MENTOR: PROF. DR. JÜRGEN LERNER

ADITYA KRISHN

MASTER PROJECT: MID TERM PRESENTATION

ANALYSIS OF THE ENRON EMAIL DATA
NLP OF EMAIL CONTENT

TABLE OF CONTENTS

COMPANY HISTORY

ENRON

Enron Corporation was an American energy, commodities, and services company based in Houston, Texas. Fortune named Enron "America's Most Innovative Company" for six consecutive years.

Manipulation

special purpose vehicles (SPVs)
Mark-to-Market: Skilling

Key People

Kenneth Lay: Chairman
Jeffrey Skilling: COO/ CEO
Andrew Fastow: CFO

Some Numbers

Share price: 90.75 to 0.26
Staff(2001): 29,000
Revenue: 101 Billion

The Timeline

1985	Enron is formed following a merger between Houston Natural Gas Co. and InterNorth Inc.
1995	Enron is named "America's Most Innovative Company" by Fortune. The firm goes on to win this award for six consecutive years.
1998	Andrew Fastow is promoted to CFO, he ultimately spearheads the creation of a network of companies that hide Enron's losses.
2000	Enron's shares skyrocket to an all-time high of $90.56.
Feb. 12, 2001	Jeffrey Skilling replaces Kenneth Lay as CEO. However, Lay remains a member of the board of directors.
Aug. 14, 2001	Skilling resigns suddenly, and Lay takes over once again. Enron's broadband division also reports a massive $137 million loss. Analysts became weary of the company and subsequently drop their ratings for Enron's stock. In turn, the company's share price dives to $39.95, a 52-week low.
Oct. 12, 2001	Arthur Andersen legal counsel tells auditors to destroy all Enron files, except Enron's most basic documents.
Oct. 16, 2001	Enron reports a $618 million loss and $1.2 billion value write off. Enron's stock drops further to $38.84.
Oct. 22, 2001	Enron announces it's facing a SEC probe. Shares fall to around $20.75 that day, following the announcement.
Nov. 8, 2001	Enron admits it has been inflating its income by around $586 million since 1997.
Nov. 29, 2001	Arthur Andersen becomes another casualty of the Enron scandal as the SEC expands its investigation.
Dec. 2, 2001	Enron files for Chapter 11 bankruptcy. Its stock closes at $0.26
Jan. 9, 2002	The Justice Department launches a criminal investigation.
Jan. 15, 2002	Enron is suspended from the NYSE.
June 15, 2002	Enron's accounting firm, Arthur Andersen is convicted of obstructing justice.

Source:
https://www.investopedia.com/updates/enron-scandal-summary/

Before and After

Identifying Difference in Behaviour before and after scandal

Topic Modelling

Sentiment and Emotion Analysis

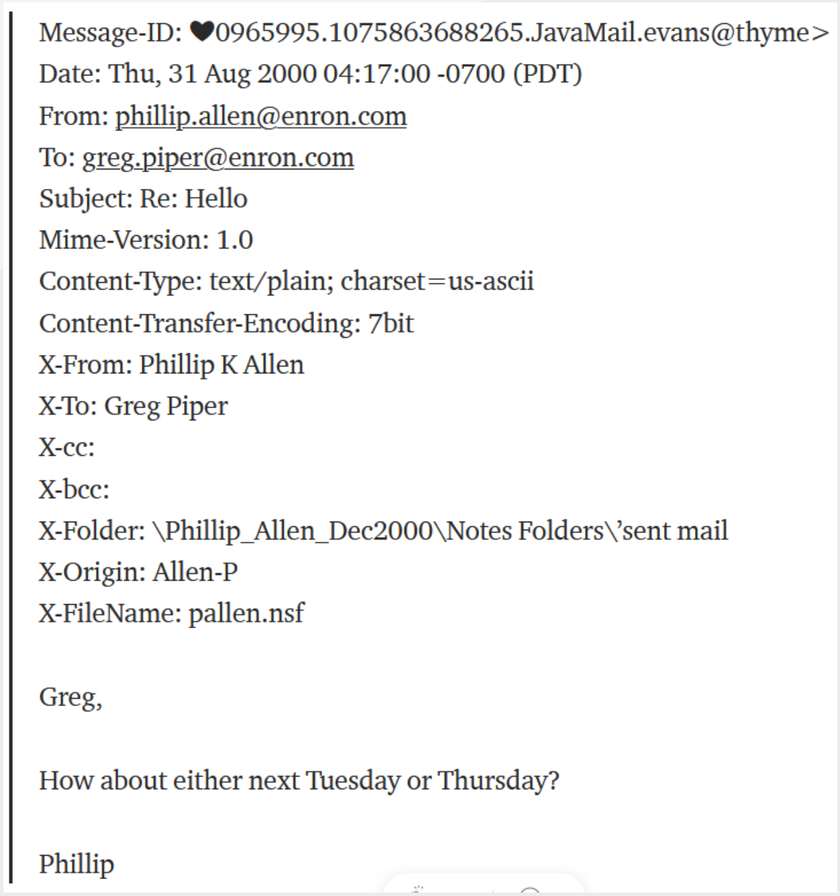
OBJECTIVE

Ethical Aspect

Detection of Moral and Blame Language

DATA DESCRIPTIVES

- Email Structure
- Summarized Communication Data
- Monetary and POI Information



Message-ID: ♥0965995.1075863688265.JavaMail.evans@thyme>
Date: Thu, 31 Aug 2000 04:17:00 -0700 (PDT)
From: phillip.allen@enron.com
To: greg.piper@enron.com
Subject: Re: Hello
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Phillip K Allen
X-To: Greg Piper
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Dec2000\Notes Folders\'sent mail
X-Origin: Allen-P
X-FileName: pallen.nsf

Greg,

How about either next Tuesday or Thursday?

Phillip

APPROACH FOR BEHAVIOURAL DIFFERENCE

Latent Dirichlet Allocation (LDA)

Fit topic model word collection with document word collection

NMF

Non-Negative Matrix Factorization

Supervised

Find labelled dataset of emails annotated with sentiment score

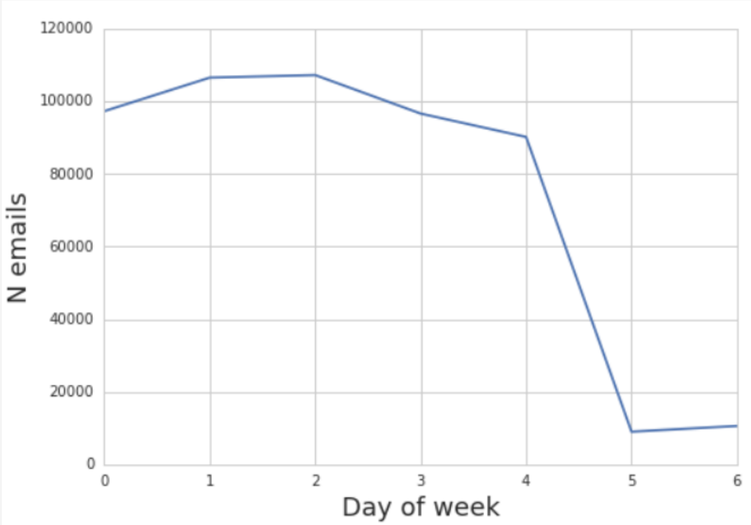
Word Embedding

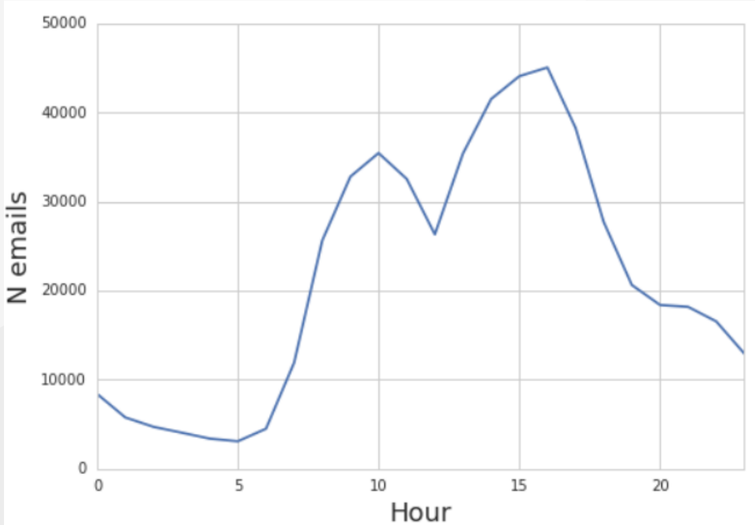
Convert text to embeddings and calculate similarity to positive and negative sets

Transfer Learning

Using pre-existing libraries like StanfordCoreNLP, pattern or text-blob

EXPLORATORY ANALYSIS

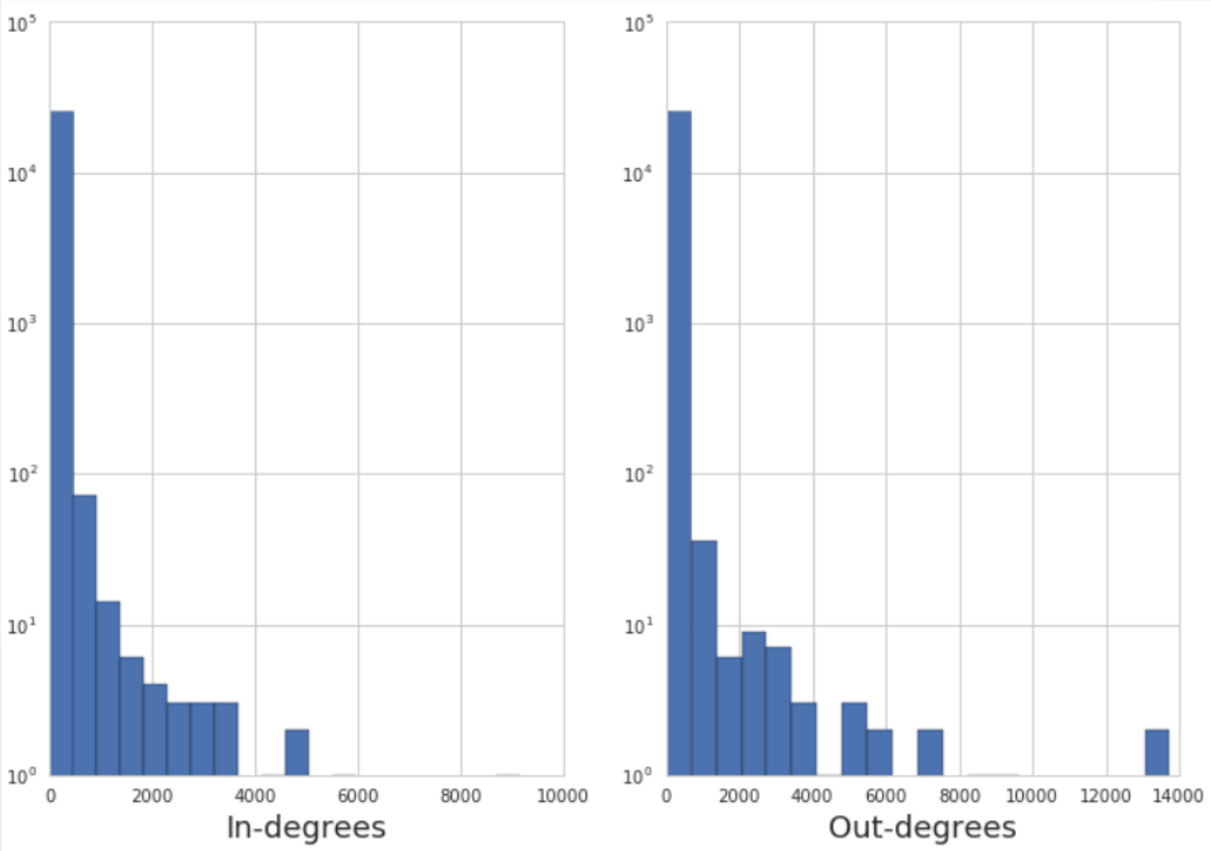




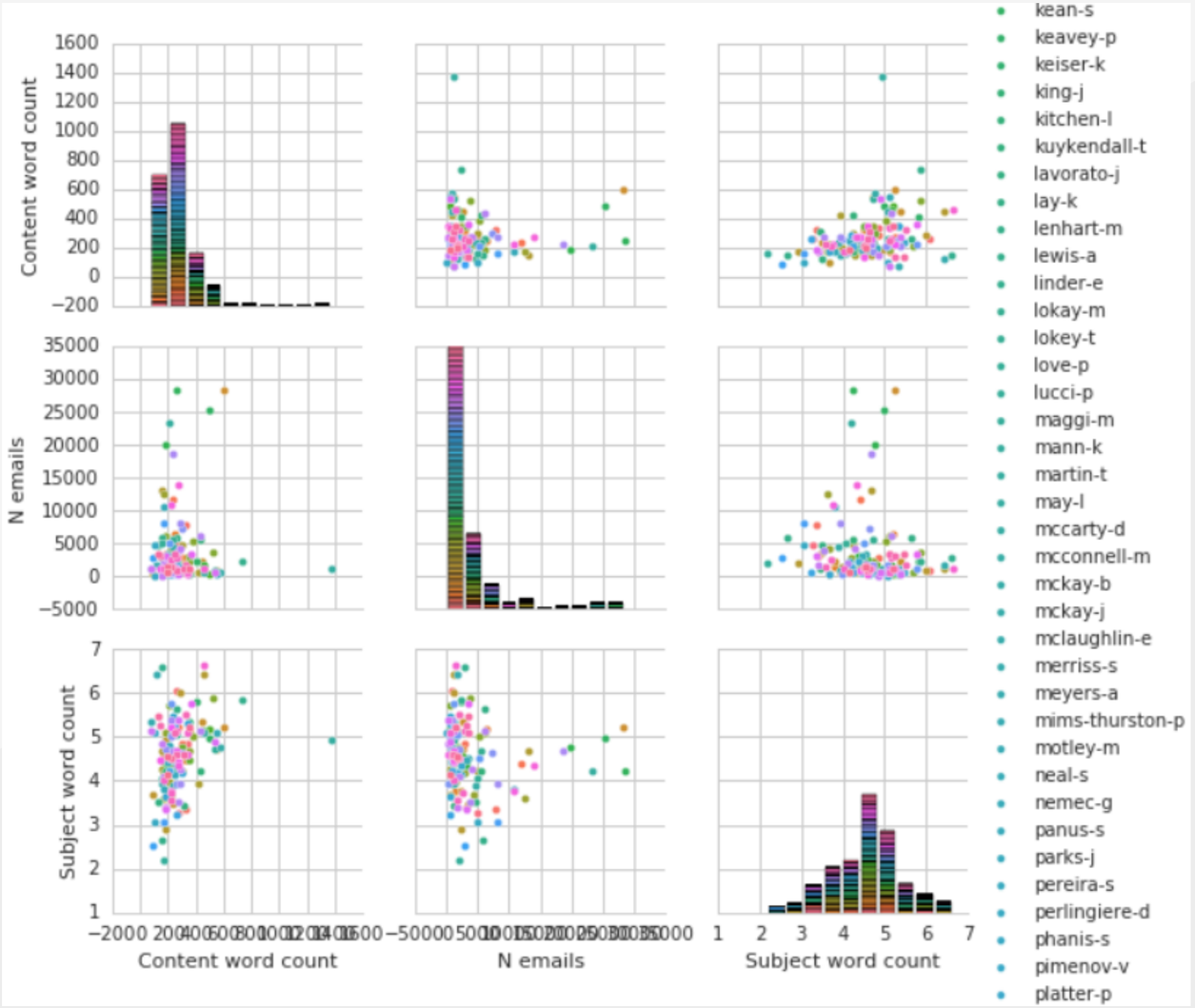
	Content word count	N emails	Subject word count
user			
kaminski-v	255.906025	28465	4.217530
dasovich-j	603.393391	28234	5.237373
kean-s	490.837561	25351	4.959331
mann-k	207.195501	23381	4.211796
jones-t	185.462607	19950	4.748221

	From	To	count
17908	pete.davis@enron.com	pete.davis@enron.com	9141
38033	vince.kaminski@enron.com	vkaminski@aol.com	4308
28920	enron.announcements@enron.com	all.worldwide@enron.com	2206
28935	enron.announcements@enron.com	all.houston@enron.com	1701
26510	kay.mann@enron.com	suzanne.adams@enron.com	1528
38031	vince.kaminski@enron.com	shirley.crenshaw@enron.com	1190
14564	steven.kean@enron.com	maureen.mcvicker@enron.com	1014
26309	kay.mann@enron.com	nmann@erac.com	980
18926	kate.symes@enron.com	evelyn.metoyer@enron.com	915
18930	kate.symes@enron.com	kerri.thompson@enron.com	859

EXPLORATORY ANALYSIS



EXPLORATORY ANALYSIS



SPEED UP PREPROCESSING

What is Modin?

Modin is a drop-in replacement for pandas. While pandas is single-threaded, Modin lets you instantly speed up your workflows by scaling pandas so it uses all of your cores. Modin works especially well on larger datasets, where pandas becomes painfully slow or runs out of memory.

By simply replacing the import statement, Modin offers users effortless speed and scale for their pandas workflows:

```
import pandas as pd
import modin.pandas as pd
```

In the GIFs below, Modin (left) and pandas (right) perform *the same pandas operations* on a 2GB dataset. The only difference between the two notebook examples is the import statement.

MODIN

pandas

EVALUATION STRATEGY

- For sentiment analysis hand-labelled data will be used and tested against(Accuracy of more than 80% is targetted)
- For topic modelling analysis of both a level-A executive and a low level employee will be done(level-A executive will have words such as bankrupcy and litigation around the date of scandal and low level will have commonly used corporate lingo and department specific words
- Email content around and after scandal will have presence of moral language for a low-level employee

SENTIMENT ANALYSIS CLASSIFICATION

Sentiment Analysis of Kay Mann: kay.mann@enron.com (Head of Legal)

- VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.
- The VADER sentiment lexicon is sensitive to both the polarity and the intensity of sentiments expressed in social media contexts
- The heuristics of VADER go beyond what would normally be captured in a typical bag-of-words model. They incorporate word-order sensitive relationships between terms.
- VADER belongs to a kind of sentiment analysis that depends on lexicons of sentiment-related words. In this methodology, every one of the words in the vocabulary is appraised with respect to whether it is positive or negative, and, how +ve or -ve.
- Eg.: (tragedy,-3.4), (rejoiced, 2.0), difference between extremely good and marginally good

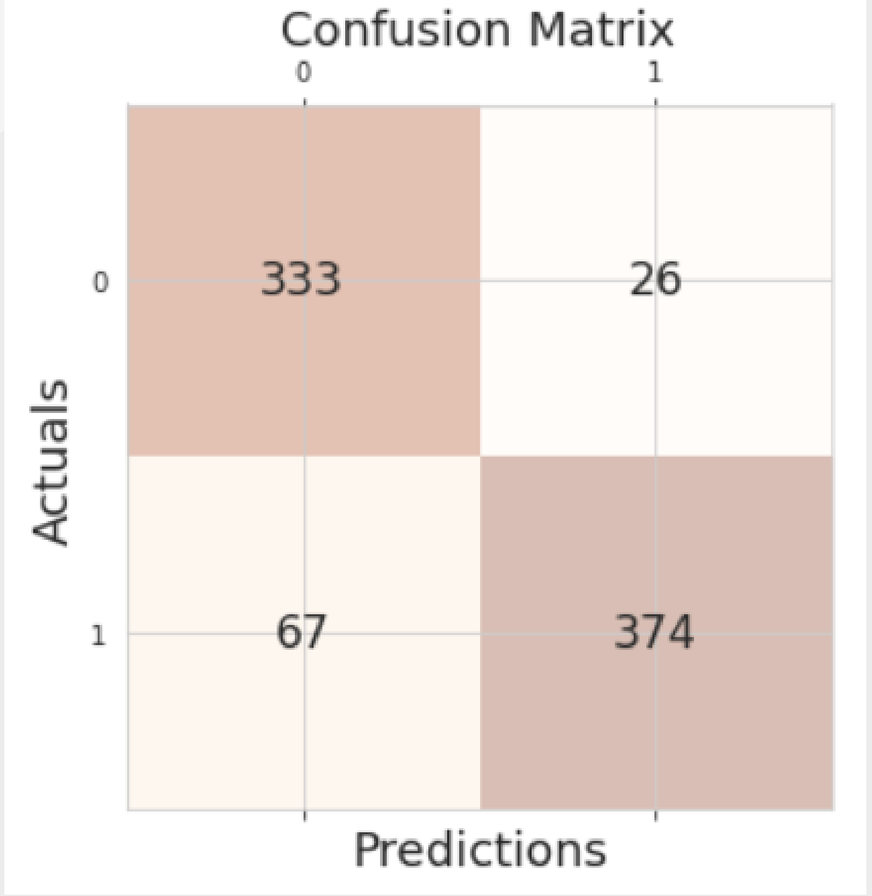
content_clean	scores	compound	comp_score	comp_score_user
accomplishment jumped window yet slapped cursed anyone enron entity significant way would understand vision value significant accomplishment lm deal revising structure figuring sell westlb stuff ge closing relatively short period time come vacation scatter day plan consecutive day kay	{'neg': 0.0, 'neu': 0.814, 'pos': 0.186, 'compound': 0.7184}	0.72	pos	pos
advice specific scope guy run bill worse anyone carlos sole enron development kay mann corp enron enron illinois project update briefly glanced pmrw update illinois project jumped indication working andrew kurth due diligence livingston kendall project presume mistake part reference king spalding get started due diligence u waiting get fax going send morningiand point correction presuming understanding correct	{'neg': 0.091, 'neu': 0.909, 'pos': 0.0, 'compound': -0.6705}	-2.01	neg	neg
afraid say anything anything said night gee anything right guess little upset comment thing even try fold clothes pack car unpack grocery load dishwasher etc say something something get upset say anything right feel guilty saying anything anything gotta go talk later want love kay	{'neg': 0.14, 'neu': 0.756, 'pos': 0.104, 'compound': -0.2297}	-0.69	neg	neg
al project must break contract deal struck ge lender change order master contract relating specific project break contract must entered soon practicable appproach taken austin lv co gen esa contract	{'neg': 0.011, 'neu': 0.816, 'pos': 0.173, 'compound': 0.9783}	0.98	pos	pos
Total		**201.60**		

SENTIMENT ANALYSIS CLASSIFICATION

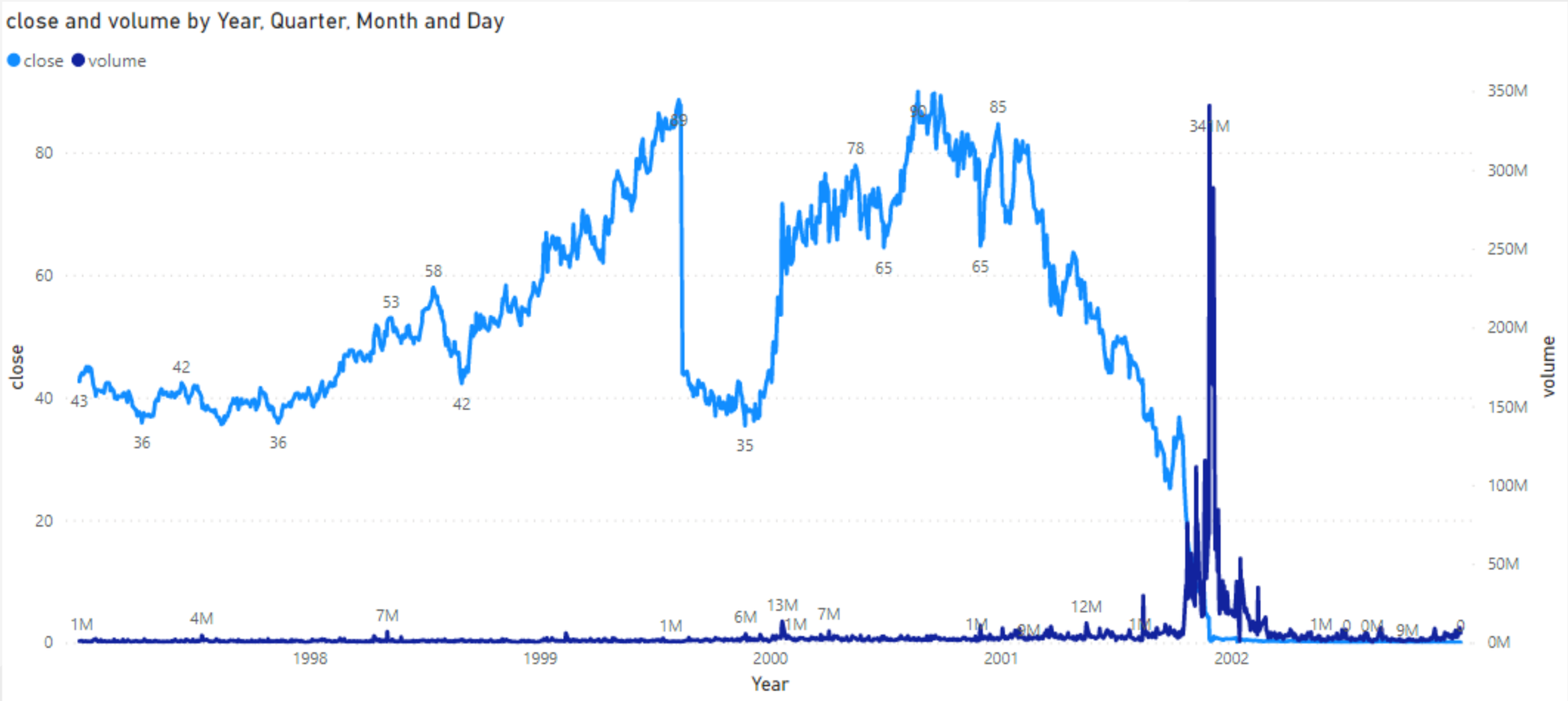
Sentiment Analysis of Kay Mann:
kay.mann@enron.com(Head of Legal)

- VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.
- The VADER sentiment lexicon is sensitive to both the polarity and the intensity of sentiments expressed in social media contexts
- The heuristics of VADER go beyond what would normally be captured in a typical bag-of-words model. They incorporate word-order sensitive relationships between terms.
- VADER belongs to a kind of sentiment analysis that depends on lexicons of sentiment-related words. In this methodology, every one of the words in the vocabulary is appraised with respect to whether it is positive or negative, and, how +ve or -ve.

```
Precision: 0.833
Recall: 0.928
Accuracy: 0.884
f1-score: 0.877
```



