Implementation of Machine Learning Approach to Detect Click Baits in Online News

In partial fulfillment of “Machine Learning,Micro Degree in Artificial Intelligence”

Prajwol Lamichhane
Student, Microdegree in Artificial Intelligence
Fuse Machines Inc.
Kathmandu, Nepal
prajwollamichhane11@gmail.com

Kamal Shrestha
Student, Microdegree in Artificial Intelligence
Fuse Machines Inc.
Kathmandu, Nepal
kamal.shrestha@student.ku.edu.np

Motivation

- How many people here have clicked in a youtube video and realised that it is nothing like its title?
- How many people here have clicked on an exciting link in Facebook,like “Was It an Alien or Something?! Can't Wait to Find Out!”… just to end up in a random blog article or something
- I am sure many of us has faced malware attacks because of randomly clicking in baited articles and downloadable links
- I am sure some of us has atleast one irrelevant email address to sign up in random websites, because we are so tempted to know what the website contains, all because of awesome clickbaited headlines.

From Youtube to Facebook to Blogs, ClickBait exists everywhere.

Introduction

Problem Definition:

- golden rule of journalism, taught to any news reporter at the beginning of their career - your introduction should grab the reader straight away.
- Without interest in a headline, there will be no interest in the story.
- every media outlet has to compete with many such outlets for reader attention and make their money from the clicks made by the readers
- catchy headlines accompanying the article links, which lure the readers to click on the links = Clickbaits
- this approach clearly entices readers, it perpetuates gossip and mistruth, and journalistic integrity becomes ever weakened.
- Slant, an online magazine, pays writers $5 for every 500 clicks on a story on top of their monthly salary* and this is a pattern followed by more and more websites.

*According to the Columbia Journalism Review

Project Goals and Milestones

1^0

- Detection of Click baited news headlines from different online news outlets
- Create user friendly web platform to perform classifications.
- To avoid any user agitations and prevent unauthorised redirecting to different websites phishing personal informations.

2^0

- Implementation of Machine Learning Algorithms (Supervised) to make proper classifications.
- To understand the concepts of loss function, optimization, regularisations, overfitting and more.
- Proper use of Python Tools, visualisations and Sklearn Libraries.

Proposed Solution

- Implementation of Supervised Machine Learning Classification Algorithm like Logistic Regression to learn patterns and make predictions based on the lexical or syntactic features extracted from manually labelled news headlines

- Implementation of Semantic Features like Word Embeddings, N-grams, Similarity and Source categorization to improve the prediction accuracy

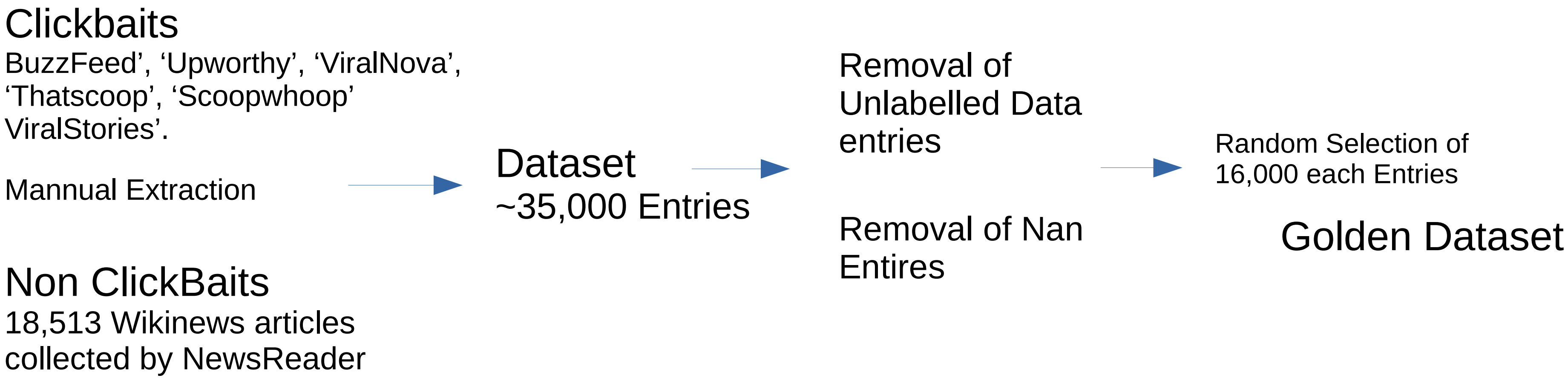
- Implementation of Deep Learning Techniques like Convulated Neural Networks to make classifications

- Create a web platform to enter and check if the given news headline is Clickbait or not OR create a chrome extension that automatically flags or highlights clickbaited news headlines

Implementation Details

- Data Pre-Processing
- Data Visualization
- Features Extraction and Selection
- Model Generation
- Results Evaluations

Dataset and Preprocessing



Clickbaits

BuzzFeed’, ‘Upworthy’, ‘ViralNova’, ‘Thatscoop’, ‘Scoopwhoop’ ViralStories’.

