Gender Bias Detection

समान - सम्मान

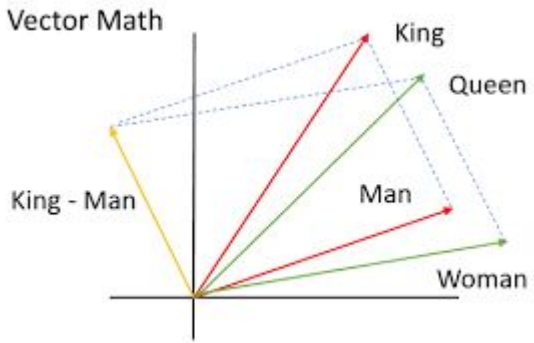
Suyash and Jainit

Word embedding system

- The most important feature of word embeddings is that similar words in a semantic sense have a smaller distance (either Euclidean, cosine or other)
- “Ped” and tree will have almost the same word

Embedding despite being in different languages

due to sense similarity



How Word Embeddings are Created

1. Read the text

How Word Embeddings are Created

1. Read the text
2. Preprocess text

How Word Embeddings are Created

1. Read the text
2. Preprocess text
3. Create (x,y) data points

How Word Embeddings are Created

1. Read the text
2. Preprocess text
3. Create (x,y) data points
4. Create one hot encoded (X,Y) matrices

How Word Embeddings are Created

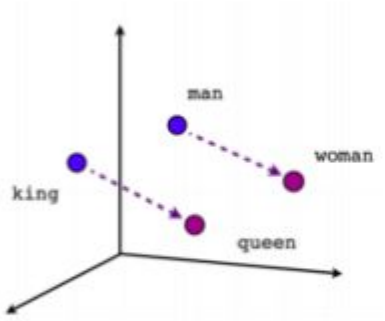
1. Read the text
2. Preprocess text
3. Create (x,y) data points
4. Create one hot encoded (X,Y) matrices
5. Train a neural network

How Word Embeddings are Created

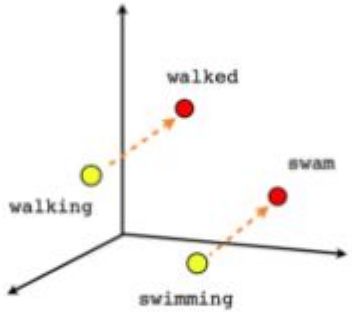
1. Read the text
2. Preprocess text
3. Create (x,y) data points
4. Create one hot encoded (X,Y) matrices
5. Train a neural network
6. Extract the weights from the input layer

Code*(vectors)