ENRON STOCK PRICE AND VOLUME



ENRON STOCK PRICE AND SENTIMENT SCORE CORRELATION

From	Correlation	Count
:({'chairman.enron@enron.com'})	0.44	285
:({'suzanne.adams@enron.com'})	0.40	263
:({'b..sanders@enron.com'})	0.38	302
:({'robert.cotten@enron.com'})	0.30	320
:({'feedback@intcx.com'})	0.29	607
:({'legal <.taylor@enron.com>'})	0.27	430
:({'newsletter@rigzone.com'})	0.26	277
:({'louise.kitchen@enron.com'})	0.26	1274
:({'pete.davis@enron.com'})	0.26	9148
:({'public.relations@enron.com'})	0.25	412
:({'kaminski@enron.com'})	0.25	250
:({'kay.mann@enron.com'})	0.25	13693
:({'kevin.hyatt@enron.com'})	0.23	566

Email	Correlation	Count
:({'jkeffer@kslaw.com'})	0.41	303
:({'controllers.dl-ets@enron.com'})	0.29	293
:({'kay.chapman@enron.com'})	0.28	340
:({'all.downtown@enron.com'})	0.26	289
:({'judy.hernandez@enron.com'})	0.26	344
:({'gregg.penman@enron.com'})	0.25	386
:({'bill.williams@enron.com'})	0.24	448
:({'suzanne.adams@enron.com'})	0.23	1800
:({'jennifer.fraser@enron.com'})	0.23	328
:({'bruce.mills@enron.com'})	0.23	294
:({'dl-ga-all_enron_worldwide2@enron.com'})	0.22	400
:({'pete.davis@enron.com'})	0.22	9155
:({'kenneth.lay@enron.com'})	0.21	1397

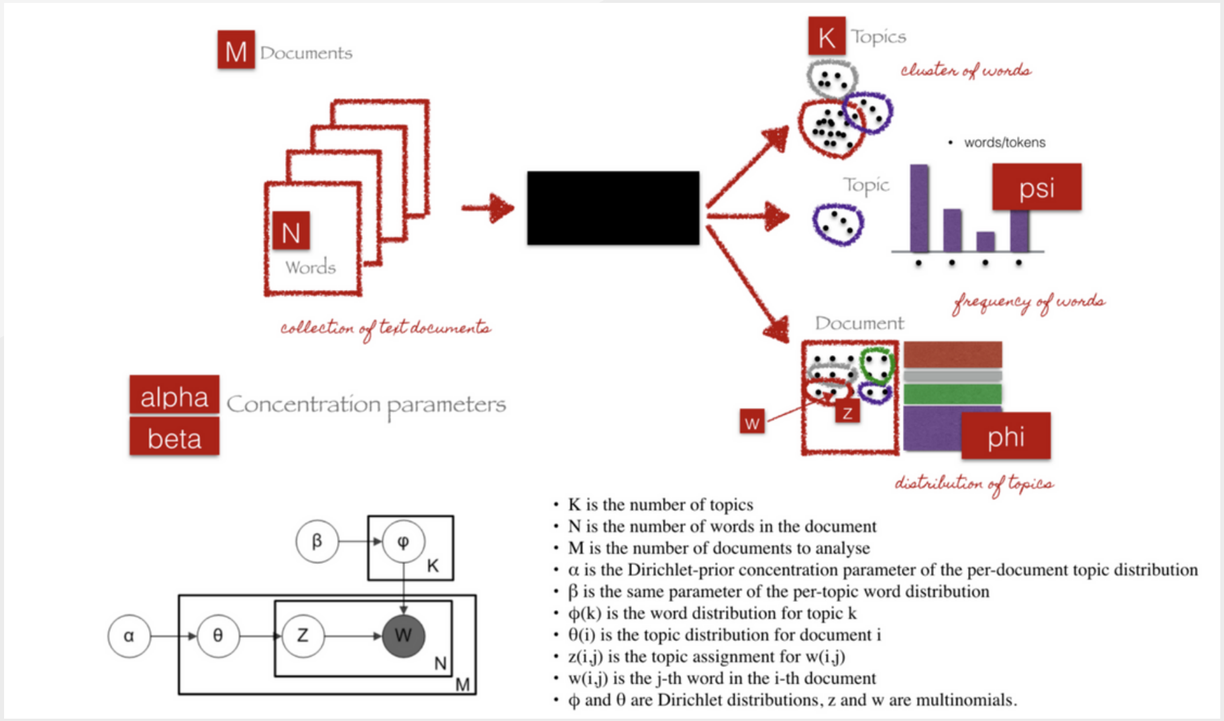
Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance.

TOPIC MODELLING

- Dimensionality Reduction, where rather than representing a text T in its feature space as {Word_i: count(Word_i, T) for Word_i in Vocabulary}, you can represent it in a topic space as {Topic_i: Weight(Topic_i, T) for Topic_i in Topics}

- Unsupervised Learning, where can be compared to clustering, as in the case of clustering, the number of topics, like the number of clusters, is an output parameter. By doing topic modelling, we build clusters of words rather than clusters of texts. A text is thus a mixture of all the topics, each having a specific weight

- Tagging, abstract “topics” that occur in a collection of documents that best represents the information in them.

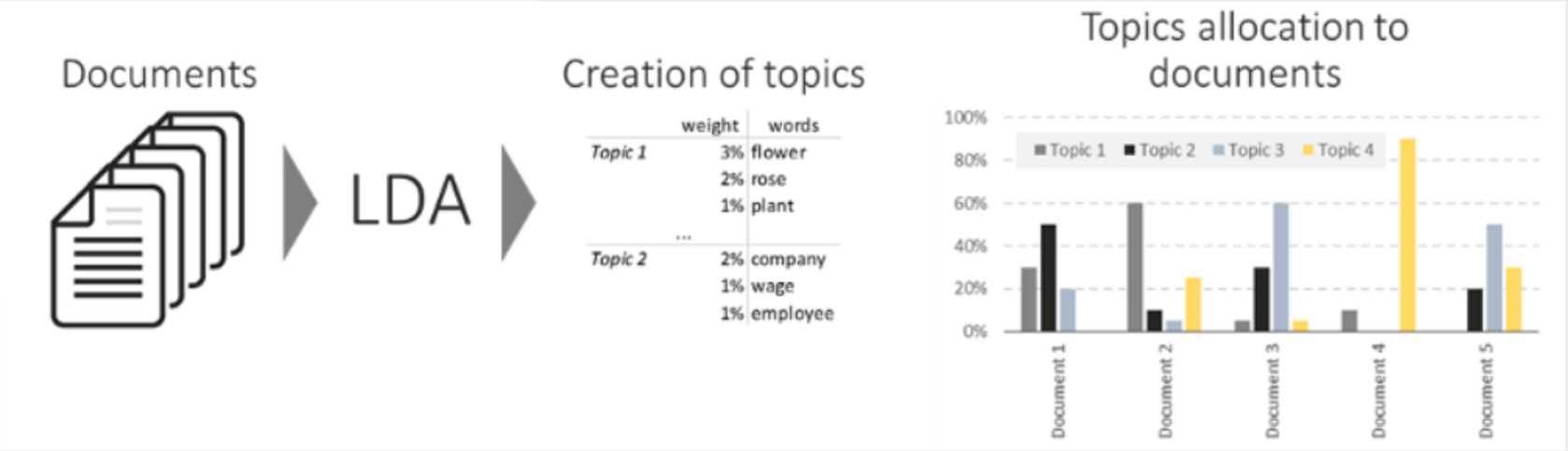


TOPIC MODELLING WITH LDA

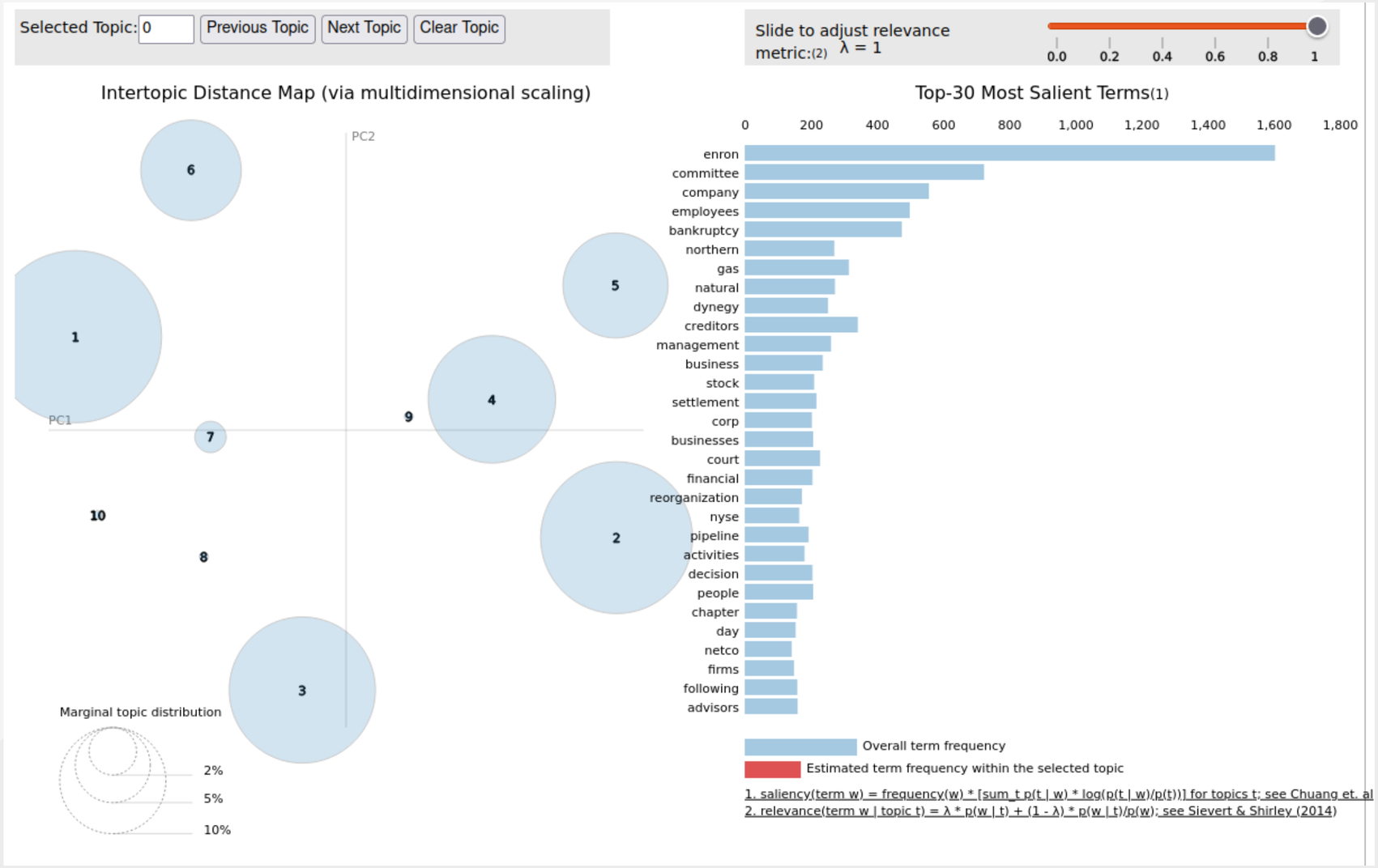
- In LDA, latent indicates the hidden topics present in the data then Dirichlet is a form of distribution.
- Dirichlet distribution is different from the normal distribution. When ML algorithms are to be applied the data has to be normally distributed or follow Gaussian distribution.
- The normal distribution represents the data in real numbers format whereas the Dirichlet distribution represents the data such that the plotted data sums up to 1.

LDA has three important hyperparameters:

1. ‘alpha’ which represents the document-topic density factor
2. ‘beta’ which represents word density in a topic
3. ‘k’ or the number of components representing the number of topics you want the document to be clustered or divided into parts.



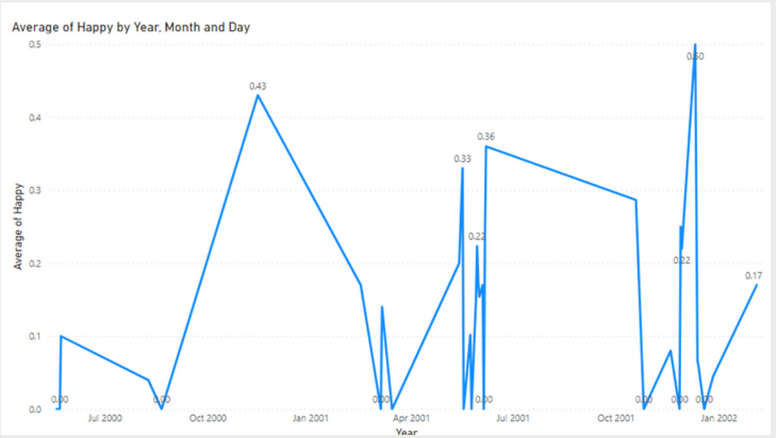
TOPIC MODELLING WITH LDA



EMOTION CLASSIFICATION

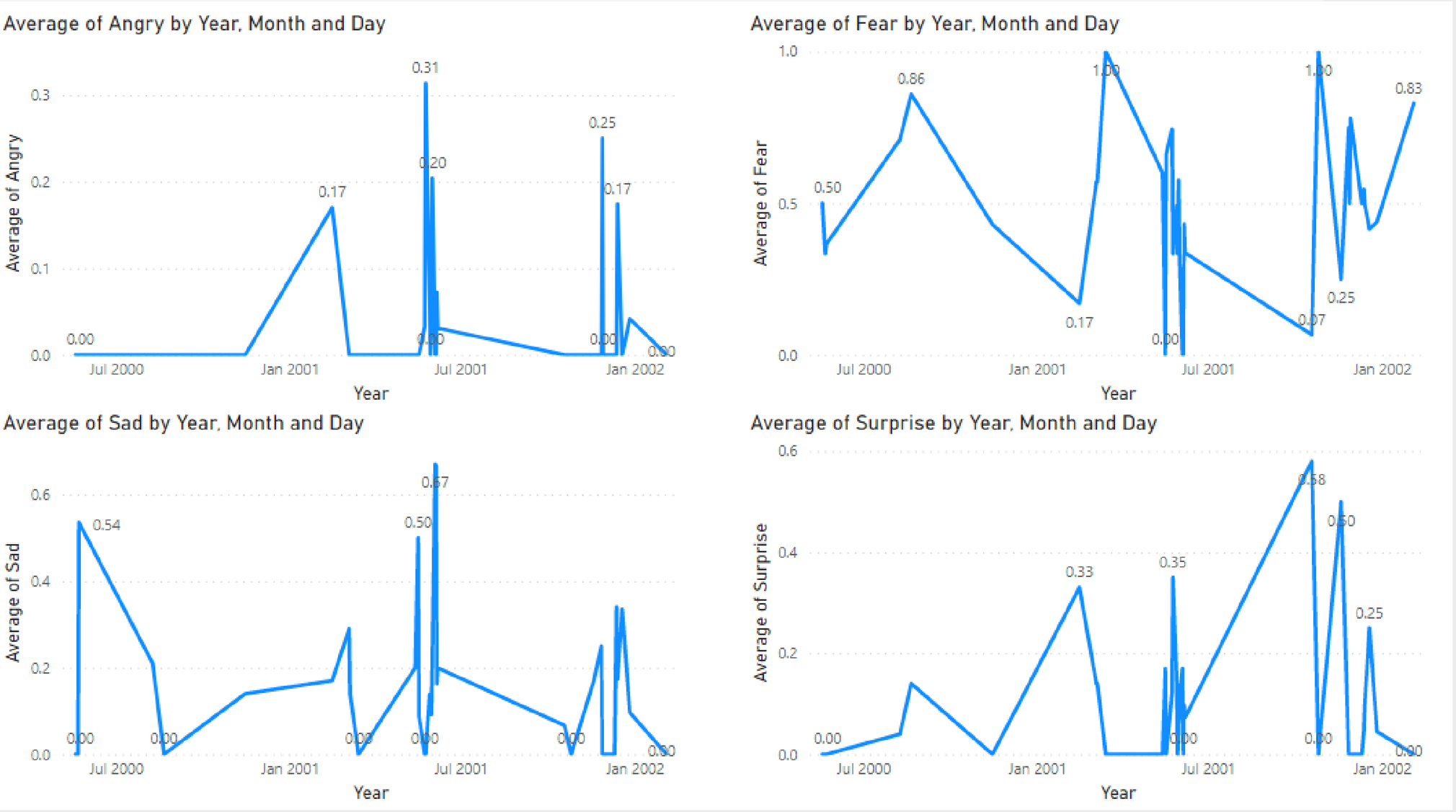
Year	Month	Day	content_clean	compound	emotion
2001	December	14	oh pissed mess around people money think mistake correcting trying freak yet original message mann kay friday december adam suzanne opinion unreal blackberry wireless handheld www blackberry net	-0.87	{'Happy': 0.0, 'Angry': 0.67, 'Surprise': 0.0, 'Sad': 0.0, 'Fear': 0.33}
2000	August	9	mark understand moving going vacation starting august th entire next week also august th finish dental work already asked joya let know soon regarding firm date get carol packed know send thing please advise thanks suz forwarded suzanne adam hou ect carol st clair suzanne adam hou ect ect mark taylor hou ect ect office build suzanne fyi clerk handle packing office may able help want make commitment make hopefully take long file need marked perhaps put legal assistant area file near window move timing slip happens vacation would prefer handled hand floater worry let know think sorry impose tried much could left carol st clair eb phone fax carol st clair enron com forwarded carol st clair hou ect mark taylor carol st clair hou ect ect office build look like phase ii construction project julia office used finished th moving new office weekend plan phase iii turn office included construction area phase iii probably start th could always slip wanted give much notice possible hope well getting sleep yet	2.88	{'Happy': 0.04, 'Angry': 0.0, 'Surprise': 0.04, 'Sad': 0.21, 'Fear': 0.71}
2001	June	6	funny kay mann enron suzanne adam hou ect ect weekly ge conference call think god gift contract suzanne adam ect kay mann corp enron enron weekly ge conference call attempt humor email list	2.96	{'Happy': 0.06, 'Angry': 0.06, 'Surprise': 0.11, 'Sad': 0.28, 'Fear': 0.5}
Total				**176.40**	

- Text2Emotion is the python package which will help you to extract the emotions from the content.
- Compatible with 5 different emotion categories Happy, Angry, Sad, Surprise and Fear.
- Involves text pre-processing, emotion investigation and emotion analysis



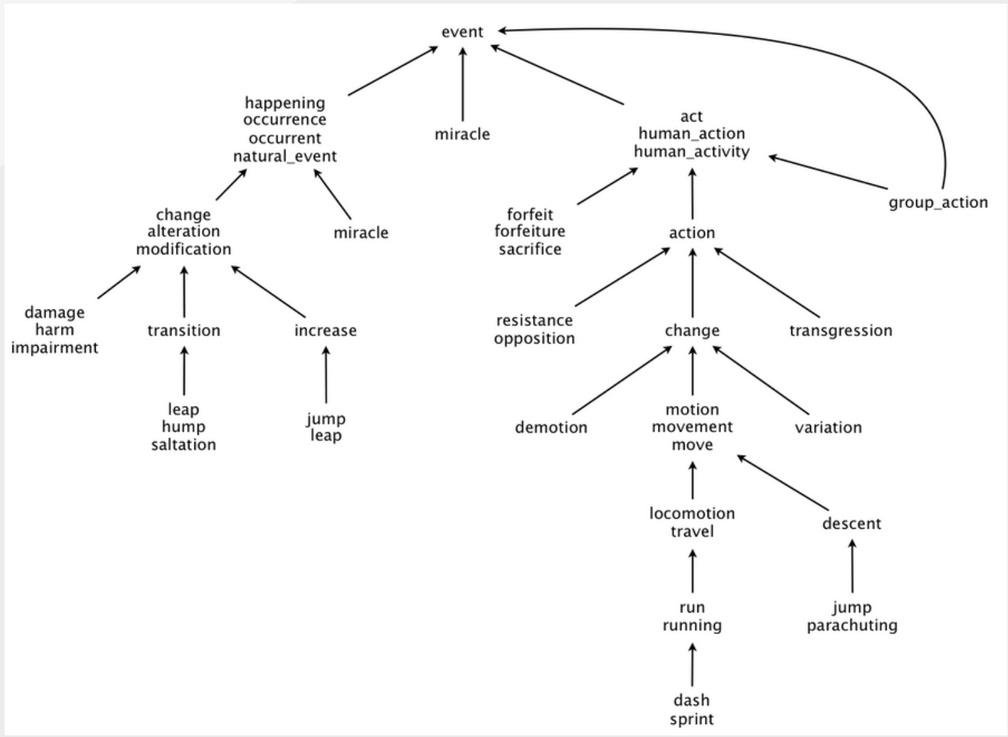
Note: A future possibility might be to use GoEmotions which is a corpus of 58k carefully curated comments extracted from Reddit, with human annotations to 27 emotion categories or Neutral based on a pre-trained BERT-based(Bidirectional Encoder Representations from Transformers) model

