**Predicting heart-warming and inspiring unsung stories for The Positive India**

**Problem statement** :

The positive india is news portal focussed on writing about unsung heroes, heartwarming stories and community development initiatives. Finding leads without a journalist network around the country is really challenging. Hence, we came up with solution of developing an AI based model which predicts heart warming or relevant stories. At the end of the project, the deliverable will be a script which when triggered scraps articles from multiple websites over a time and prints top 20% of stories which are more likely to be either heart warming or community development etc.

**Sub Tasks for the project**

1. Building a web scraper/ crawler to pick stories from top websites like TOI and Hindu which will used for training set. Existing open source crawlers can be modified according to need. Most of top websites block crawlers/ips.
2. Collecting target variable - we need to get articles which we know for sure are as per our context so that they become labelled set.
3. Preprocessing - collecting from multiple sites means each have different formats, cleaning, taking headlines or only side headlines or comments also etc.
4. Train topic modelling. Since most of our articles can be filtered with topics. Our topics are mostly community development, women empowerment, donation drives , NGOs. These topics are a big indicator while political, sports , business etc are a general No-N0. I believe topic modelling will become a crucial part.
5. Feature engineering - experimenting . Sentiment score of adjectives, most frequent and relevant words picked for count vectoriser in positive labelled articles, deploying sentence 2 vec, distance metrics etc are which I am looking for feature engineering
6. ML, AML advanced modelling techniques to be applied

This is a draft proposal beings sent for taking it course project for NLP Applications course.