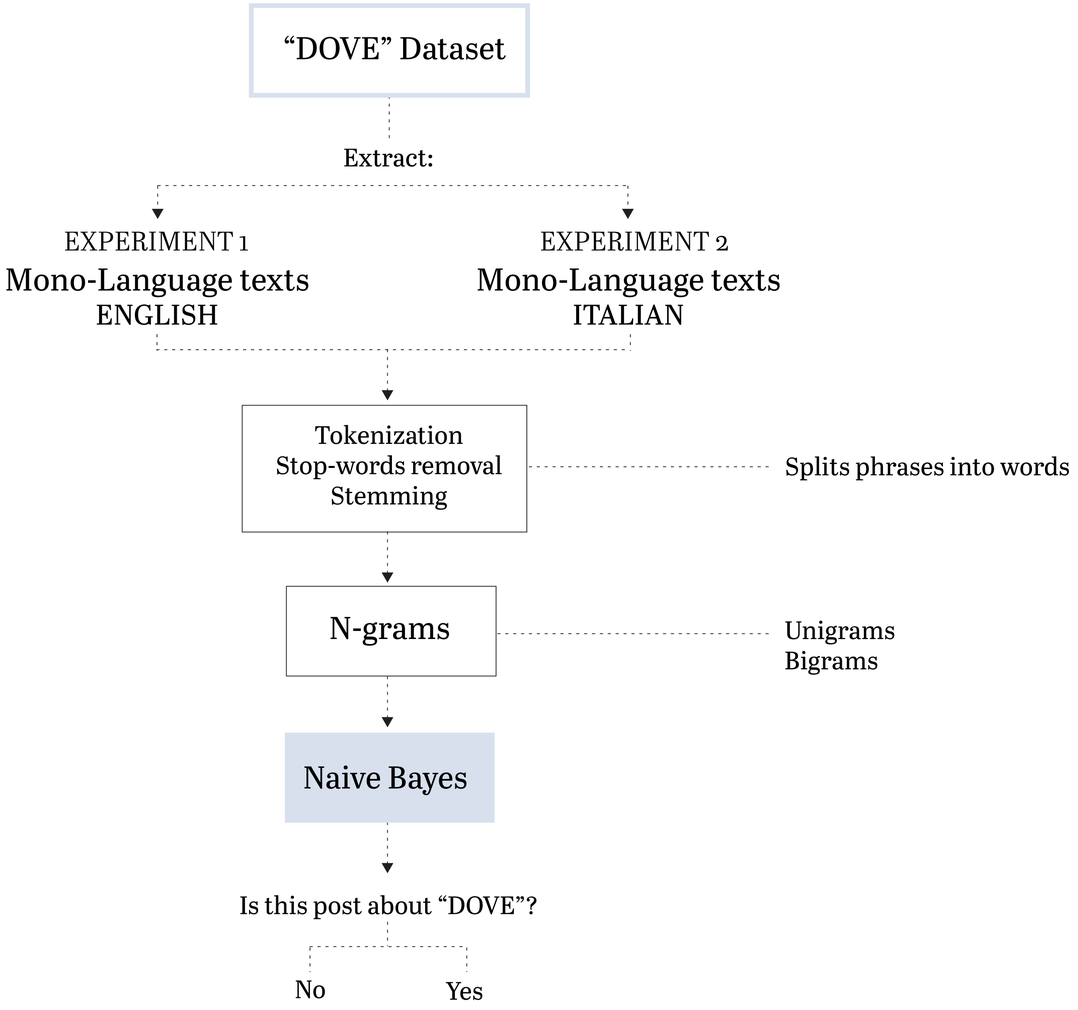
ROC curve AUC = **0.82** +/- **0.13**

Average precision score = **0.62** +/- **0.25**

Models - Second Approach

N-grams Naive Bayes

- Starting Dataset
- Pre-Processing
- ML Model



N-grams Naive Bayes

Most relevant features for the monolingual classification task

Ranking based on the empirical log probability of features given a class (P(x_i|y))

