Introduction to NLP

Think *tokens*, not words

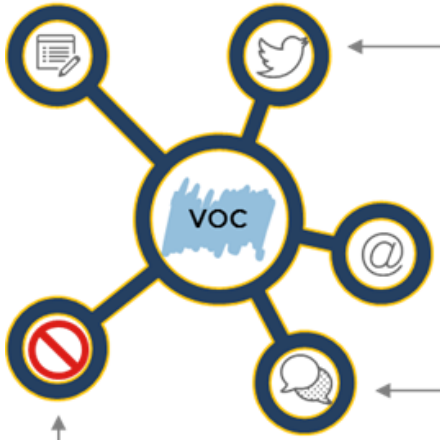
Manas Ranjan Kar

What's NLP ?

Natural Language Processing (NLP) is processing information contained in human language, applying computational techniques to language domain.

It started off as a branch of AI, borrows from linguistics, psycholinguistics, cognitive science and statistics. It makes computers learn our language than we learn theirs. NLP & Computational Linguistics (CL) are used interchangeably

There’s a deluge of information



Internal Reviews & NPS scores, collected via tablet/paper/emails

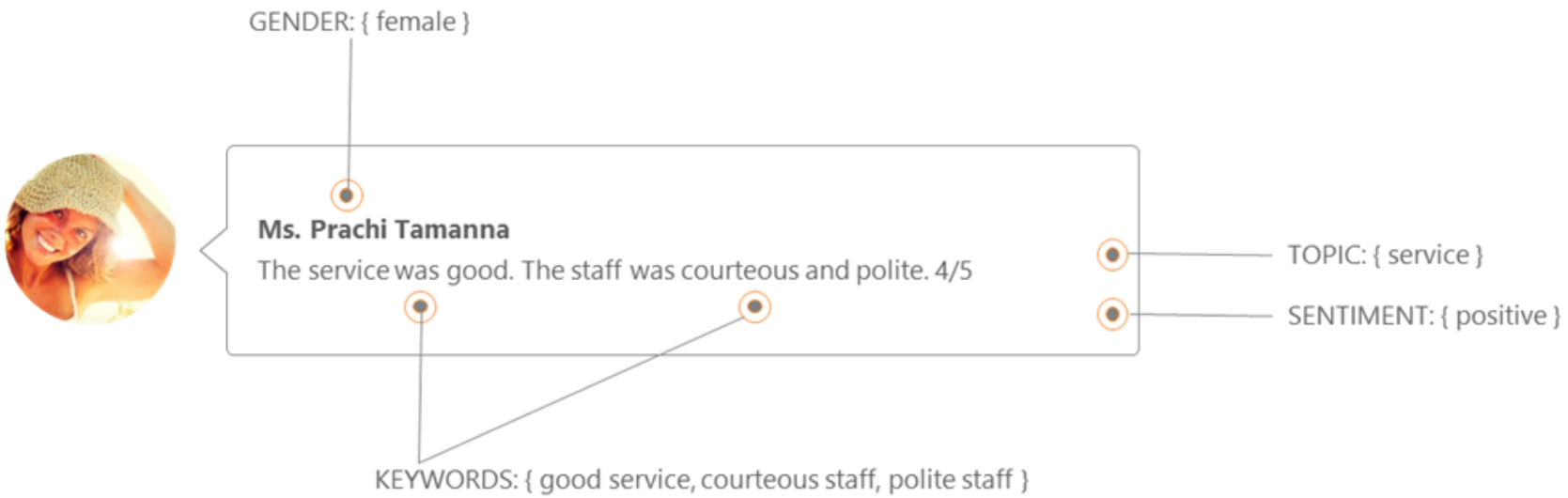
Social posts, on FB/Twitter, directed towards the brand

Emails, mostly for complaints or seeking information

External reviews – on sites like Zomato, Tripadvisor, Mouthshut

NOISE – Reviews which are either ‘fake’ or ‘planted’

Analyzing information



The client had thousands of customer reviews which they wanted to analyse - to understand customer feedback and identify improvement opportunities.

The **broad questions** we focused on;

PRIMARY FOCUS AREAS

What did they say about the restaurant?

Keywords & topics of discussion across the comments

What elements of the restaurant would they want improved? – service, staff behaviour, ambience etc.

SECONDARY FOCUS AREAS

When did the customer visit the store?

How is client's traffic distributed over time?

