Department of Computer Science & Engineering
University of Dhaka

Project Report

Mathematics for Computer Science(CSE-3205)

Submitted to

Dr. Amin Ahsan Ali

Group Members

Redwan Ahmed Rizvee (Rol l : 09)
Md. Shahadat Hossain Shahin (Rol l : 35)
Mehreen Rahman (Rol l : 61)

Table of Contents

Project Category:

Natural Language Processing with the help of word to vector and and K means Clustering

Project Short Details(What we wanted to do):

1. Giving a machine verdict about a bengali sentence whether it's expresses a positive meaning or negative meaning.
2. Trying to find vector difference semantics about a word which is both used in positive and negative sentence.
3. Using K means clustering to interpret sentence topic(both good interpretation meaning and false interpretation meaning)

Analysis of our task 1:

"Giving a machine verdict about a bengali sentence whether it's saying a positive meaning or negative meaning."

This part was the core part of our project, to determine if a bengali sentence is giving a positive type meaning or a negative type meaning.

The Data part:

Our Data was collected from "**Prothom Alo, annotated comments.csv"** provided by Sir Dr. Muhammad Asif Hossain Khan and Tamim Ad Dari vai from 19th Batch. There were about 12000 lines of prothom alo comments including about 30000 bengali new words.

Training Session:

We couldn't provide enough GPU and we faced several memory related and stack architecture related issues. Though we had a huge dataset we couldn't perfectly use all the datas.
Our procedure was, first we classified the positive, negative and neutral sentences in three different files from the **annotated comments file** in three different files named "**Positive Sentence**", "**Negative Sentence**" and "**Neutral Sentence**". Now these three types of files only

had their own specific meaning related sentences. Ex- Positive Sentence file only had positive meaning sentences.

We almost used 6000 words per file for training. At this period we trained computer for per files and generated word to vector for each word. There were about 5 feature selection (EMBEDDED_DIMENSION), 100 Training iterations, learning gradient descent rate optimizer by 0.1 with the base of skip gram model and along with window size of two.

After the end of calculation, we got vector presentation for each word in each file.

"Our Basic hypothesis was to have three different vectors for each word of their positive meaning vector, negative meaning vector and neutral meaning vector depending on their context"

Hypothesis 1:

"We have chosen 5 dimension or characteristics for each word vector representation. So as word to vector learning uses its context to learn and determine its vectors. So when same characteristic feature will be learned by machine they will be correlated and two different feature will be independent. Suppose

আমি ভাত খাই । is our sentence.

Now আমি vector is = [a b c] and ভাত vector representation is [d e f] and খাই representation is [g h i].
Now when vector will be created these three vectors will be selected based on their context and for these reason (a,d,g) should be related as they are same feature, similarly for(b,e,h) and (c,f,i).

Now our first hypothesis was to take all type of vectors first take all positive vectors for each word, then take for negative vectors and then take for neutral vectors for each word. And calculation will be
*"Same feature multiplication" and "different feature summation" . Like in our example, the hypothesis result was, (a*d*g + b*e*h + c*f*i) for one type of vector. Similarly take for negative vectors and neutral vectors. And take the absolute minimum to be the answer. And if positive was minimum it's a positive sentence,if negative was minimum then negative sentence otherwise neutral sentence. Why taking minimum because when two vectors are related their cosine difference should be minimum. If words are related their vectors came dependently and so when their modified-dot product is taken it should be minimum. "*

Result: minimum satisfactory

Hypothesis 2:

We took a sentence and generated word to vector depending on the context only for that sentence. Then we compared for this new vector of this word with other three different type of vectors and for each word we gave decision either it’s a positive word, or a negative word of a neutral word and counted. The maximum count solution gave the verdict of the sentence. Like if there is positive words are dominating then it’s a positive sentence either others.

Result : Not Good

Hypothesis 3:

Now when we faced a sentence then first we calculated all the possible pair combination maintaining the window restriction. Now for each pair we looked if this pair is found in positive file then we calculated the vector difference/Norm among them, similarly if this pair is found in negative file then again calculated vector difference /Norm and same goes for neutral also. Now we calculated the minimum vector difference among them and from that we choose the word type. And using all of them the dominated type determined the sentence type. Why it should work because words are correlated so correlated words vector difference should be minimum.

