Artificial Intelligence

Project report (SentiMax Hindi-English)

By:

Muhammad Bilal

17i0317

Sec-B

Introduction

The Problem

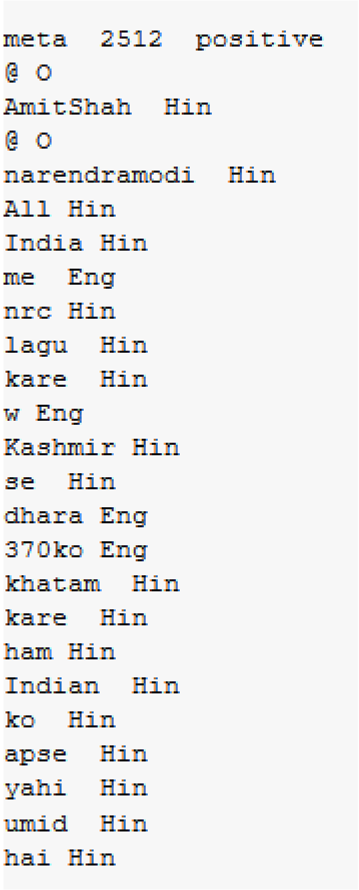
The problem pertains to sentiment analysis of mixed Hindi-English tweets. The sentiments are classified as: Positive, Negative and Neutral

The Data

The tweets have been directly scraped from Twitter. There are a total of 17000 tweets. These have been split into train data (14000 tweets) and test data (3000 tweets). Some details are:

- The tweets are in conll format
- Each tweet has been labelled into positive, negative and neutral sentiments
- Each word is labelled into hindi and English. Words that are not from these languages are labelled as 0.
- The tweets contain emojis, urls and usernames.
- The tweets are both lower and upper case.

The tweets looked like this:



```
meta  2512  positive
@ O
AmitShah  Hin
@ O
narendramodi  Hin
All Hin
India Hin
me  Eng
nrc Hin
lagu  Hin
kare  Hin
w Eng
Kashmir Hin
se  Hin
dhara Eng
370ko Eng
khatam  Hin
kare  Hin
ham Hin
Indian  Hin
ko  Hin
apse  Hin
yahi  Hin
umid  Hin
hai Hin
```

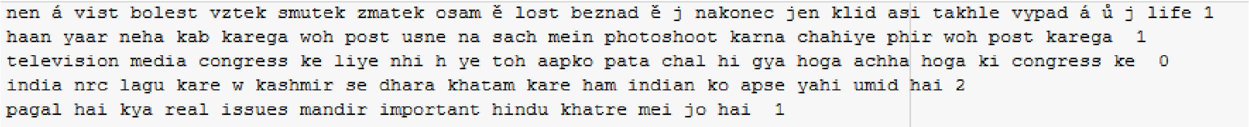
Proposed Approach

Data Cleaning

Considering that the tweets are still in raw format. It was necessary to first clean the data. The sequence followed was:

- The tweets were first joined to form sentences.
- URLs, usernames and emojis were removed.
- Sentences were converted to lower case.
- Unnecessary punctuation was removed from the sentences.
- Stop words (from nltk corpus) were removed
- Sentences were lemmatized and stemmed (depended on experiment being run)

After all these steps (except lemmatizing or stemming) the tweets looked like this:



```
nen á vist bolest vztek smutek zmatek osam ě lost beznad ě j nakonec jen klid asi takhle vypad á ů j life 1
haan yaar neha kab karega woh post usne na sach mein photoshoot karna chahiye phir woh post karega  1
television media congress ke liye nhi h ye toh aapko pata chal hi gya hoga achha hoga ki congress ke  0
india nrc lagu kare w kashmir se dhara khatam kare ham indian ko apse yahi umid hai 2
pagal hai kya real issues mandir important hindu khatre mei jo hai  1
```

Vectorization

Once that data was cleaned it was time to vectorize the tweets. I tried a number of approaches which revolved mainly around Bag of Words approach.
Basically, what I did was, I created a vocabulary of distinct words that appeared in the entire data at least K number of times (K became a hyper parameter, set at 15). This vocabulary became the template of each vector.

Each tweet became a vector of size vocabulary, where words of the vocabulary which were present in the tweet were given a non-zero numeric value. Words that did not appear were marked with 0. This vector than becomes our tensor that we use.

Another approach I tried, (and which is in the final submission) is I used Bigram in bag of words. So, while creating the vocabulary I take into consideration bigrams as well. This way the vocabulary contains both single words and bigrams. In turn the vectors themselves also follow the same convention.

Deep Learning

In the paper I followed an interesting finding was put forward. It said that the accuracy of the task did not depend on the amount of training of the model and the architecture of the model itself. (ref: https://arxiv.org/abs/2007.13061)

With this is mind I created a very simple CNN with 2 dense layers and a relu and softmax layer. For this network I used a Cross Entropy Loss along with a simple Stochastic Gradient optimizer. The optimizer was initialized with a 1x10^-3 learning rate and 9X10^-1 momentum. I trained this network for 40 epochs with a batch size of 6.

I experimented with some more complicated architectures, but there was no significant improvement in results.