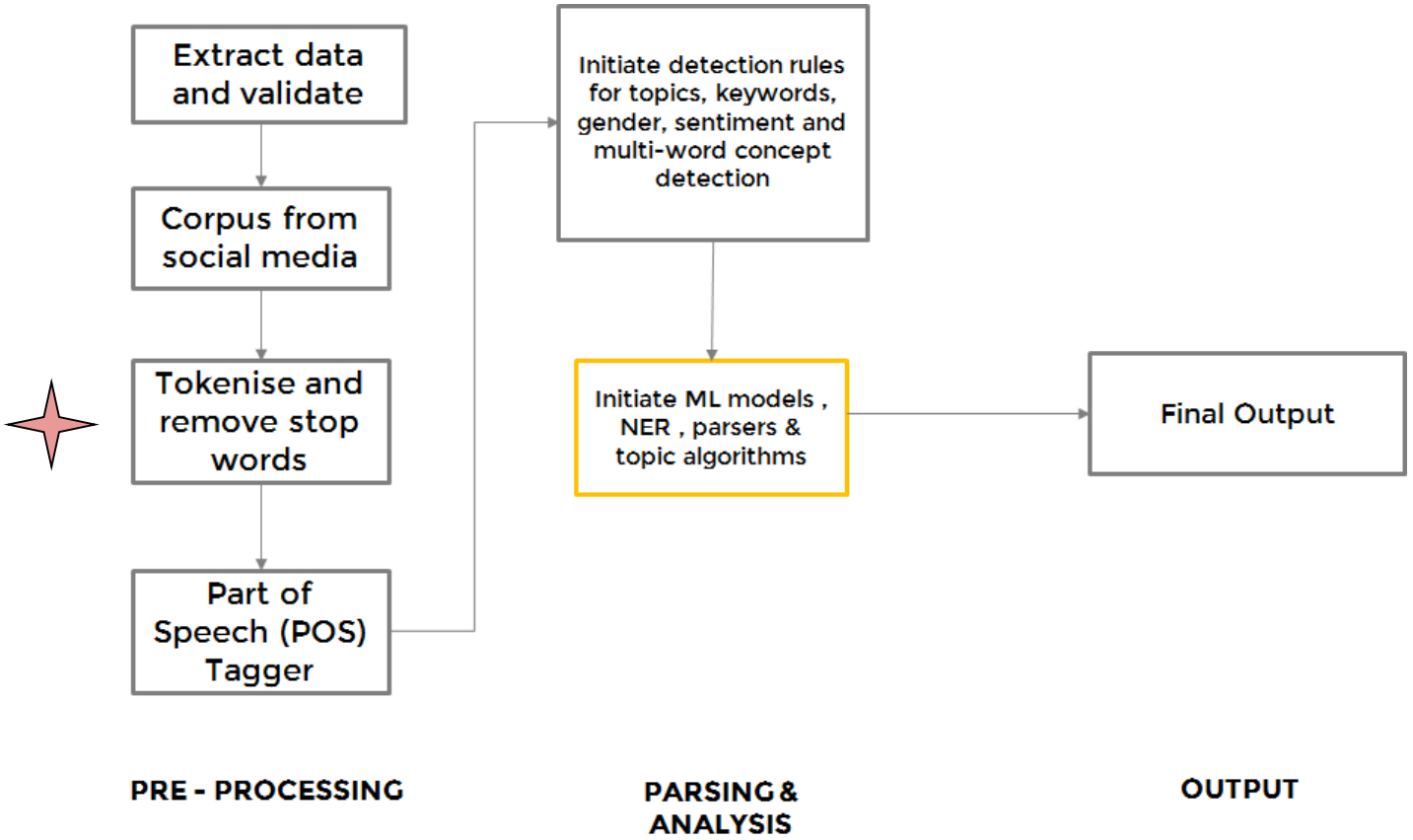
Ticket sizes across multiple customer dimensions – age, gender, ratings, location, time of visit etc.

Overall customer sentiments & views about UCH

Why is NLP difficult

1. No single architecture works everywhere
1. Pipelines need to be modified for different domain
1. Languages - US vs UK vs Indian English
1. Preprocessing kills people
1. Clustering techniques need maximum love and care
1. **Lack of training data**
1. Turnaround times can be higher for clients
1. CS + **Linguistics** + Algorithms + Databases -> Beat that !