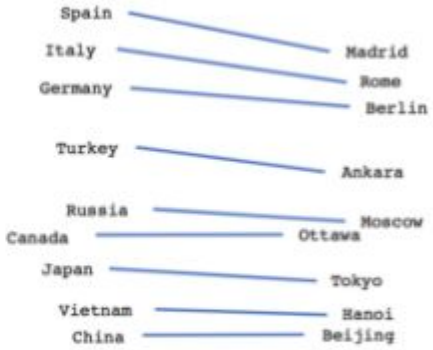
How Word Embeddings are Created



Male-Female

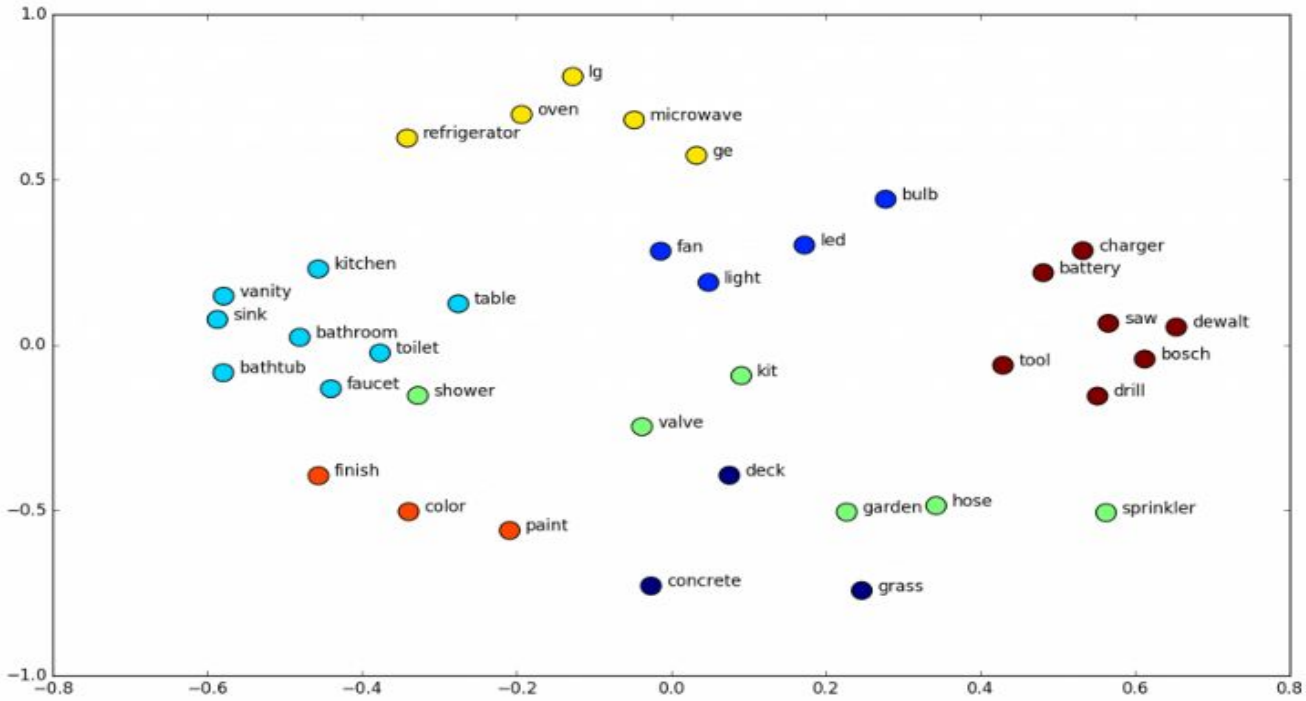


Verb tense



Country-Capital

Vector space



Our Methodology for Hindi Word Embeddings

1. Extraction of Data through Web Scraping and creation of Corpus for different Data sets like Wikipedia , News Articles
2. Creation of Word embeddings Using Python Library
3. Formation of Text Library of different domains like , intelligence , strength , Arts , Sports ,etc .
4. Formation of Text libraries of inherently Male or Female words .
5. Comparison of Euclidean and Cosine Similarites of Vectors To analys Possibilty of Gender Bias in every domain taking inheritent library as base
6. Error Analysis Based on Neutral Gender Words and already Exhisting Word Embedding Models .

Corpus Creation

1. Use of Libraries Beautiful , html5lib, re
2. Corpus Cleaning Use re for removal of disallowed characters and emojis .
3. Different from english as Hindi Inflections contain Gender information unlike English .

   Eg. Kamlesh Went to eat .
   कमलेश खाना खाने गया ।

4. Lemmatization and Stemming was not done due to the scare of losing important information

```
news_corpus = ""
string= "/home/turning/Desktop/CL_project/CL_PROJECT_CODE/Kerasmodel/aclImdb/train/pos/in"

for i, url in enumerate(all_links):

  try:
    req = requests.get(url)
    html_content = req.content
    soup = BeautifulSoup(html_content, 'html5lib')
    news_corpus_set = soup.find_all('p')
    news_corpus=""
    for element in news_corpus_set:
      news_corpus += element.text

    path_file=string+str(i%100)+'.txt'

    f = open(path_file,'a')

    f.write(str(news_corpus))
    f.close()
  except:
    pass
```

Corpus Creation

5. Removal of Stop Words like '_मैं','_मेरे','_मुझे','_उसने','_हमारे','_हमें', ,etc

- Special care was taken to not remove stop words which contained Gender Information like उसका उसकी .
- Neutral gender words were retained like उसके
- Stop words were removed after the formation of Word Embeddings so that no information is lost .

6. Statistics :-

- We collected a total of 1.2 lakh sentences to train our Model from ABP news site and approximately 8 lakh words

```
#getting the stopwords from file
f= open('/home/turning/Desktop/temp/final_stopwords.txt', 'r')
stop_words = f.read()
stop_words=stop_words.split("\n")
#stop_words = ['_','_मैं','_मेरे','_मुझे','_उसने','_हमारे','_हमें', '_मुझको', '_मेरा', '_अपने आप को', '_हमने',....

#removing the stop words
def clean_file(file_path):
    f= open(file_path, 'r')
    data= f.read()
    data = data.split('\n')
    for i in range(len(data)):
        data[i] = data[i].strip()
        data[i]=re.sub('\[\]',"",data[i])
        data[i]=data[i].split(' ')
    data_in_string= ""
    for data_list in data:
        for data_word in data_list:
            if data_word is not None and data_word not in stop_words:
                data_in_string+=data_word.strip()
                data_in_string+=' '
        data_in_string+='\n'
    return data_in_string
```

Creation of Word embeddings

1. Then input files were created to be fed into the model.

2. Word embeddings were created using two different libraries,

- Gensim library
- Keras library

Gensim Library

- This has modules that implement the word2vec family of algorithms, using highly optimized C routines, data streaming and Pythonic interfaces.
- Gensim module also has some pretrained models. They are in english, ofcourse.
- We made our word embedding using this library in hindi. Using around 50Mb of corpus around.
- We scraped the corpus and then give the data in required format.
- The words were then transformed to vectors

Keras Library

- Keras offers an Embedding layer that can be used for neural networks on text data.
- It requires that the input data be integer encoded, so that each word is represented by a unique integer.
- The Keras Embedding layer can also use a word embedding learned elsewhere.
- Keras provides the one_hot() function that creates a hash of each word as an efficient integer encoding.
- Then the encodings go through the embedding layer and they are transformed to vectors.

Code*

Challenges during designing the word embeddings

- The major challenge was training the corpus.
- We chose corpus which was only about 120000 sentences.
- We therefore had a bad frequency distribution.
- Therefore we had to compromise with the accuracy.
- The morphology in Hindi was another challenge because there were so many words having the same root and we couldn’t run lemmetiser or stemmer as the morpheme contains valuable information on Gender .
- How many dimensions to use?
- How many Epochs ?

Text Libraries

Text Libraries were created with regard to different domains which we suspected to have Gender bias .

- Strength vs Weakness
- Family vs Career

  We Made 2 Lists of Inherentetly_male_terms and inherently_female_terms which act as a reference to calculate the similarity .

  Male terms = नर,आदमी,लड़का .etc

  Female terms = नारी,औरत,लड़की ,etc

- GITHUB

Comparisons and Analysis

- We now compare the vectors formed for any bias

Euclidean Distance Analysis

- We first select a domain where we want to analyse the Gender Bias .
- We then calculate the average distance of all the vectors of that domain with all vectors of inherently male and female terms and take its ratio .

Hypothesis Testing

Hypothesis 1= the corpus biased towards male gender (Ratio>1.1)

Hypothesis 2 = the Corpus is Biased Towards Female Gender . (Ratio<0.9)

Hypothesis 3 = the Corpus is not biased (0.9<Ratio<1.1)

The Ratio 1.1 and 0.9 have been calculated by the use of hypothesis Testing by converting the Function into a Normal distribution Function .

Cosine Similarity

In this method we use 2 domains to compare Gender bias

Eg. Family vs Career

We create an average (Male - Female) Vector and We create and average (Career - Family) vector and take their cosine similarity as the basis for comparison .

Euclidean Distance

- So we measured the euclidean distance of average of all male terms vector with different domains in weat list.
- We did the same with the female average vector.
- We then compared the euclidean distance across different domains.
- We took ratio of the distance of domains with male to euclidean distance with female vector.

  Eu_ratio = eu_dist(male , domain)/ eu(female, domain)

Results

Domain	Male/Female	Comment
Intelligence	0.957744	No conclusion as hypothesis 3
Math	1.127872	Male are more into maths.
Arts	0.803066	Female are more inclined towards arts field.
Strength	1.94027	male are shown more powerful.
Weakness	0.963201	Hypothesis 3
Appearance	1.11305	Good appearance words are used for males.
Family	1.07047	Hypothesis 3 but still more inclined towards the hypothesis Men are more family person , which was unexpected. (opposite of the hypothesis)
Career	1.356091	Men are more exposed to career than Women.
Science	0.924645	Women are inclined towards science field.

Observations

- There are some domains where the news channel seems to be biased like the ‘intelligence’ field , ‘career’ field , ‘Math’.
- Interesting thing to observe is that strength related terms are also used and also weakness represented terms are used which indicates that some articles empower women but some just try to weakenise women by using weakness terms.
- Another possibility is that it is just noise in the data as the data set is very small.

Conclusions

- We conclude that Gender Bias is Present in most of the domains but many domains are still Neutral or unbiased in the ABP news Data .

- Many Hypothesised Domains which were considered to be Bias towards one Gender were found to Neutral .

Error Analysis

We analysed the errors in this system using 2 methods

1. Using Gender Neutral Words
2. Using another Word Embedding Model Trained on a larger Corpus and Known to be Gender Unbiased .

Using Gender Neutral Words

We use Gender neutral terms and compare their euclidian distance from average man vector - average women Vector .

Ideally it should be same but this would show us the error .

Using INLTK embeddings

We would chose a set of words to compare the embeddings between our embeddings and INLTK embedding and their comparison would give us error in our analysis assuming that INLTK is unbiased .

Upgradation …

- This Analysis could be extended to all the Indian Languages with the information of inflection of different languages.

- Various news channels, blogs can be checked for bias using this model.

- Removal of Gender Bias using a suggestion model to change certain words , a system which prescribe the correct vocabulary.

धन्यवाद !!