- Feature
- Product
- Target
- Market

English texts

Italian texts

N-grams Naive Bayes

Performance

Model: Multinomial NB(fit_prior=False)

N-grams: (1,2)

Preprocessing: SnowballStemmer (ignore_stopwords=False) + TfidfTransformer()

Train-Test split: 60-40

Evaluated on "**DOVE**" **dataset** (lacking Tribe's model predictions)

ENGLISH

Number of observations: **3623**
Imbalance ratio: **0.45** [True=**1385** ; False=**788**]

Performance:

Accuracy = **0.85**

ROC curve AUC = **0.95**

Average precision score = **0.97**

ITALIAN

Number of observations: **902**
Imbalance ratio: **0.95** [True=**30** ; False=**511**]

Performance:

Accuracy = **0.96**

ROC curve AUC = **0.76**

Average precision score = **0.35**

Observations

What can we take away from these baseline models?

1. Really good performance even with simple models
2. Class imbalance is a relevant issue
3. Text pre-processing is key (hashtags and mentions give performance boost)

Observations

Why do we need more complex models?

Solutions:

1. N-grams are cumbersome and not scalable
   Vertical classification might not be easy just with hashtags

   Compact representation of words with context
   Understanding sentence context might be key

2. One model per language has performance variance across languages
   [Might not have enough data to train one model per language]

   Train a model and transfer model features to another language space
   Latent generative models can work with unlabeled corpus of target data (saves MTurk cost)

3. Interesting academic exercise

Future Implementation

Idea 1: **Word Embeddings**

Across the languages there are words with similar intent

Model learning

Clustering words with similar intent together

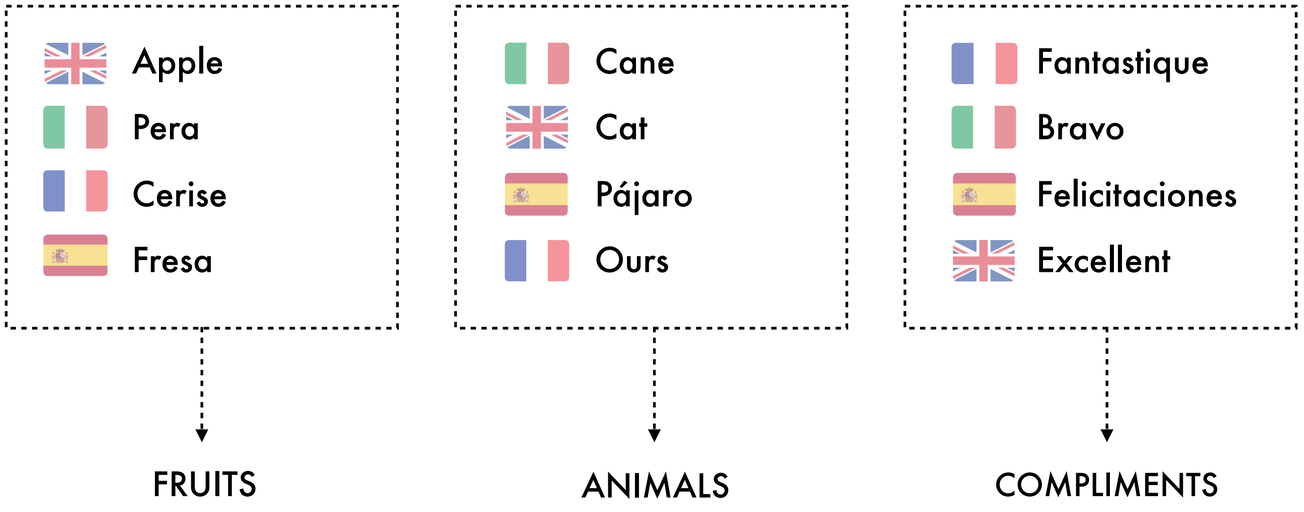
[one model for all languages]

