# Scope of Work for Tribe Dynamics

Version 1.1, 2018-02-01

|  |  |
| --- | --- |
| **Prepared by** | Srivatsan Srinivasan, srivatsansrinivasan@g.harvard.edu  Andrea Porelli, anp398@g.harvard.edu  Alessandro Bianchi, alessandro16.bianchi@mail.polimi.it  Ginevra Terenghi, ginevra.terenghi@mail.polimi.it |
| **Prepared for** | Pavlos Protopapas, pavlos@seas.harvard.edu |
| **Summary of changes** | * Version 1.0, 2017-02-13, Initial draft |

## Background

Tribe Dynamics is a consulting company that provides measurements of social media engagement for beauty, fashion and lifestyle brands. They have been consulting for the last four years (largely in the American market) and provide web-based tools and periodic reports to track the clients’ social media engagement. Their marquee offering is a metric called Earned Media Value (EMV) which is derived out of various engagement markers such as likes comments, shares etc. on their clients’ social media posts.

To calculate these metrics, they scrape popular social media platforms for textual posts using customized search phrases that are provided by their in-house experts. Once these posts are scraped and cleaned, it is imperative to filter these posts by identifying if a post is talking about a particular brand or a particular class of product before they analyze the engagement of these posts. Tribe has been highly successful in this task with American brands whose posts are largely in English with their proprietary classification model (n-grams with logistic regression).

Over the last couple of years, they have been expanding rapidly into Europe and Asian markets and naturally, the language repertoire of their posts has been growing fast. Their models do not perform with the same efficiency across languages as the linguistic structure of these new languages are largely different and the models need sufficient training data in new languages before they train well. This problem is at the center of what the project tries to accomplish - provide a model that scales well across languages with limited amount of training data by exploring the structural similarity across languages.

## Problem statement

### Goal

We intend to improve upon Tribe’s proprietary existing classification model to essentially develop a brand and vertical classification model that scales well across languages with limited training data. The model is expected to learn and adapt the latent language structure behind social media posts across languages. To this effect, we intend to evaluate and develop the most cost-benefit viable alternative that best fits Tribe’s business needs and constraints among the following themes - a.) Improving upon existing n-gram logistic model; b.) Standard classifiers with multi-language aligned word embedding representations; c.) Machine translation API with a single classifier in source language (English); d.) Well-constructed source model with model translation/domain adaptation across target languages with semi-supervised learning. The goal is also to provide a well-constructed evaluation framework that incorporates extreme examples for testing and post training, summarizes key model performance attributes and constraints. The developed model should be cognizant of the following factors: a.) The new model should either improve the current performance on the source language directly or bring a lot of performance scale across languages with marginal loss of source accuracy, thus still providing better overall performance across the training data corpus; b.) The model should be scalable across any language that may populate the data and c.) Computational complexity of the suggested model should be feasible within Tribe’s business resources, timelines and demands.

### Resources available

Data available includes:

* Of the order of a few 1000 records of posts gathered for each of 106 different brands from different social media platforms (i.e. Facebook, Twitter, Instagram) which are either manually labeled by Tribe’s employee (<10%) and through on-demand crowd services (> 90% posts) structured on a CSV containing:
  + Brand id
  + True or False label by employee (one column for Mturker and one for Tribe’s employee plus worker id)
  + Hash to identify the post
  + Text of the post in original language
  + Date when the text was added to the data set
  + Model classification for brand decision (belong or not to the brand)
  + Model classification for language (which language the post is written)
  + Timestamp whether or not the record was used to train the model

## Deliverables

The deliverables will be all necessary code, assets, and documentation necessary for the Client to run on their system fulfilling the following requirements:

|  |  |
| --- | --- |
| **Deliverable 1** | A **predictive model** along with an exhaustive model document trained on Tribe Data which :  ESSENTIAL FEATURES   * Provides either a.) better classification performance (Accuracy, AUC etc.) on source language with comparable performance across target language or b.) provides better performance scale across target languages with comparable performance in source language. * Transfers internal representation across languages (through one/ensemble of word translation, model translation or word vectors) and fine-tunes performance in target language with limited data on semi-supervised learning.   WISHLIST FEATURES   * Provides Cross-language aligned word-embeddings that compresses similar text across languages to a fixed dimensional vector space to allow for scale via finite and consistent model input dimensions. * Is finely optimized for non-aligned objectives of training constraints (time to train, frequency of training, complexity/interpretability) and out-of-sample test performance. |
| **Deliverable 2** | Python **module** which:   * Can be run as a standalone script or imported programmatically into other parts of the Client’s codebase. * Import a frozen trained model and directly use it to predict product/brand categories. * Runs training on different variants of the best model identified from structured trials and experiments. * Generates summary statistics and performance metrics (Accuracy, AUC etc.) of the trained model. |
| **Deliverable 3** | Thorough cost benefit analysis on different alternatives among existing n-gram models, deep NLP models, paid translation across languages and domain adaptation/transfer learning models. |
| **Deliverable 4** | Brief **recommendations** on model alternatives/improvements and further courses of action. |

## Project timeline

|  |
| --- |
| **Note:** It is okay if this changes along the way, but it should be completely filled out. Once the scope is approved, the TF should be consulted before changes are made. The most important part at any given time is the next few milestones; the TF will use these to set expectations for what the team is working on and hoping to achieve. |

## 

|  |  |
| --- | --- |
| **Sprint ending** | **Tentative milestone or goal** |
| 2018-02-11(Completed) | * Receive toy data from clients * Exploratory Data Analysis & Results * Request for Real Data with Client (pending with client) * Choice of broad themes |
| 2018-02-18 | * Literature Survey on embeddings and model transfer/domain adaptation - SRI * Baseline Models for language transfer(Tested on standard data) - AP/AB * Black-box Evaluation Pipeline initial ideas - Ginevra |
| 2018-02-25 | * Compare pre-trained vocabulary for embeddings with data - SRI * Choose top 3 model ideas and create model documents - Math, implementation details (Model building costs, pros, cons, possible test cases, etc.) - Discuss with client - SRI * EDA on new data, Semi-supervised evaluation ideas - Ginevra * Evaluate ground truth reliability in English and Italian - AP/AB |
| 2018-03-02 | * Create text processing pipeline from CSV files * Find alternative ground truth mechanisms if needed (Work with Rohan) - AP/AB * Implement and test the perceived top model from literature survey * Evaluate pros/cons of Topic Models vs. One-vs-Rest for multi-label classification for baseline English. Implement the better alternative - AP/AB |
| 2018-03-09 | * Implement and test the second and third model alternatives if needed |
| 2018-03-18 | * Evaluation and Re-iteration |
| 2018-03-25 |  |
| 2018-04-01 |  |
| 2018-04-15 |  |
| 2018-04-22 |  |
| 2018-04-29 |  |