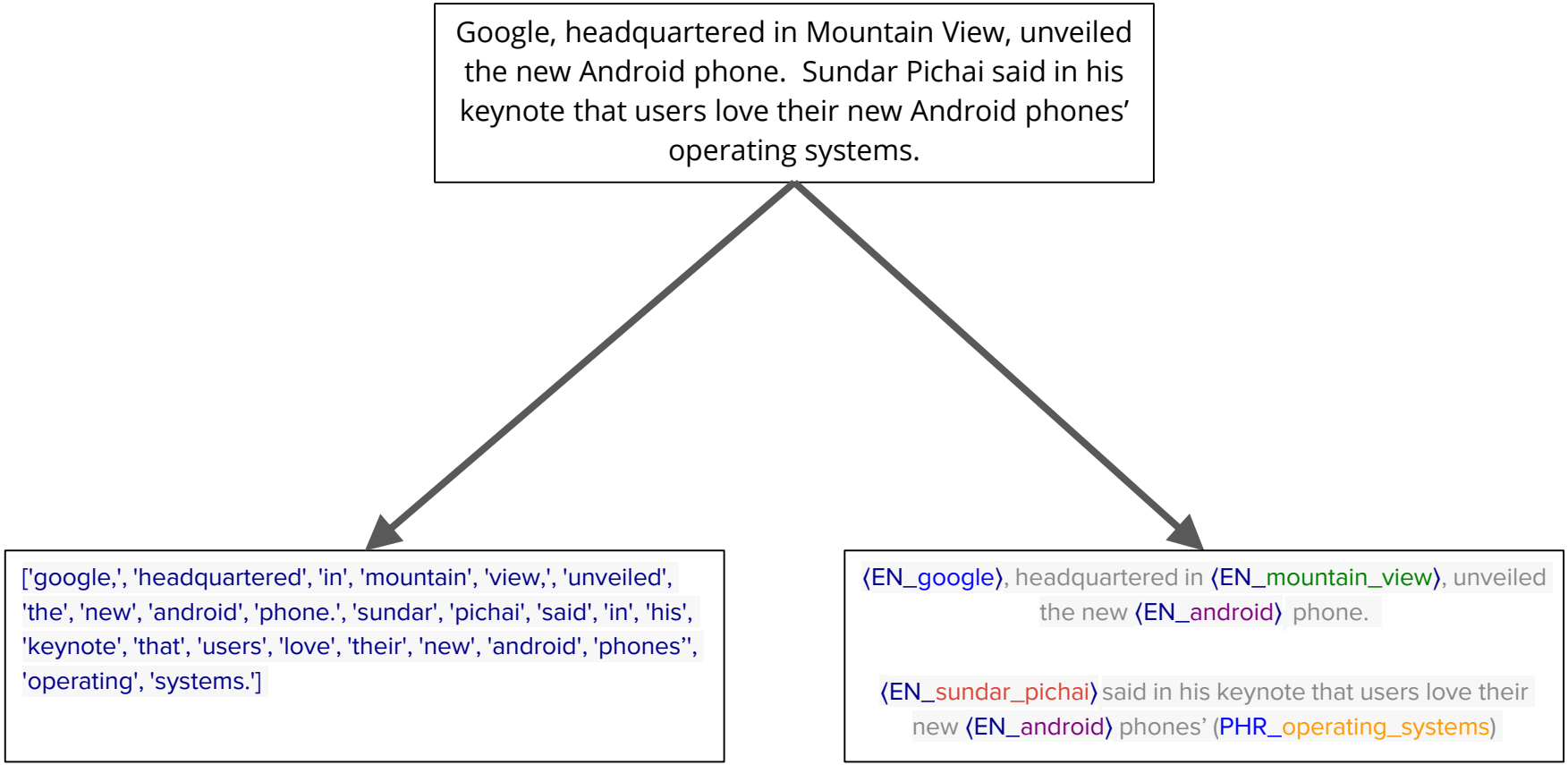
Typical NLP Pipeline



Stepwise into NLP

Sentence segmentation ***Identify sentence boundaries***	Frank met the president. He said: "Hi! What's up – Mr. President?"	**Sentence 1:** Frank met the president. **Sentence 2:** He said: "Hi What's up – Mr. President?"
Tokenization ***Identify word boundaries***	My phone tries to change 'eating' to 'dating'. #hateautocorrect	[My] [phone] [tries] [to] [change] ['] [eating] ['] [to] ['] [dating] ['] [.] [#hateautocorrect]
Stemming/lemmatization	eating, ate, eat	eat, eat, eat
Part-of-Speech tagging	If you build it, he will come	If you build it , he will come IN PRP VBP PRP , PRP MD VB
Parsing	Jon and Frank went into a bar.	(S (NP (NP John) and (NP Frank)) (VP went (PP into (NP a bar))))
Named entity recognition	Let's meet John in DC at 6pm.	Let's meet **John** in **DC** at **6pm**. Pers Loc Time
Co-reference resolution	John drank a beer. He thought it was warm.	**John** drank a **beer**. **He** thought **it** was warm.

Possible Transformations for tokens



Makings of a good NLP Scientist

1. Maximum enrichment - add more information
1. Minimum information loss - minimal stopwords, minimal stemming
1. Assume word dependencies - do BOW, don't think BOW
1. Features engineering is key !
1. Make a pipeline - clean, tokenize, NER, phrase extraction, ML - reuse architectures*
1. Create your datasets - Crowdflower, Amazon MT or scrape !
1. Learning to create grammars for parsing
1. Understanding Linguistics and how it affects your pipeline
1. Be close to the business problem - don't make it an academic hunt

Data Resources

RESOURCE	DESCRIPTION	LINK
WORDNET	Large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.	https://wordnet.princeton.edu/
SENTIWORDNET	SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity.	http://sentiwordnet.isti.cnr.it/
LDC	LDC's Catalog contains hundreds of datasets for research, both free and paid.	https://www.ldc.upenn.edu/
WIKIPEDIA	Wikipedia offers free copies of all available content to interested users. These databases can be used for mirroring, personal use, informal backups, offline use or database queries	https://dumps.wikimedia.org/enwiki/
MOVIE REVIEWS	Dataset of 25,000 highly polar movie reviews for training, and 25,000 for testing.	http://ai.stanford.edu/~amaas/data/sentiment/

Questions?

Manas Ranjan Kar	manasrkar91@gmail.com GitHub: manasRK @manasrnkar