Words are clustered by intent

Using:

NEURAL NETWORKS

Functional form of similarity, end-to-end learning



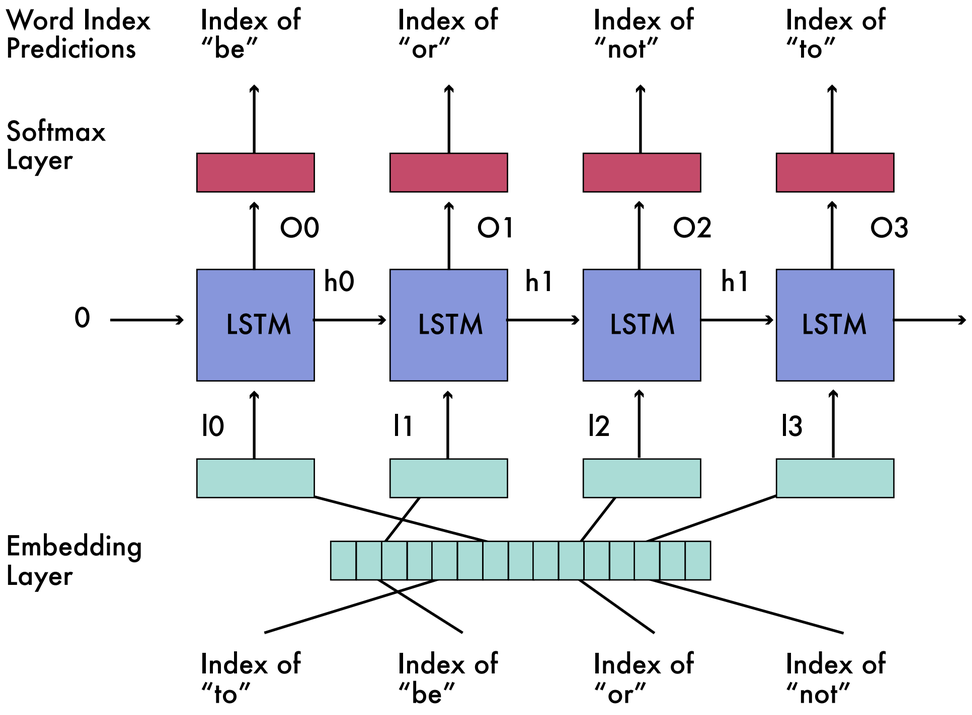
xx

Idea 1: Word Embeddings

How do we train embeddings?

- NNLM [Neural Network Language Modelling]
- NNLM + Skip-gram connections
- RNN [Recurrent Neural Network]

  (Variants: LSTM, GRU, etc.)



Part of the Neural Network architecture
http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

Word Embeddings - MONOLINGUAL

What words have embeddings closest to a given word?

[Pre-trained Demo]

France	Jesus	Xbox	Reddish	Scratched	Megabits
Austria	God	Amiga	Greenish	Nailed	Octets
Belgium	Sati	Playstation	Bluish	Smashed	Bb/s
Germany	Christ	Msx	Pinkish	Punched	Bit/s
Italy	Satan	IPod	Purplish	Popped	Baud
Greece	Kali	Sega	Brownish	Crimped	Carats
Sweden	Indra	psNumber	Greyish	Scraped	Kbit/s
Norway	Vishnu	Hd	Greyish	Screwed	MegaHertz
Europe	Ananda	Dreamcast	Whitish	Sectioned	MegaPixels
Hungary	Parvati	Geforce	Silvery	Slashed	Gbit/s
Switzerland	Grace	Capcom	Yellowish	Ripped	Amperes

From Collobert et al (2011)