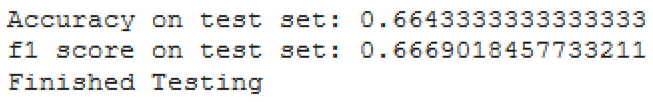
After training this model was tested using the provided test dataset.

Results

The final accuracy: 0.6643333333333333

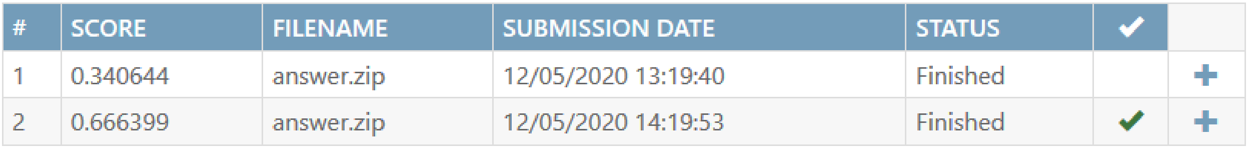
The final F1-score: 0.6669018457733211

Image from notebook:



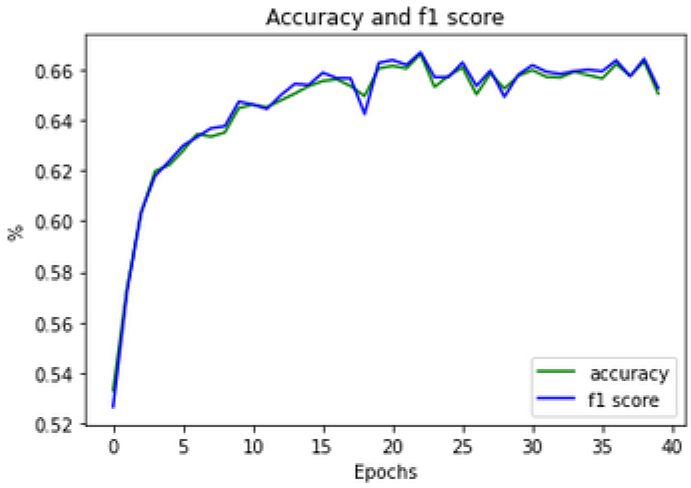
```
Accuracy on test set: 0.6643333333333333
f1 score on test set: 0.6669018457733211
Finished Testing
```

Codelab score

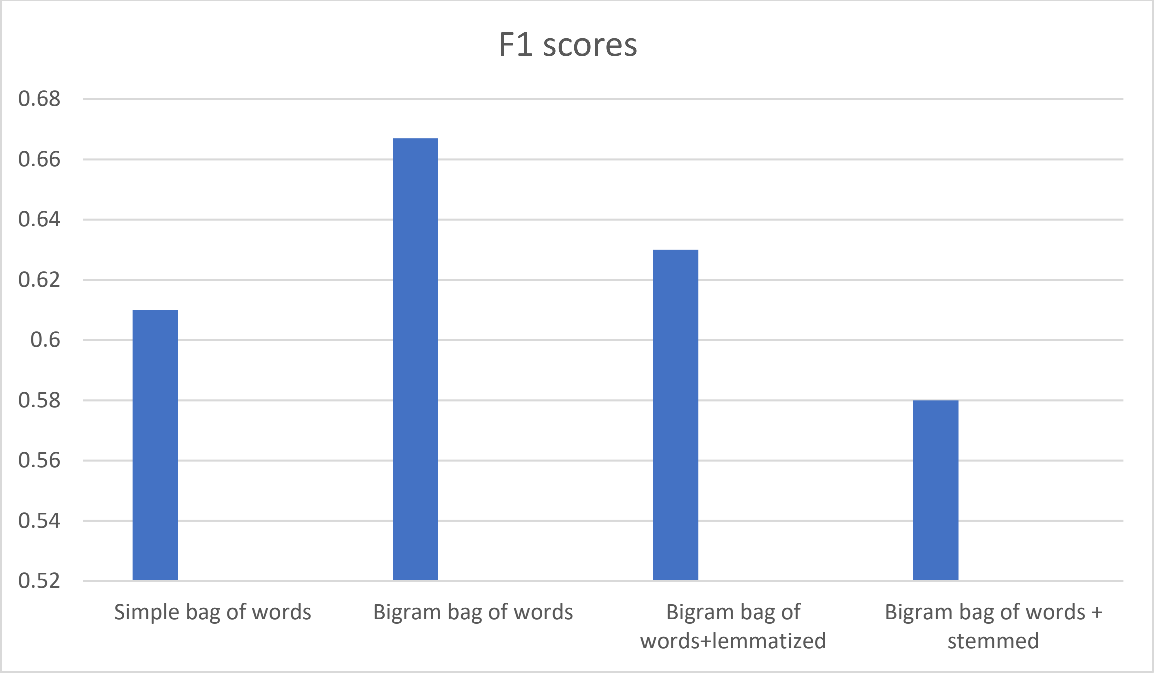


#	SCORE	FILENAME	SUBMISSION DATE	STATUS	✔	
1	0.340644	answer.zip	12/05/2020 13:19:40	Finished		+
2	0.666399	answer.zip	12/05/2020 14:19:53	Finished	✓	+

Training graph of final submission:



The different results from different approaches:


F1 scores
Simple bag of words
Bigram bag of words
Bigram bag of words+lemmatized
Bigram bag of words + stemmed