Mannual Extraction

Non ClickBaits

18,513 Wikinews articles collected by NewsReader

Dataset

~35,000 Entries

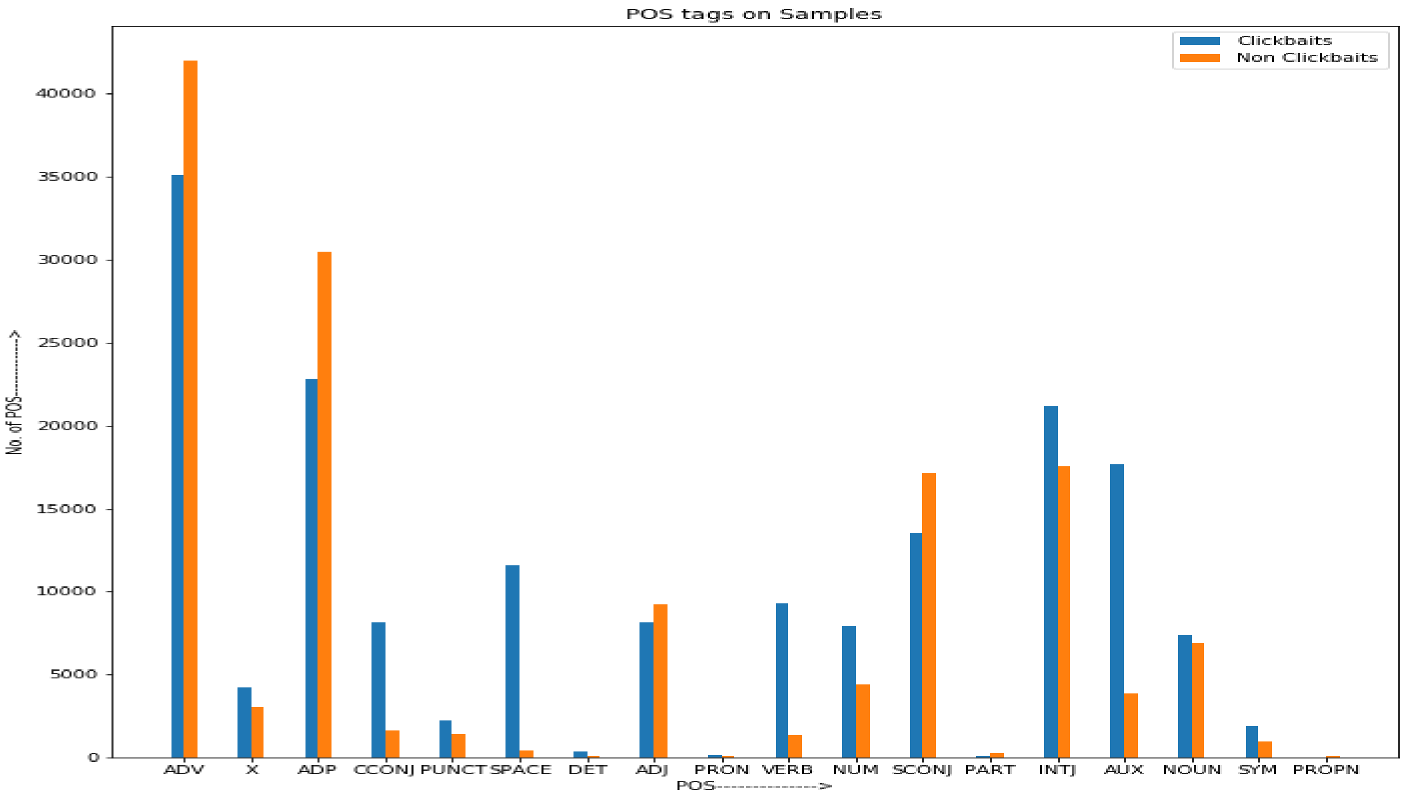
Removal of Unlabelled Data entries

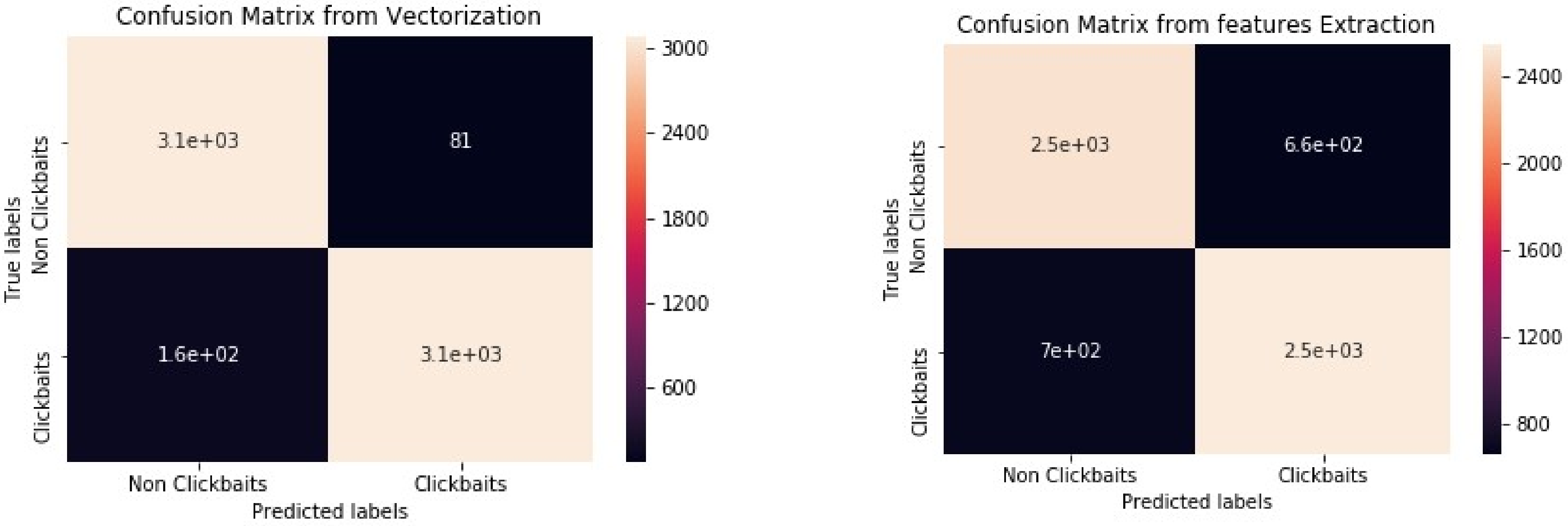
Removal of Nan Entires

Random Selection of 16,000 each Entries

Golden Dataset

Data Visualizations





Confusion Matrix from Vectorization

Confusion Matrix from features Extraction

Confusion Matrix Plots based on different approaches to train and test Logistic Regression Model