Idea 1: Word Embeddings

Transfer Learning - BILINGUAL ALIGNED EMBEDDINGS

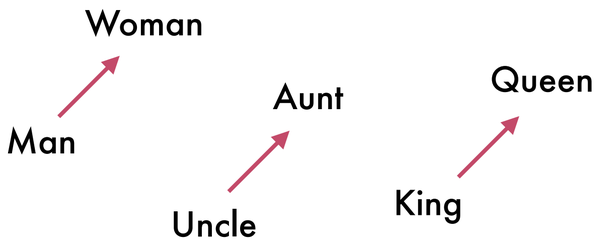
1. Monolingual embeddings

Bilingual dictionaries to translate

http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

2. Train bilingual embeddings online

Alignment to translate



From Mikolov et al. (2013a)

Idea 1: Word Embeddings

Basic Results and Concerns

Pre-trained embeddings issues

Fashion Language - Brand Names [#katewinslet #katemiddleton #katetokyo]
Get mapped to the same epsilon neighborhood

They all have similar vector value
because they are not recognized as valid vocabulary

Should we train our own embeddings?

Problems:		Solutions:
Data Size - particularly in foreign languages	----------→	Scrape similar social media fashion data
Cross-Lingual Embedding alignment		

Future Implementation

Idea 2: **Model Translation - Transfer Learning**

Concept

Expressed idea

↓

Latent variables

↓

Latent variables transfer across languages

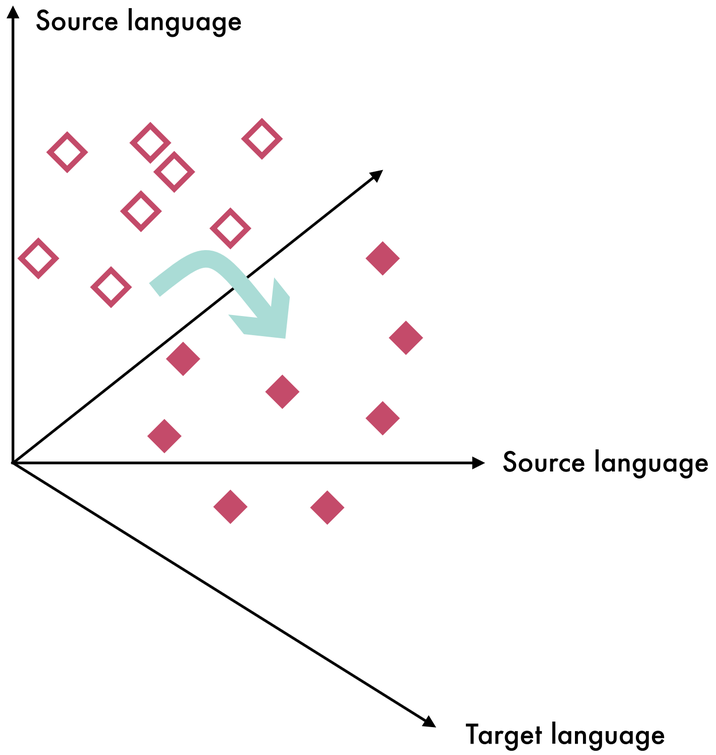
↓

All languages have underlying similarity in expressing ideas

Idea of
MODEL TRANSLATION // DOMAIN ADAPTION

Debate

Latent model learning:
Exact Inference vs. Variational Inference



http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

References

1. http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/\
2. http://web.eecs.umich.edu/~mihalcea/papers/shi.emnlp10.pdf
3. https://ai.stanford.edu/~wzou/emnlp2013_ZouSocherCerManning.pdf
4. Special Thanks: FAIR – FastText and Muse Embeddings

Plan

Project Time-line

March

Day **1**

Baselines Development

20

Word Embeddings

31

Domain Transfer

April

Day **10**

Finetune model results

20

Evaluate cost-benefits and rank model alternatives. Build model visualization tools

30

Deliver Project

Thank you

TRIBE DYNAMICS

POLITECNICO
MILANO 1863

IACS
INSTITUTE FOR APPLIED
COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY