Seminar Report on “Word Embedding in Natural language Processing”

1. Date of the Seminar/Workshop: 13-04-2022

2. Title of the Seminar/Workshop: “Word Embedding in Natural language Processing”

3. Name of the Speaker/Resource person with Affiliation with the institute/industry:

Mr. Cletus Agnello Dsouza

Industry: - Mastek

Designation: - Software Engineer

4. Venue of the Seminar/Workshop: Platform used-Google meet

5. Duration of the Seminar: 1 hour (9.00 am to 10.00 am)

6. Conducted For: Students of final year computer Engineering

7. Objective of the Seminar/Workshop /Curriculum Gap identified/Other than that:

The objective of the Seminar was basically to ensure that the students get acquainted to new methodologies of word embedding techniques so that they can use this information for developing projects in NLP. This workshop was conducted to draw the attention of all towards the challenges in word embedding methodologies and use some of them in extracting features from text data so that they can be represented in vector form and be fed into any machine learning training algorithms for extracting information from textual data.

8. Contents of the Seminar/Workshop:

The one hour period was very interactive and aroused key enthusiasm among students.
The speaker started with

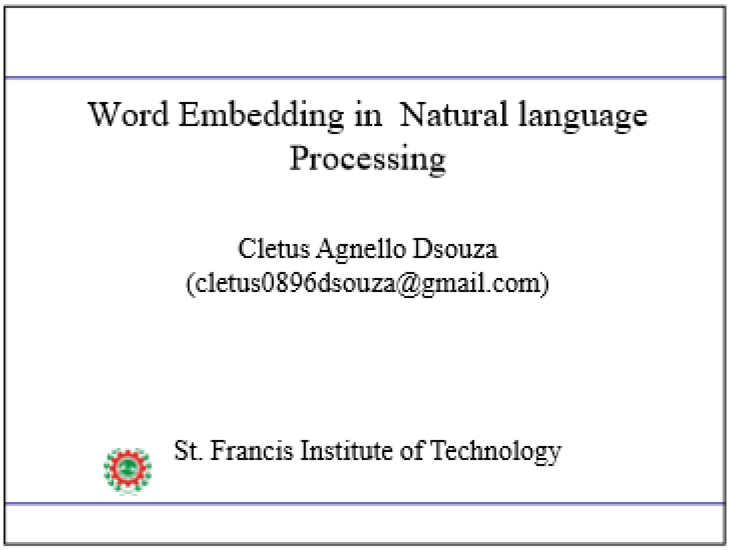
- A layman’s perspective on personal and professional ethics
- Why am I (Dr Vikram V. Shete) qualified to give this seminar?
- What is Ethics?
- Why do I need any ethic?
- Miracle in the Andes @3750 meters on Oct 13th, 1972
- Ethical Dilemma
- A quick survey on Ethical Dilemma
- Paradigms of Ethical Dilemma
- Professional Ethics
- Professional Ethical Dilemmas
- Closing thoughts

9. Description of the Entire Event

The Computer Department at St. Francis Institute of Technology hosted a seminar on “Word Embedding in NLP” on Monday, April 13 , 2022, from 9.00 am to 10.00 am, online via Google meet.

There were approximately 100 student participants and 2 faculty participants who attended the seminar. The opening remark was given by Ms. Vincy Joseph where she introduced the Speaker, Mr. Cletus Dsouza.

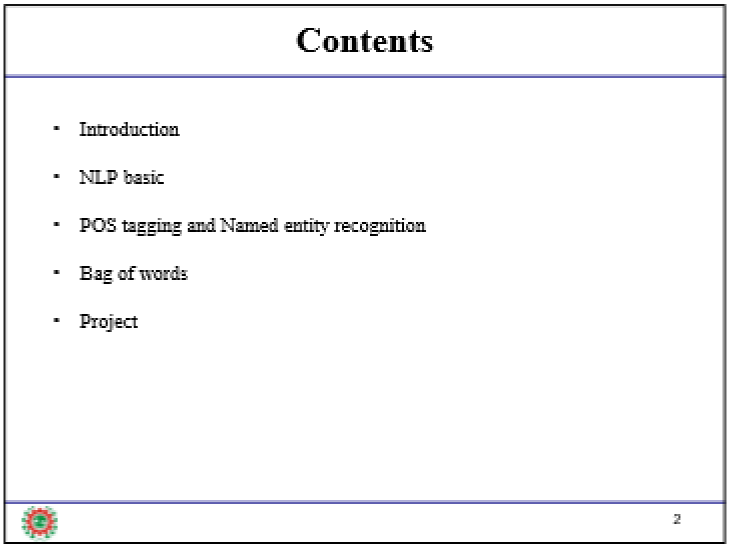
Further, the speaker took over the session and started off the session. The summary of the session can be interpreted from the slides given below:



Word Embedding in Natural language Processing

Cletus Agnello Dsouza
(cletus0896dsouza@gmail.com)

St. Francis Institute of Technology



Contents

- Introduction
- NLP basic
- POS tagging and Named entity recognition
- Bag of words
- Project

2

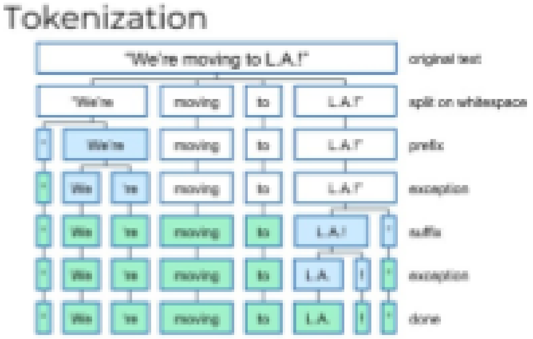
Introduction

- NLP is an area of computer science and artificial intelligence concerned with interaction between computer and human languages , in particular how to program computers to process and analyze large amount of language data
- For Example Movie Reviews , product reviews etc

NLP Basic

Tokenization

- Tokenization is the process of breaking up the original text into component pieces(tokens).
- Tokens helps us understand the meaning of the text and their relationship with one another



Tokenization

"We're moving to L.A.!" original text

"We're | moving | to | L.A.!" split on whitespace

" | We're | moving | to | L.A.!" prefix

" | We | 're | moving | to | L.A.!" exception

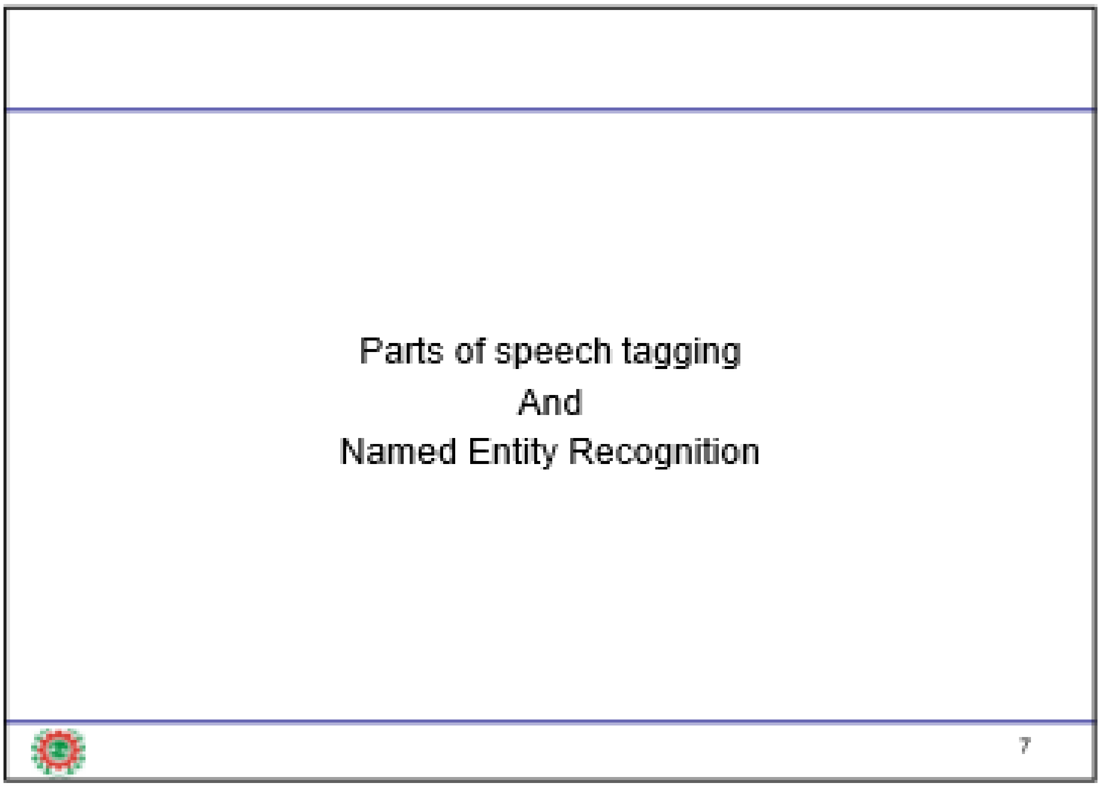
" | We | 're | moving | to | L.A.! | " suffix

" | We | 're | moving | to | L.A. | ! | " exception

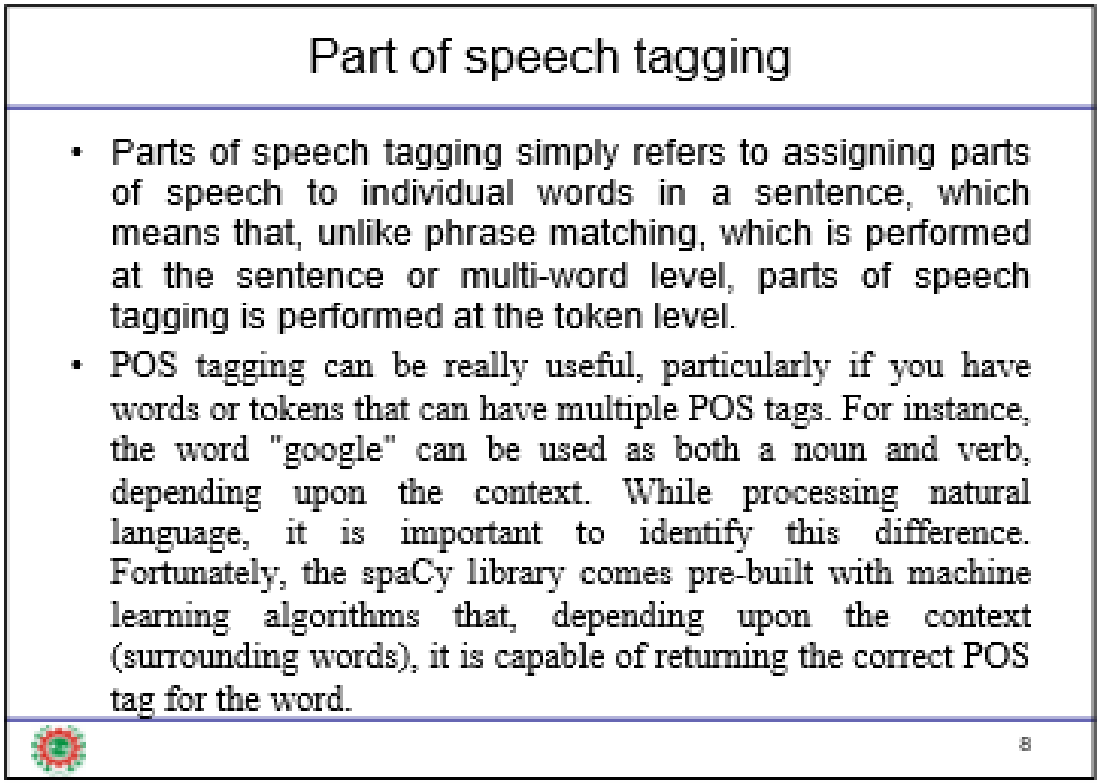
" | We | 're | moving | to | L.A. | ! | " done

Stemming and Lemmatization

- Stemming is a technique used to extract the base form of the words by removing affixes from them. It is just like cutting down the branches of a tree to its stems. For example, the stem of the words ***eating, eats, eaten*** is ***eat***.
- Lemmatization technique is like stemming. The output we will get after lemmatization is called 'lemma', which is a root word rather than root stem, the output of stemming. After lemmatization, we will be getting a valid word that means the same thing.



Parts of speech tagging
And
Named Entity Recognition



Part of speech tagging

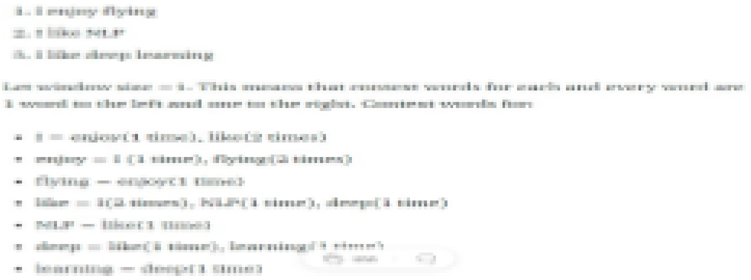
- Parts of speech tagging simply refers to assigning parts of speech to individual words in a sentence, which means that, unlike phrase matching, which is performed at the sentence or multi-word level, parts of speech tagging is performed at the token level.
- POS tagging can be really useful, particularly if you have words or tokens that can have multiple POS tags. For instance, the word "google" can be used as both a noun and verb, depending upon the context. While processing natural language, it is important to identify this difference. Fortunately, the spaCy library comes pre-built with machine learning algorithms that, depending upon the context (surrounding words), it is capable of returning the correct POS tag for the word.

Named Entity Recognition

- Named entity recognition refers to the identification of words in a sentence as an entity e.g. the name of a person, place, organization, etc. Let's see how the spaCy library performs named entity recognition

- Word Embedding Module

- Many Machine Learning algorithms and almost all Deep Learning Architectures are incapable of processing *strings* or *plain text* in their raw form. They require numbers as inputs to perform any sort of job, be it classification, regression, etc. in broad terms.
- Lets take an example

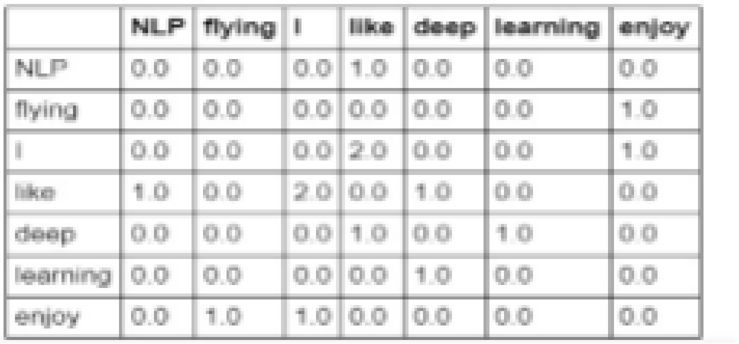


1. I enjoy flying
2. I like NLP
3. I like deep learning

Let window size = 1. This means that context words for each and every word are 1 word to the left and one to the right. Context words for:

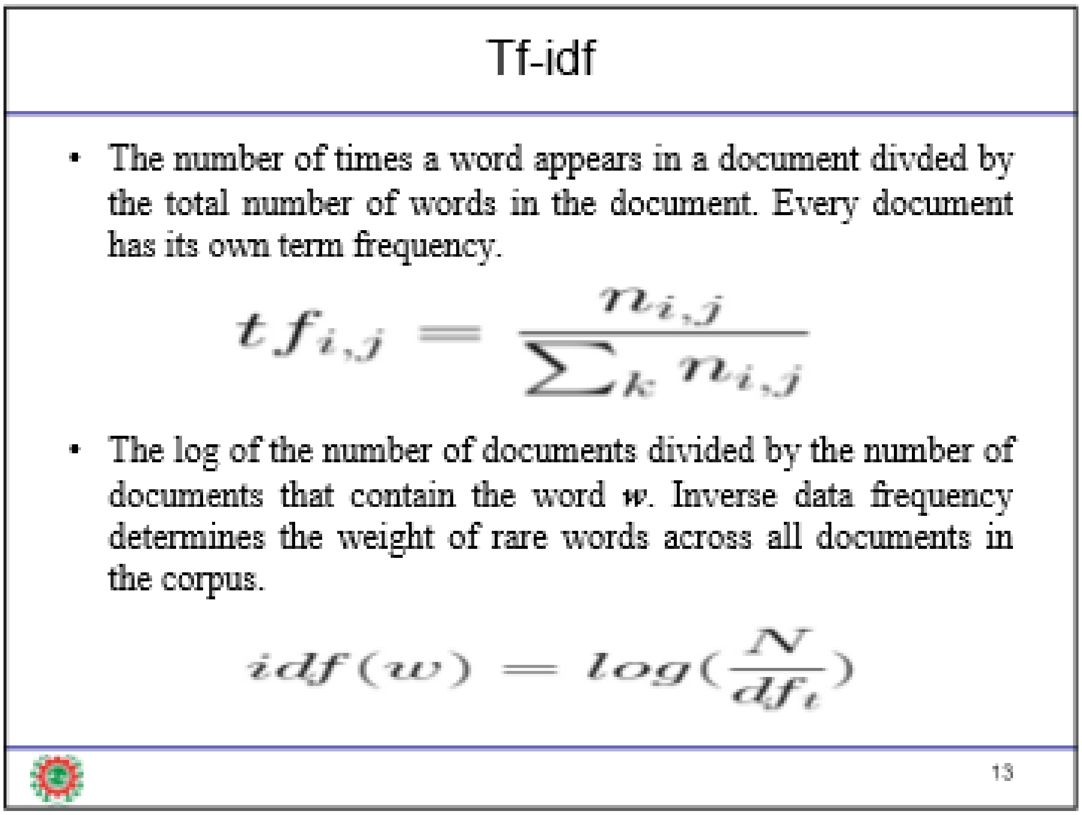
- I = enjoy(1 time), like(2 times)
- enjoy = I (1 time), flying(2 times)
- flying — enjoy(1 time)
- like = I(2 times), NLP(1 time), deep(1 time)
- NLP — like(1 time)
- deep = like(1 time), learning(1 time)
- learning — deep(1 time)

Co-occurrence



	NLP	flying	I	like	deep	learning	enjoy
NLP	0.0	0.0	0.0	1.0	0.0	0.0	0.0
flying	0.0	0.0	0.0	0.0	0.0	0.0	1.0
I	0.0	0.0	0.0	2.0	0.0	0.0	1.0
like	1.0	0.0	2.0	0.0	1.0	0.0	0.0
deep	0.0	0.0	0.0	1.0	0.0	1.0	0.0
learning	0.0	0.0	0.0	0.0	1.0	0.0	0.0
enjoy	0.0	1.0	1.0	0.0	0.0	0.0	0.0

Co-occurrence matrix A



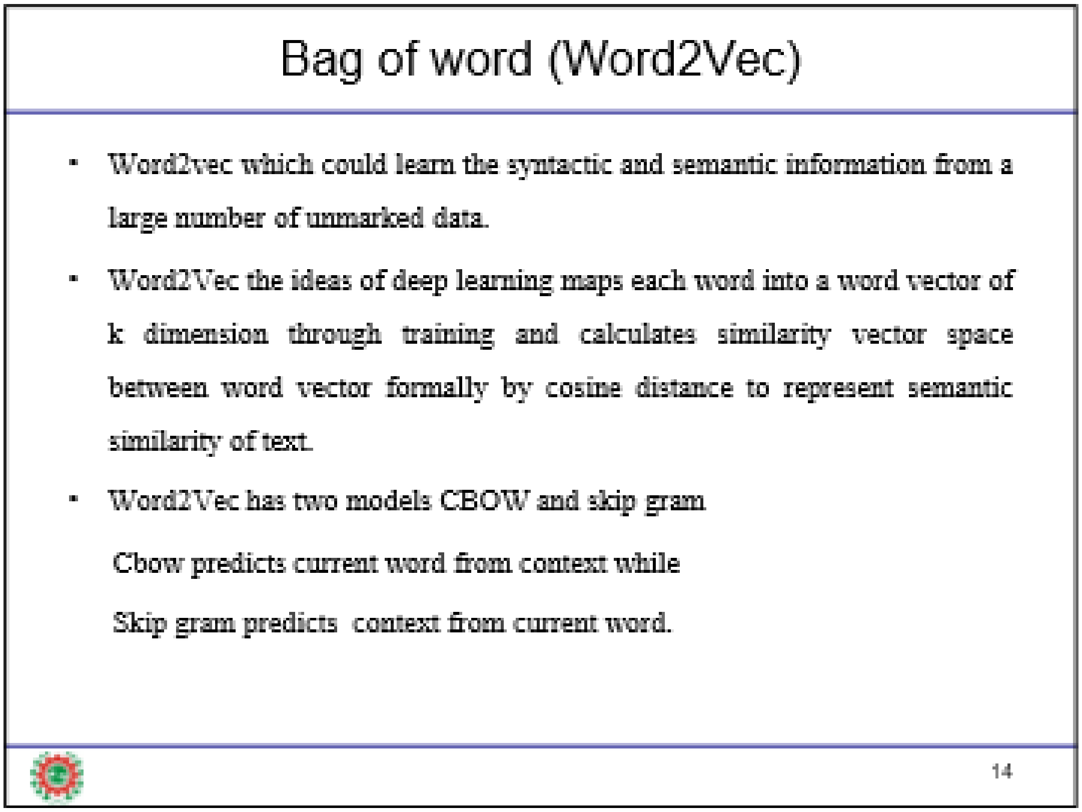
Tf-idf

- The number of times a word appears in a document divded by the total number of words in the document. Every document has its own term frequency.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

- The log of the number of documents divided by the number of documents that contain the word *w*. Inverse data frequency determines the weight of rare words across all documents in the corpus.

$$idf(w) = log(\frac{N}{df_t})$$



Bag of word (Word2Vec)

- Word2vec which could learn the syntactic and semantic information from a large number of unmarked data.
- Word2Vec the ideas of deep learning maps each word into a word vector of k dimension through training and calculates similarity vector space between word vector formally by cosine distance to represent semantic similarity of text.
- Word2Vec has two models CBOW and skip gram

Cbow predicts current word from context while

Skip gram predicts context from current word.

Word2Vec Continue

- Word2Vec treats text set as input, and generates the corresponding words vector quickly and efficiently through the training .Because of the word vector capturing semantic characteristics between words in a natural language.

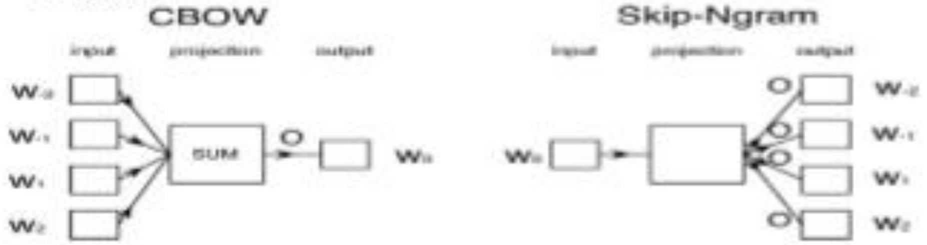


Fig No 1 : word2vec diagram

Word2Vec Cont….

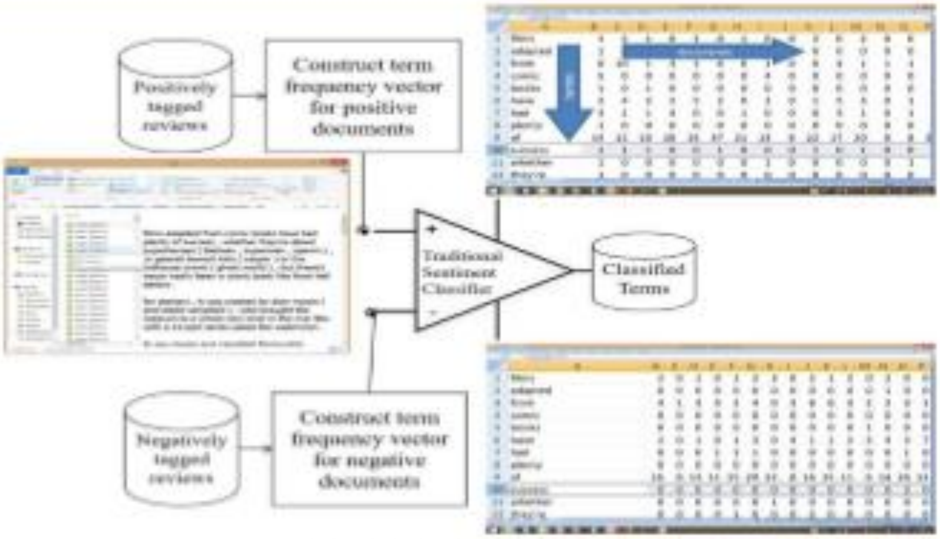
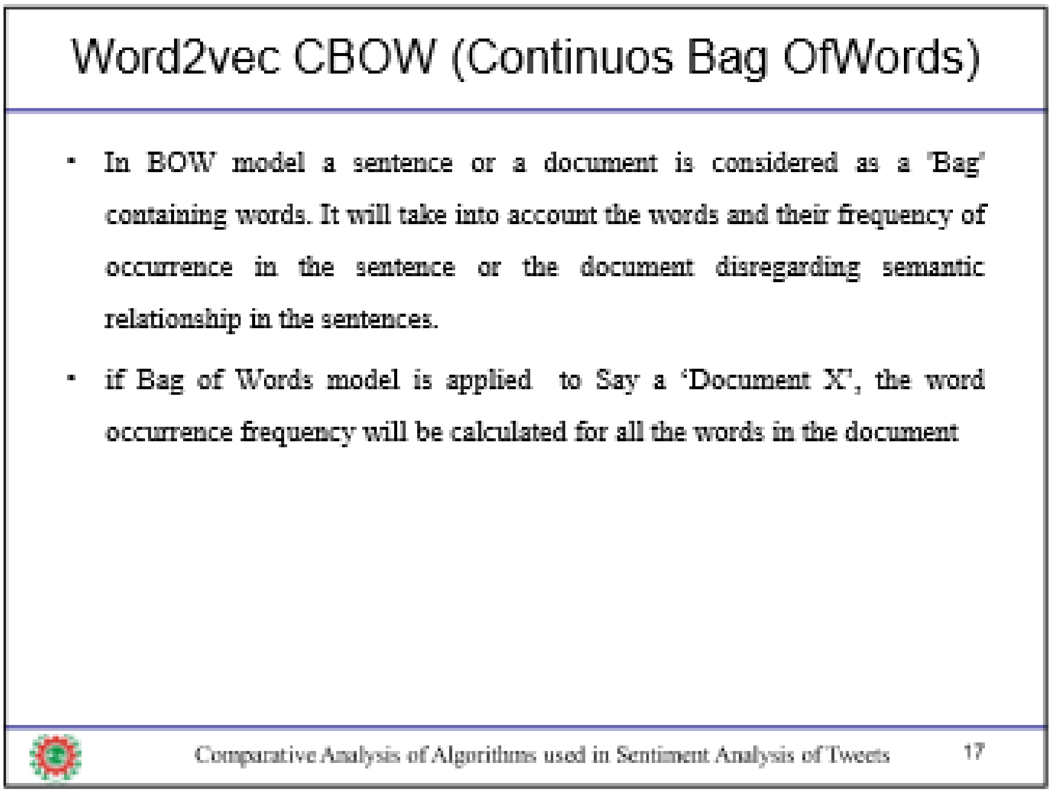
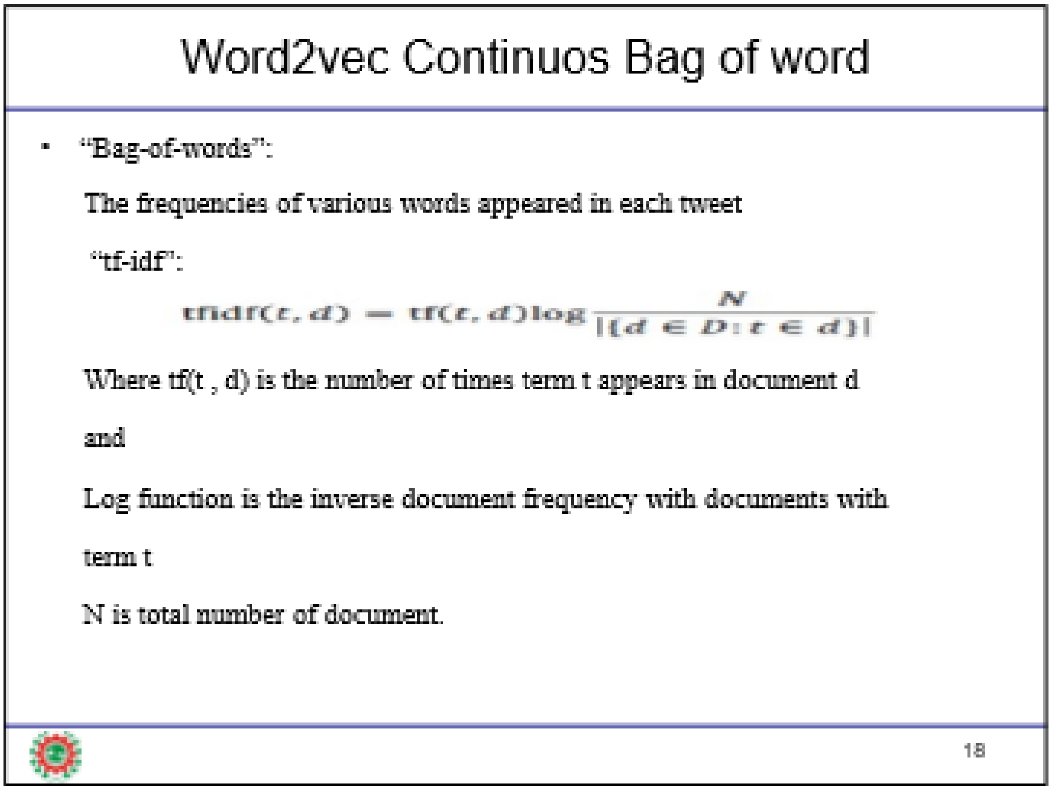

Positively tagged reviews
Construct term frequency vector for positive documents
+
Traditional Sentiment Classifier
-
Classified Terms
Negatively tagged reviews
Construct term frequency vector for negative documents

Fig no 2 : word2vec example



Word2vec CBOW (Continuos Bag OfWords)

- In BOW model a sentence or a document is considered as a 'Bag' containing words. It will take into account the words and their frequency of occurrence in the sentence or the document disregarding semantic relationship in the sentences.
- if Bag of Words model is applied to Say a 'Document X', the word occurrence frequency will be calculated for all the words in the document



Word2vec Continuos Bag of word

- "Bag-of-words":

  The frequencies of various words appeared in each tweet

  "tf-idf":

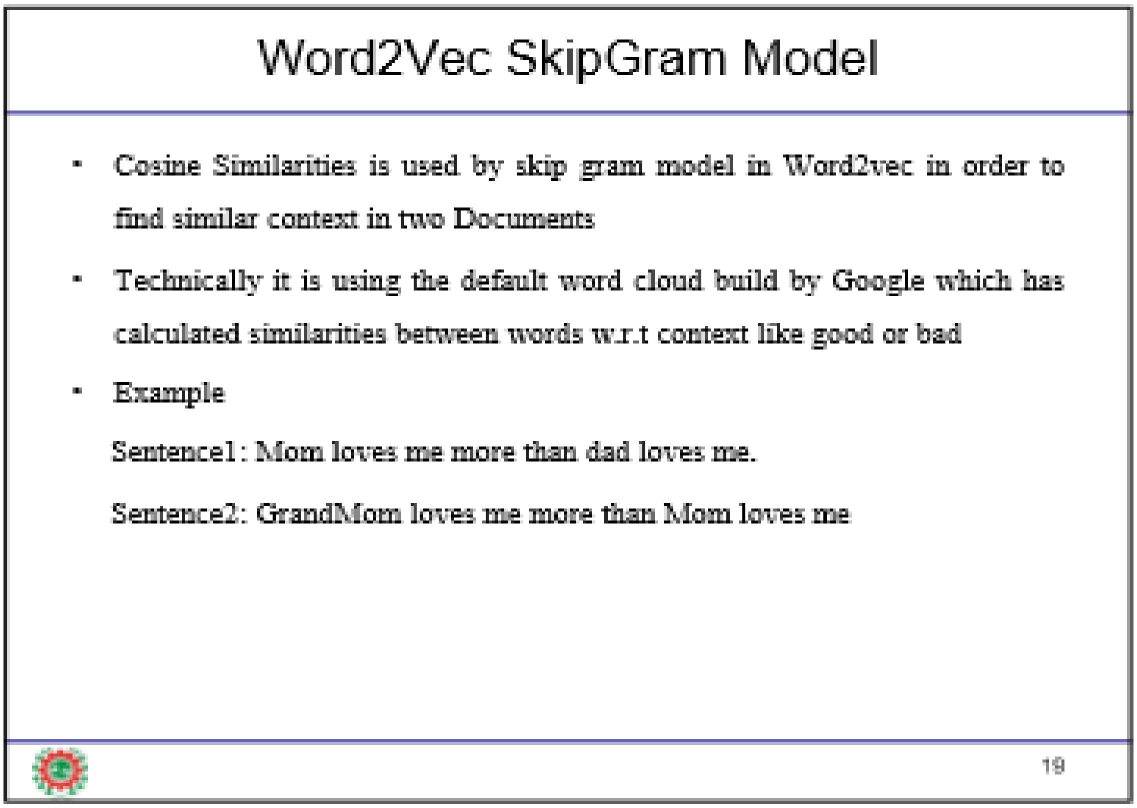
$$\mathrm{tfidf}(t, d) = \mathrm{tf}(t, d) \log \frac{N}{|\{d \in D : t \in d\}|}$$

Where tf(t , d) is the number of times term t appears in document d

and

Log function is the inverse document frequency with documents with term t

N is total number of document.

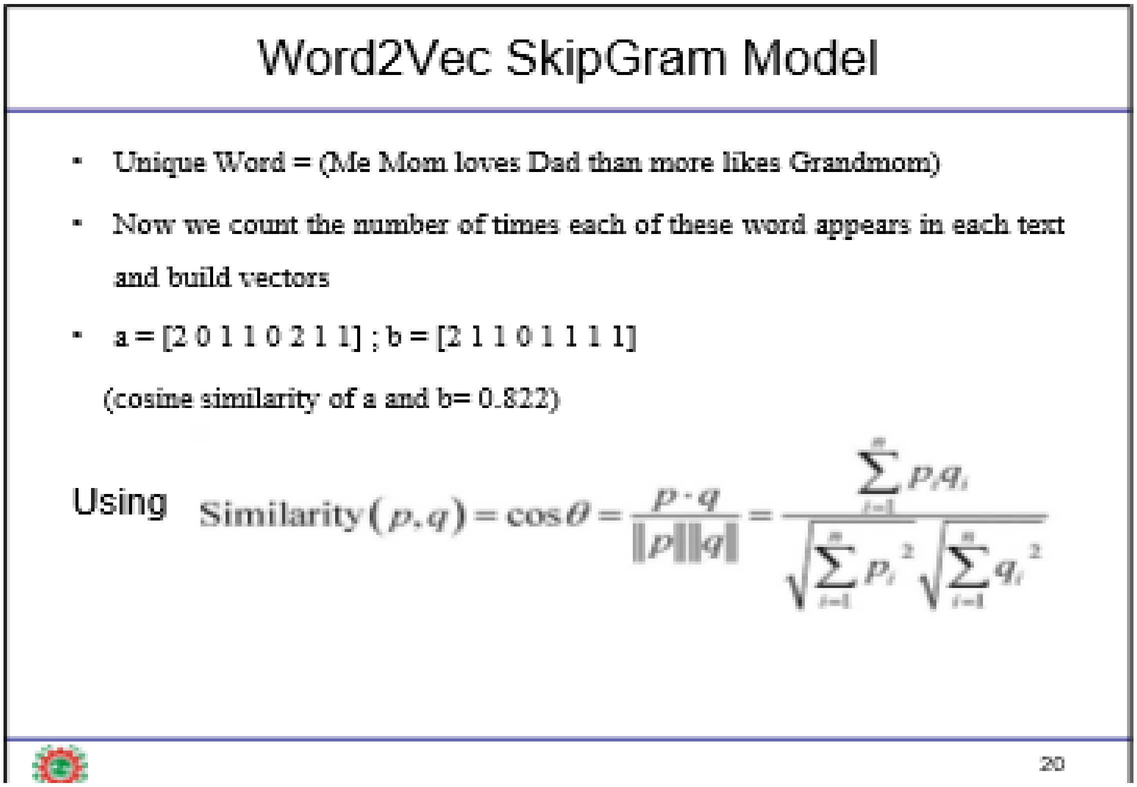


Word2Vec SkipGram Model

- Cosine Similarities is used by skip gram model in Word2vec in order to find similar context in two Documents
- Technically it is using the default word cloud build by Google which has calculated similarities between words w.r.t context like good or bad
- Example

  Sentence1: Mom loves me more than dad loves me.

  Sentence2: GrandMom loves me more than Mom loves me



Word2Vec SkipGram Model

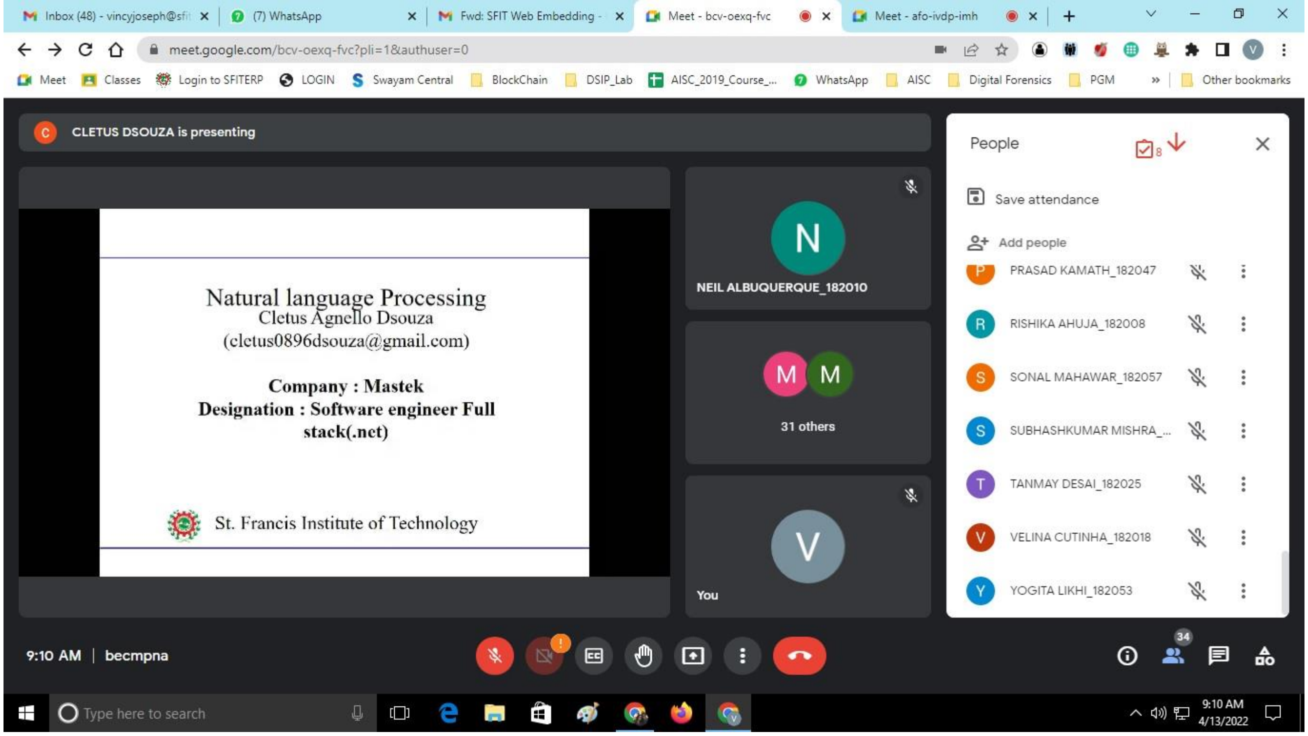
- Unique Word = (Me Mom loves Dad than more likes Grandmom)
- Now we count the number of times each of these word appears in each text and build vectors
- a = [2 0 1 1 0 2 1 1] ; b = [2 1 1 0 1 1 1 1]

  (cosine similarity of a and b= 0.822)

Using $$\text{Similarity}(p,q) = \cos\theta = \frac{p \cdot q}{\|p\|\|q\|} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}}$$

At the end of the session, everyone turned on their cameras and took a virtual group photo. Ms. Vincy Joseph gave our speaker a vote of gratitude and useful input on this information-packed webinar, and the webinar was ended on this note. The students had a better understanding of word embedding a result of this session.

Session Photographs

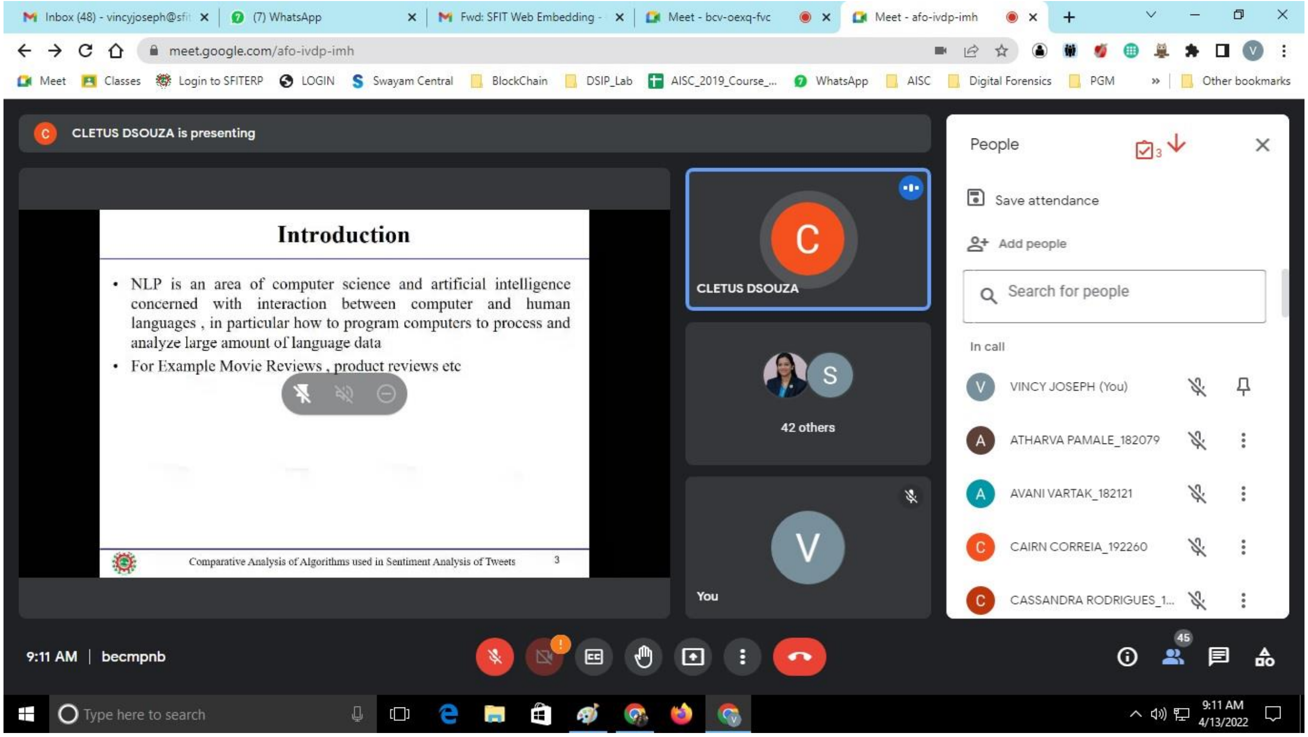


Natural language Processing
Cletus Agnello Dsouza
(cletus0896dsouza@gmail.com)

Company : Mastek
Designation : Software engineer Full stack(.net)