Result - Best among two

হুমায়ূন আহমদের খুবই সুন্দর আর আমার ফেভোরিট একটা বই "নিতু তোমাকে ভালবাসি"

positive

বালডাইও হইল না এতক্ষণে।

negative

আমরা পারছি ফাহিম ভাই !!!!

postive

সুবোধ অনেক আগে দেশ ছেড়ে পালায়ে গেছে।

negative

আমাদের সারাদিনের কাজ অবশেষে কাজ করছে। আলহামদুলিল্লাহ।

positive

ওয়াহ ম্যান দিস অ্যাকুর্যাসি ওয়ার্কস।

positive

ট্যুরের যাই না কেন আমরা।

negative

কম্পিউটার তুমি জান, আলগো যে কাজ করতেছে।ইইইইয়া

positive

তুমি আমার বালডা বুঝছ।

negative

ফাহিম ভাই, জানেন আলগো কাজ করে।

positive

আমরা ডিসেম্বরে আবার ট্যুরে যাব। ইইইইইইই

positive

ফাহিম ভাই আপনি জোস।

positive

স্বাধীন মিরপুরের প্রথম প্রেসিডেন্ট আমাদের প্রিয় ফাহিম ভাই।

positive

আল্লাহ চাইলে শাহিন রা ওয়ার্ল্ড ফাইনালসে যাবে।

positive

বাপ্পা মজুমদার আমার আদর্শ।

positive

চলেদুর সুপারহিরোর নাম ফাহিম আরেফিন।

positive

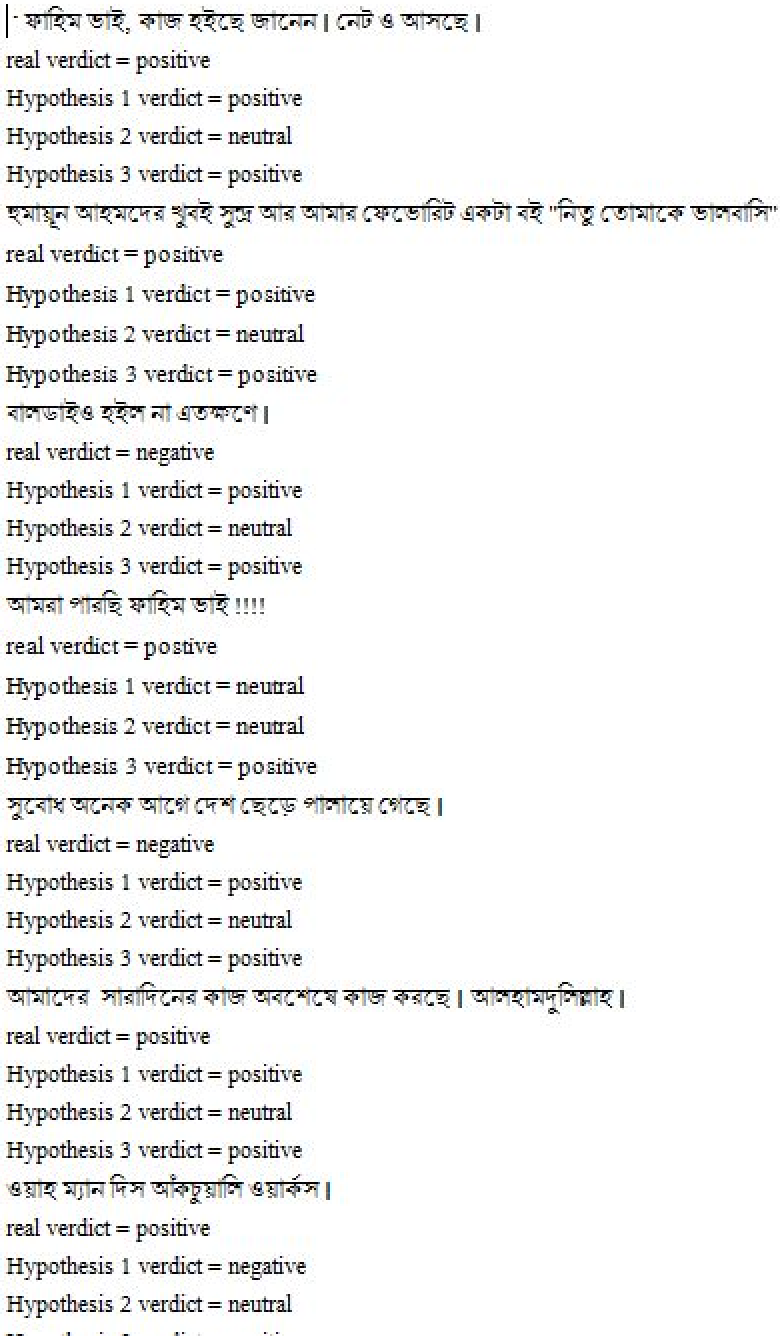
তুমি বন্ধু নামে শত্রু।

negative

আজ আমরা আমাদের জাতির পিতাকে হারিয়ে শোকাচ্ছন্ন।

neutral

[illegible]



· ফাহিম ভাই, কাজ হইছে জানেন। নেট ও আসছে।

real verdict = positive

Hypothesis 1 verdict = positive

Hypothesis 2 verdict = neutral

Hypothesis 3 verdict = positive

হুমায়ুন আহমদের খুবই সুন্দ্র আর আমার ফেভোরিট একটা বই "নিতু তোমাকে ভালবাসি"

real verdict = positive

Hypothesis 1 verdict = positive

Hypothesis 2 verdict = neutral

Hypothesis 3 verdict = positive

বালডাইও হইল না এতক্ষণে।

real verdict = negative

Hypothesis 1 verdict = positive

Hypothesis 2 verdict = neutral

Hypothesis 3 verdict = positive

আমরা পারছি ফাহিম ভাই !!!!

real verdict = postive

Hypothesis 1 verdict = neutral

Hypothesis 2 verdict = neutral

Hypothesis 3 verdict = positive

সুবোধ অনেক আগে দেশ ছেড়ে পালায়ে গেছে।

real verdict = negative

Hypothesis 1 verdict = positive

Hypothesis 2 verdict = neutral

Hypothesis 3 verdict = positive

আমাদের সারাদিনের কাজ অবশেষে কাজ করছে। আলহামদুলিল্লাহ।

real verdict = positive

Hypothesis 1 verdict = positive

Hypothesis 2 verdict = neutral

Hypothesis 3 verdict = positive

ওয়াহ ম্যান দিস অ্যাকচুয়ালি ওয়ার্কস।

real verdict = positive

Hypothesis 1 verdict = negative

Hypothesis 2 verdict = neutral

Hypothesis 3 verdict = positive