EMOTION CLASSIFICATION

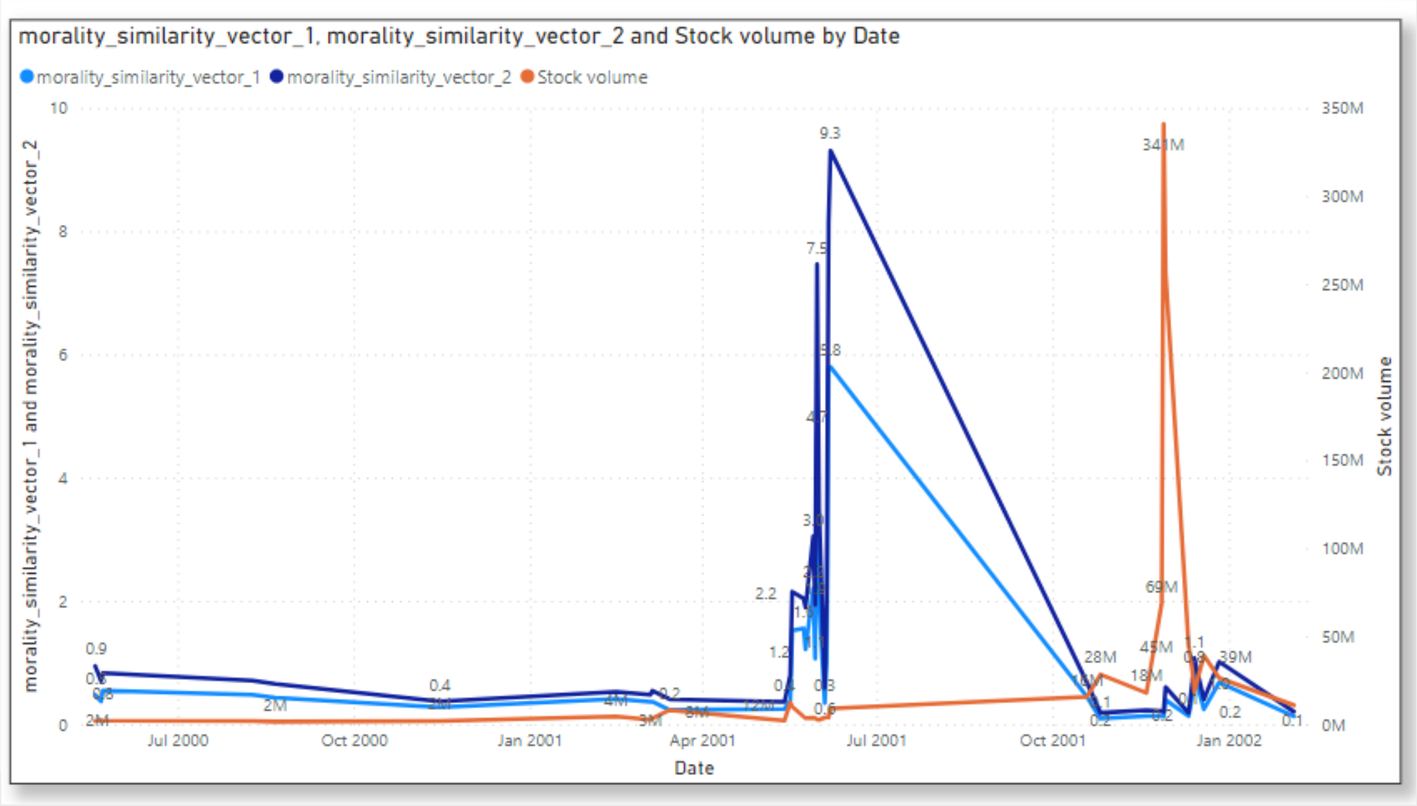


DETECTION OF MORALITY IN EMAIL CONTENT USING SYNSETS

- WordNet is the lexical database i.e. dictionary for the English language, specifically designed for natural language processing by Princeton University
- Synset is a special kind of simple interface that is present in NLTK to look up words in WordNet. Synset instances are the groupings of synonymous words that express the same concept. Some of the words have only one Synset and some have several.
- Hypernyms: More abstract terms, Hyponyms: More specific terms.
- Both come into the picture as Synsets are organized in a structure similar to that of an inheritance tree. This tree can be traced all the way up to a root hypernym. Hypernyms provide a way to categorize and group words based on their similarity to each other.
- Two kinds of similarity were calculated: Between focus word synset(wn.synset('morality.n.1')) and content, and then between content and focus word synset(wn.synset('morality.n.1'))



DETECTION OF MORALITY IN EMAIL CONTENT USING SYNSETS



Note:
quality, attribute, abstraction, entity, conscience, good, righteousness, rightness, virtue, conscientiousness, unconscientiousness, beneficence, benignity, kindness, saintliness, summum, bonum, virtue, virtue, honesty, honor, honorableness, impeccability, justice, piety, uprightness, honor, religiousness, grace, benevolence, consideration, generosity, loving-kindness, cardinal, virtue, candor, good, faith, incorruptibility, incorruptness, integrity, scrupulousness, truthfulness, nobility, respectability, venerability, fairness, right, devoutness, dutifulness, godliness, charity, attentiveness, tact, bigheartedness, bounty, charitableness, liberality, unselfishness, natural, virtue, theological, virtue, ingenuousness, probity, sincerity, sooth, veracity, high-mindedness, sublimity, decency, non-discrimination, sportsmanship, religiosity, delicacy, savoir-faire, munificence, altruism, fortitude, prudence, temperance, hope, religion, artlessness, heartiness, singleness, backbone, frugality, providence, abstemiousness, sobriety, apophatism

RESULTS

- For sentiment analysis after comparison against hand-labelled data accuracy of 88.4 and f1-score of 87.7 was achieved above target
- For topic modelling analysis of emails from chairman revealed words related to scandal such as bankrupcy and litigation while for low-level executive words related to department and corporate lingos were observed
- No significant patterns were observed in emotion analysis
- A proxy for detection of moral language was used so can't state reliable results

FUTURE PLANS

- Parallelize analysis processes used such as LDA and emotion detection
- Targetted sentiment analysis
- Develop ways for moral language detection using social science theories
- Use transformer-based models like BERT wherever possible for better accuracy

THANK YOU

Please reach out to me if you have any suggestions or ideas

CONTACT


aditya.krishn@rwth-aachen.de


+49-1782857836

www.linkedin.com/in/aditya-krishn