St. Francis Institute of Technology

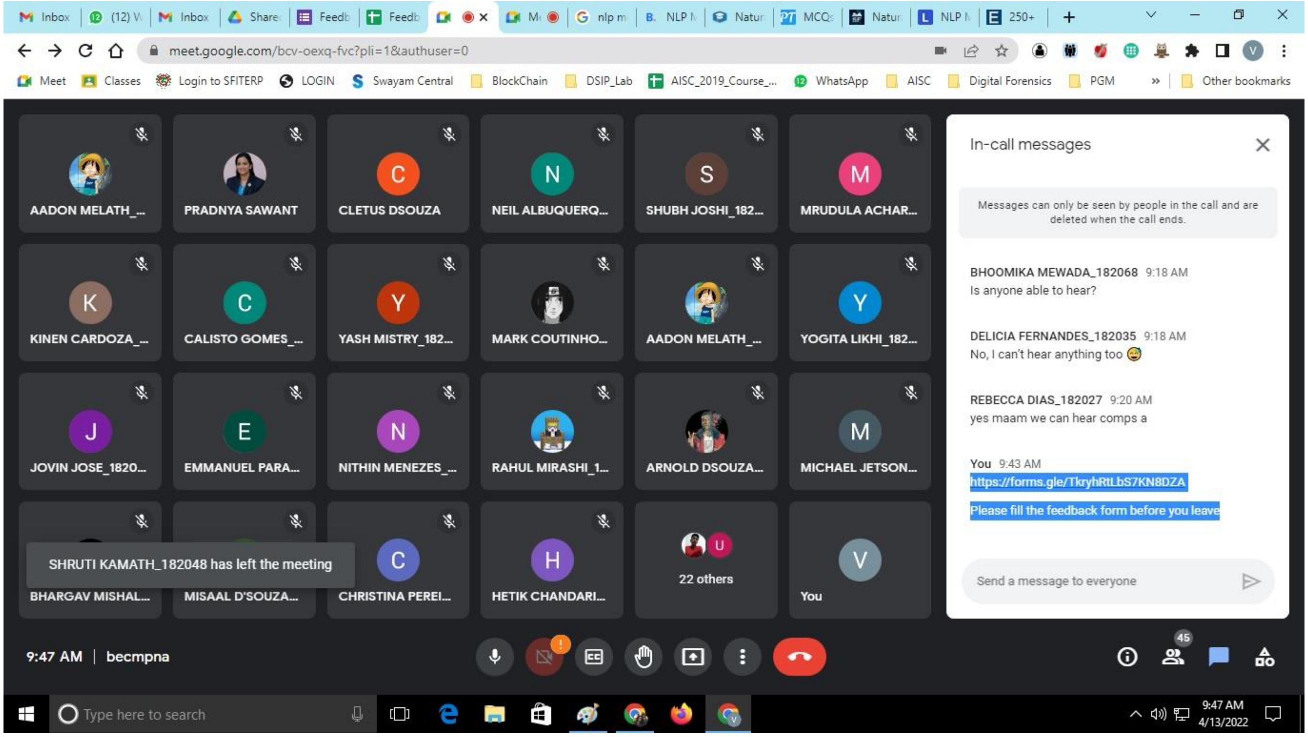
PRASAD KAMATH_182047
RISHIKA AHUJA_182008
SONAL MAHAWAR_182057
SUBHASHKUMAR MISHRA_...
TANMAY DESAI_182025
VELINA CUTINHA_182018
YOGITA LIKHI_182053



Introduction

- NLP is an area of computer science and artificial intelligence concerned with interaction between computer and human languages , in particular how to program computers to process and analyze large amount of language data
- For Example Movie Reviews , product reviews etc

Comparative Analysis of Algorithms used in Sentiment Analysis of Tweets 3



In-call messages

Messages can only be seen by people in the call and are deleted when the call ends.

BHOOMIKA MEWADA_182068 9:18 AM
Is anyone able to hear?

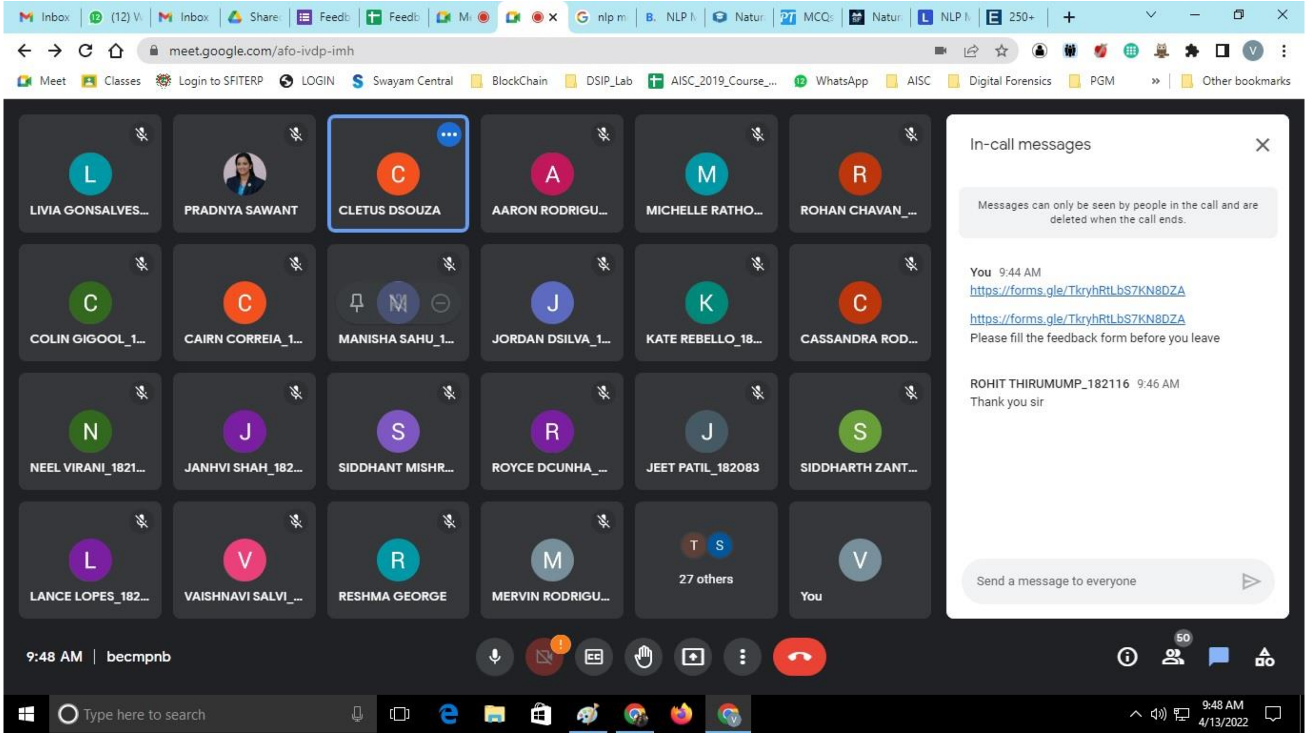
DELICIA FERNANDES_182035 9:18 AM
No, I can't hear anything too 😅

REBECCA DIAS_182027 9:20 AM
yes maam we can hear comps a

You 9:43 AM
https://forms.gle/TkryhRtLbS7KN8DZA

Please fill the feedback form before you leave

Send a message to everyone



In-call messages

Messages can only be seen by people in the call and are deleted when the call ends.

You 9:44 AM
https://forms.gle/TkryhRtLbS7KN8DZA

https://forms.gle/TkryhRtLbS7KN8DZA
Please fill the feedback form before you leave

ROHIT THIRUMUMP_182116 9:46 AM
Thank you sir

Send a message to everyone

Ms. Vincy Joseph and Ms. Pradnya Rane Sawant. **Dr. Kavita Sonawane**

Seminar Incharge **HOD, CMPN**