Features Extraction and Selection

Normal Analysis on Texts

Length

Words Counts

Average Word Length

Stop Words to Content Words Ratio: Cardinalty

POS Tag Analysis

Co-Ordinating Conjunctons

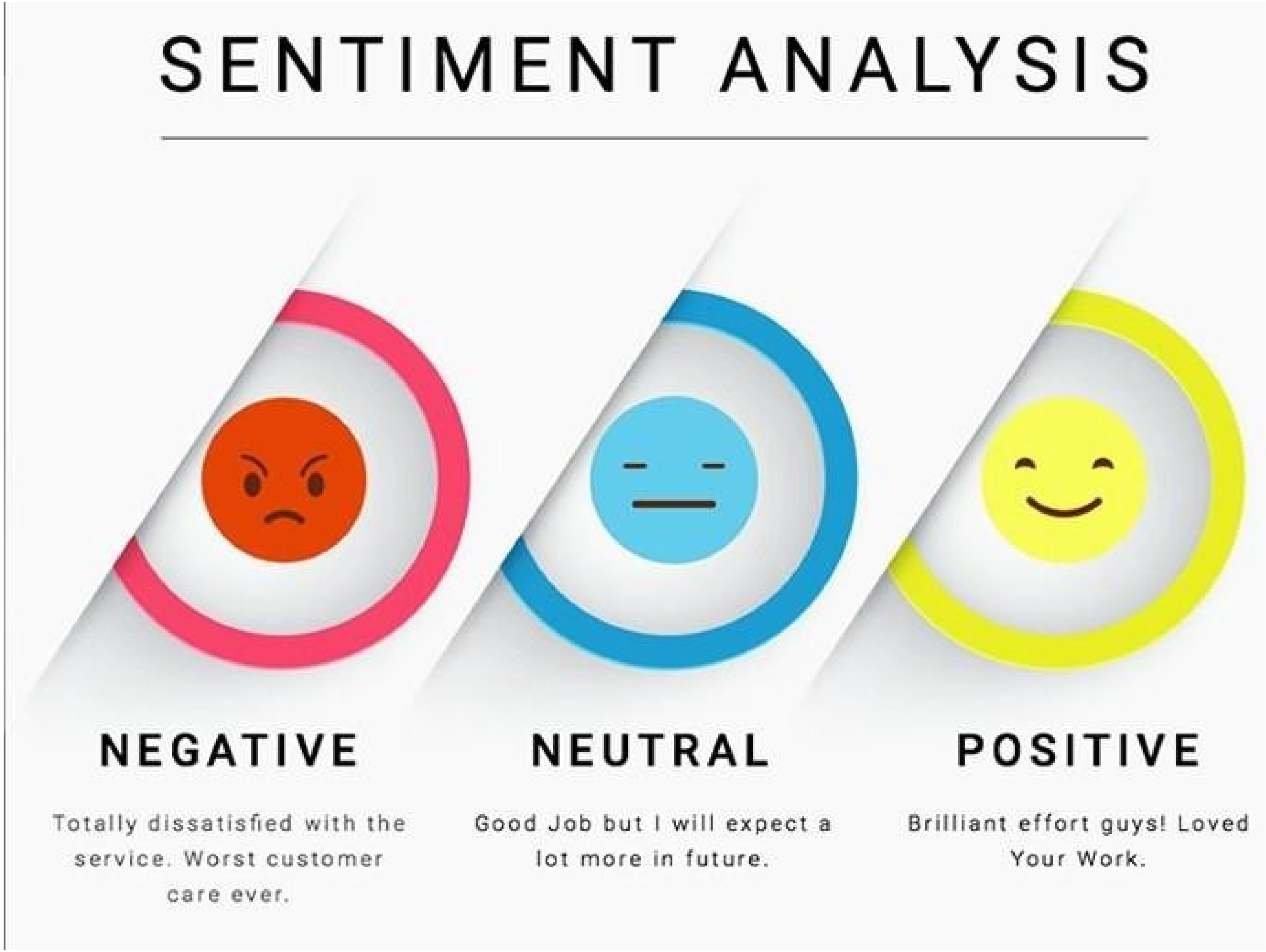
Verbs

Auxillaries

Sentiment Analysis

Polarity
Subjectivity

Example: Three Simple steps for shedding belly fat over 40.



SENTIMENT ANALYSIS

NEGATIVE

Totally dissatisfied with the service. Worst customer care ever.

NEUTRAL

Good Job but I will expect a lot more in future.

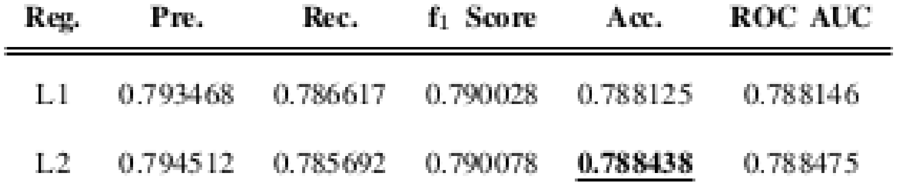
POSITIVE

Brilliant effort guys! Loved Your Work.

Model Generation

- Supervised Learning (Logistic Regression Classification Model)
- Loss Function: Cross Entropy Function
- Optimization Function: Gradient Descent
- Regularization: L1, L2
- Calculation Metrics: Precision, Recall, F1 Score, ROC AUC & Accuracy
- Predictions

Results Evaluations



Reg.	Pre.	Rec.	f_1 Score	Acc.	ROC AUC
L1	0.793468	0.786617	0.790028	0.788125	0.788146
L2	0.794512	0.785692	0.790078	**0.788438**	0.788475

An f1 score of 0.790 solely based on lexical features paves a promising pathway for the introduction and inclusion of semantic features like **word embedding** and **n-grams** in future works.

On **varied regularization** penalty, there was no any significant changes to the result indicating the model is **not over fitted** and the accuracy is **not misleading.**

With an addition of source of extraction of the news headlines itself as a feature, the results is sure to be improved.

OTHER APPROACHES

1. Cosine Similarity
2. Word Vectorization

Cosine Similarity has its own problems. There is a dilemma for making similarity comparison, either with manually labeled headlines or with randomly selected headlines from golden data set.

The word vectorization approach yielded maximum score in testing dataset, but performed poorly in further experimentation with foreign entries clearly demonstrating overfiting.

Problems Faced

1. Conduct the study for detection of clickbaits in **Nepali** online news portals but lacked any proper dataset
2. Create an browser **extension** to automatically detect and flag click baited headlines, but lacked the domain knowledge in creating extensions
3. Complicated Data Visualizations due to **large** number of data entrires
4. Basic **Features Extraction** and NLP Techniques.

Conclusion and Future Enhancements

- Compared Clickbaits and Non Clickbaits based on several highlighting differences, further used to train the model
- Clickbait Detection Model Developed

Furthermore,

- Implementation of Semantic Features and different NLP techniques
- Implementation Using Multilingual approach
- Implementation of Deep Learning Techniques

Thank You

I don't always create clickbait
But when I do, I find it contributes significantly to the success of my brand
imgflip.com

Any Queries ?