SENTIMENT ANALYSIS

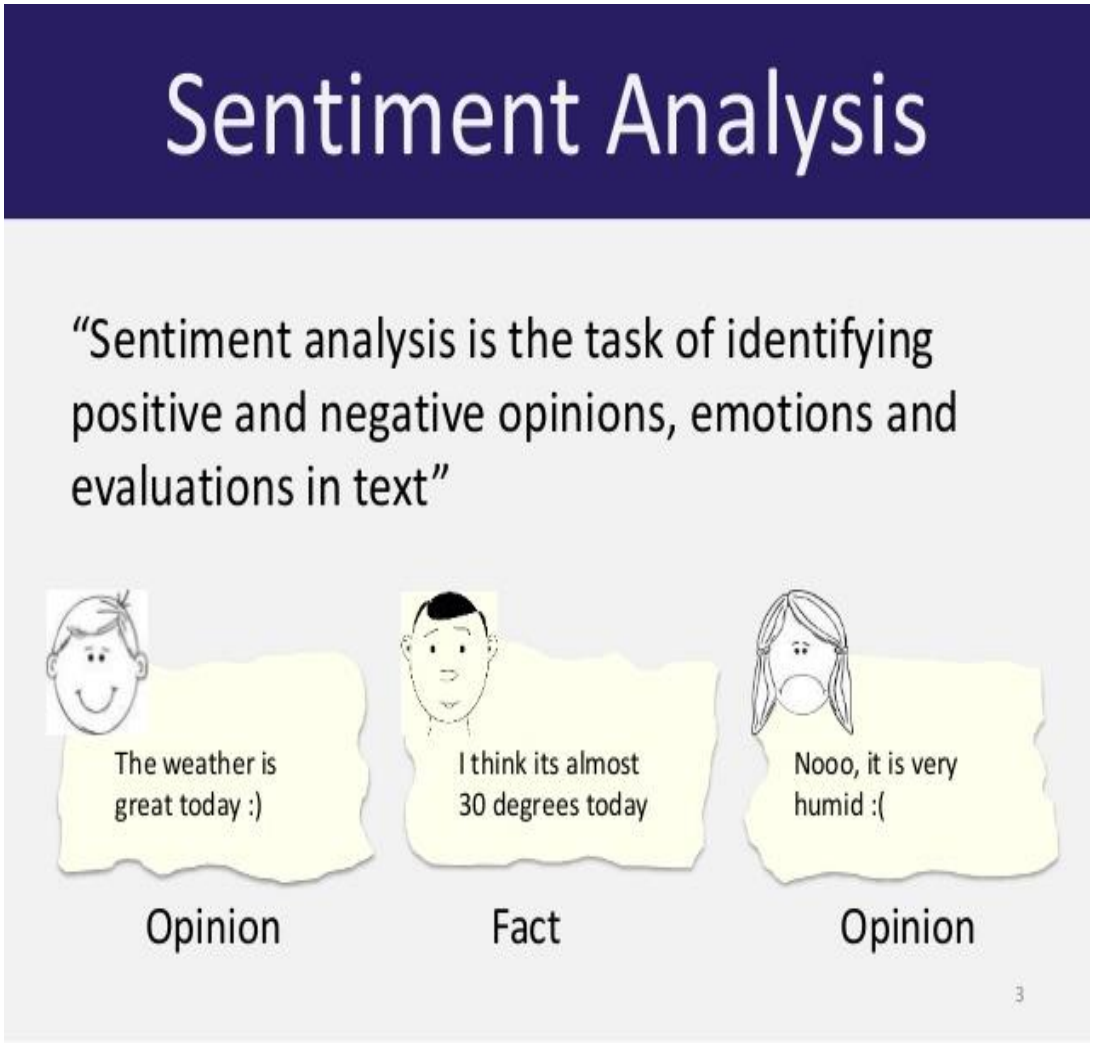
USING SEMI- SUPERVISED RECURSIVE AUTOENCODER

By : Vinay Kumar, CSE , IIT Kanpur
Guided by: Prof. Amitabha Mukerjee

Sentiment Analysis: What is it and why are people doing something about it?

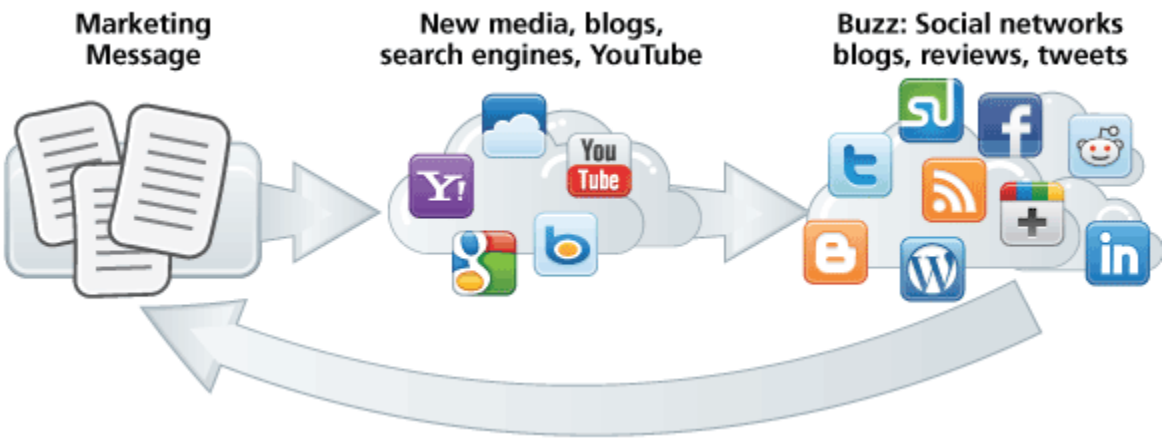
Sentiment Analysis

Definition:
"It is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes(from sources like text)."
-Pang, Bo and Lillian Lee

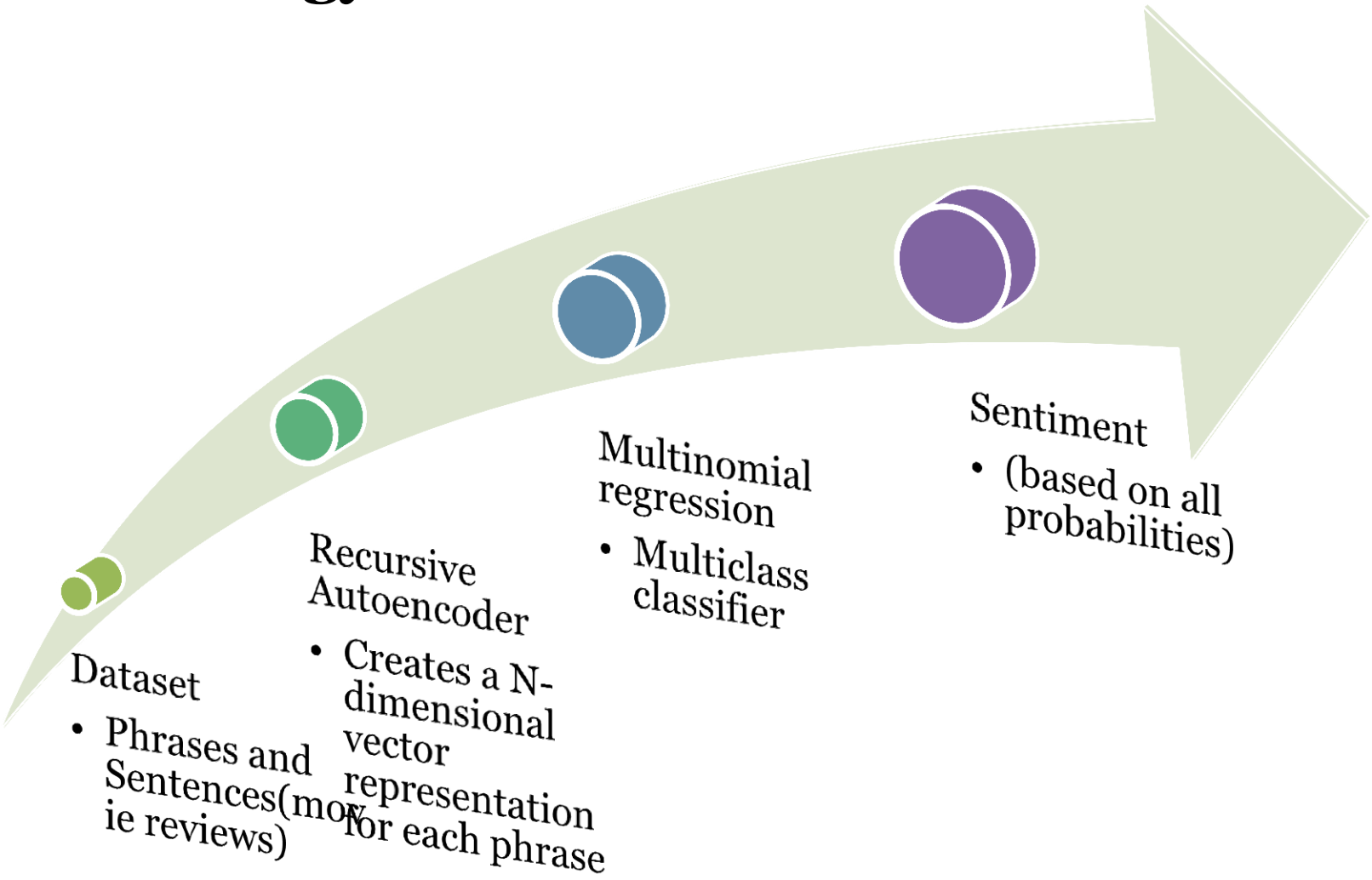


Sentiment Analysis (Applications)

- **Stock Prediction**
- **Social Media Monitoring**
- **Twitter Sentiment**
- **Movie/product reviews**



Methodology



Recursive Autoencoder

Multinomial Regression(Softmax):

Cost Function:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(i)} = j\right\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}\right]$$

Notice that this generalizes the logistic regression cost function, which could also have been written:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}(1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) + y^{(i)} \log h_\theta(x^{(i)})\right]$$

$$= -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=0}^{1} 1\left\{y^{(i)} = j\right\} \log p(y^{(i)} = j|x^{(i)};\theta)\right]$$

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(i)} = j\right\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}\right] + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{ij}^2$$

:Weight decay

Class label probablities:

$$p(y^{(i)} = j|x^{(i)};\theta) = \frac{e^{(\theta_j-\psi)^T x^{(i)}}}{\sum_{l=1}^{k} e^{(\theta_l-\psi)^T x^{(i)}}}$$

$$= \frac{e^{\theta_j^T x^{(i)}} e^{-\psi^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}} e^{-\psi^T x^{(i)}}}$$

$$= \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}.$$

Dataset

Consists of 8544 sentences which is converted to 156060 English phrases from movie reviews.

Kaggle. 2014. Sentiment Analysis on Movie Reviews. https://www.kaggle.com/c/sentimentanalysis-on-movie-reviews.

Issues and Resolutions

- Randomized N-dimensional vector : Using **word2vec** to map similar words nearby to exploit the similarity of meaning/context.
- Belong to more than one class -> Unnecessary errors : Using **Softmax regression** instead of logistic regression.
- Altering the values of random parameters as N, α, the parameter that controls the relative weighting between the reconstruction error, error of the unsupervised part, and the cross-entropy error, error of the supervised learning part.

Source code for Recursive Auto-encoder:

http://www.socher.org/index.php/Main/Semi-SupervisedRecursiveAutoencodersForPredictingSentimentDistributions

Thanks to Richard Socher (Phd fellow at Stanford University) and Stanford CoreNLP for source code of Recursive Autoencoder

References:

[1] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. 2011b. SemiSupervised Recursive Autoencoders for Predicting Sentiment Distributions. In EMNLP Url: http://ai.stanford.edu/ang/papers/emnlp11-RecursiveAutoencodersSentimentDistributions.pdf

[2] Bahareh Ghiyasian and Yun Fei Guo. 2014. Sentiment Analysis Using SemiSupervised Recursive Autoencoders and Support Vector Machines. Stanford.edu

[3] Google. 2014. word2vec. https://code.google.com/p/word2vec/

Questions?