Assalamu alaikum I am omok my id is eto my team member is omok omok omok. We complete our research under the supervision of omok sir . Now I am going to describe about my presentation

**Slide 2: Table of Content**

The table of content contains Introduction, Motivation, Literature Review, Methodology, Data collection, Preprocessing, Tokenization, Algorithm Implementation, Experimental Result, Evaluation, References.

**Slide 3: Introduction**

Newspaper is very important part in our life. With the growth of technology, the demand of news heading recommendation in Bangla language is increasing. In our work we used traditional machine learning algorithm for making recommendation engine in bangla language.

**Slide 4: Motivation**

* Nowadays the popularity of recommendation engine is increasing, there are lots of recommendation engine is available in English or other language but unfortunately there is no effective recommendation available in Bangla language this was the main point of our motivation. And for this reason we started our work.

**Slide 5: Objective**

In this work our main objective is

* To make a recommendation system in Bangla language
* To help people to reduce time while he/she search news in online.

**Slide 6: Literature Review**

This is the main two works that helps to get idea about our work, in these work authors did not use traditional machine learning algorithm with tfidf, but in our work we used traditional machine learning as well as tfidf that makes our work unique form others work.

**Slide 7: Methodology**

The methodology of our work is very complex process, there are main 5 steps in this section, that is data collection preprocessing tokenization algorithm implementation and evaluation

**Slide 8: Data collection**

We collected our required data from kaggle website, we collected our data from different newspaper website like prothomalo.com, dailynayadiganta.com bd-pratidin.com ittefaq.com. bdjugantor.com . our initially collected data was about 20000 but after preprocessing we found 10000 quality full data. the next slide represents how we got qualityfull data.

**slide 9: Preprocessing**

There are 4 steps of preprocessing html removing, non-Bangla removing, data cleaning, target manipulation, the process of preprocessing is done by lxml, bnltk, nltk, and our own created function.

**Slide 10: Selected target.**

There are about 60+ target in our initial dataset. But all target has not enough data even some target comes only one time, these data are actually outliers, so we remove them and keep targets which data comes at least 100 times. Then we found about 10 targets like Bangladesh, Crime, Cricket, Politics Education, Entertainment, life, Football, Science-tech, with total 10000 data.

**Slide 11 Dataset Balancing**

After selecting target our dataset was imbalance, Bangladesh category contains higher percentage among all the targets, and the percentage rate is about 48.0%. on the other hand, science-tech contains about 3.5%. that means our dataset set is very imbalanced. We balanced this dataset by applying up sampling and down sampling. Left site figure represents imbalanced dataset and right site figure represents balanced dataset.

**Slide 12 Tokenization:**

Our work is based on word level. That why we needed to tokenize each word form a sentence. After tokenization we used TFIDF algorithm to convert word to numerical vector.

The sample of tokenization is given in this table.

**Slide 13 Algorithm implementation**

In our work we have used 6 different machine learning algorithms. Like logistic regression, decision tree, random forest, multinomial naïve Bayes and SGD classifier. We have chosen the parameter by which algorithm produced best accuracy. And for hyper parameter tuning we used manual process.

**Slide 14 Experimental Result:**

for data usage rate we used 25% to 40% test data. when we used 25% test data then training size is 75% same way when test data is 40% then training is 60%. We apply this technique for machine learning algorithm. And we can see from this table the highest accuracy is achieved by Stochastic Gradient Descent classifier algorithm. So, we decided to use SGD classifier for final prediction.

This table represent the accuracy table with stop word dataset. For each algorithm perform better with this dataset but the highest accuracy is achieved by SGD classifier the accuracy rate is about 90.29

**Slide 15 Evaluation**

As SGD classifier perform very best everywhere so we have to decide to use SGD classifier for real life prediction. In this state we have collected around 300 data that is never seen by our model in train or in .testing phase. All of 10 target exist in this dataset. This figure represents the real and prediction graph. In this graph Blue color bar represents the real value and orange color bar represents predicted value. We can see that for most of the target our system predicted very accurately. So we can say that our model is also very good for real life data.

**Slide 16: future work**

In future we will create a Recommendation system for Multi language detection

We will use deep learning algorithm such as LSTM(Long short term memory) ,RNN(Recurrent neural network) instead of Traditional Machine Learning

**Slide 16: References:**

Here is all references that we used in our work.

**Slide 17:**

This is all about our presentation thank you every one. If you have any question, please ask.