Analysis of Task 2:

“Trying to find vector difference semantics about a word which is both used in positive and negative sentence.”

We took common words which are both present in positive sentence and negative sentence and calculated their vector difference and and cosine difference. We thought that may be the vector difference will be much higher because of the two interpretation of the same word usage,. But unfortunately the testing result was not satisfactory. May be more data usage and learning might give somewhat better.

Analysis of Task 3: (On going)

“Using K means clustering to interpret sentence topic(both good interpretation meaning and false interpretation meaning)”

Motivation of this task was to, interpret sentence category. If true interpretation and false interpretation change sentence topic or not.
Like if we cluster depending only on positive vector semantics and only for negative vector semantics,how the sentences are interpreted, do they converge to same meaning or do they converge to different meaning. Our cluster point was not very high(just used 2) and iteration rate wasn’t also(just 5) , so we couldn’t perfectly gave any certain decision. It needs more training and dataset to test.We coded the solution but we need more testing to answer that.

References used along with library:

1. **Learn word 2 vec by Implementing it in tensorflow** https://towardsdatascience.com/learn-word2vec-by-implementing-it-in-tensorflow-45641adaf2ac
2. **Lecture 2 | Word Vector Representations: word2vec** https://www.youtube.com/watch?v=ERibwqs9p38
3. **K Means Clustering How it works** https://www.youtube.com/watch?v=_aWzGGNrcic&t=153s
4. **Tensorflow Word 2 Vec implementation** https://www.tensorflow.org/tutorials/word2vec

Special Thanks:

1. **Dr. Muhammad Asif Hossain Khan.**
2. **Tamim Ad Dari vai, 19th Batch, CSEDU**
3. **Saklain Akash,21 st Batch, CSEDU**
4. **Stack Overflow**
5. https://www.mathworks.com/matlabcentral/answers/98514-how-do-i-avoid-stack-overflow-error-in-my-matlab-code-compiled-using-emlmex-in-matlab-7-7-r2008b?requestedDomain=www.mathworks.com