
# Applications of Natural Language Processing

Aditya Joshi | October 2022

Source: https://www.ngv.vic.gov.au/explore/collection/work/145612/

# Applications of Natural Language Processing

## A Research Perspective

Aditya Joshi, PhD
SEEK, Australia

aditya.m.joshi@gmail.com

# Natural Language Processing (NLP)

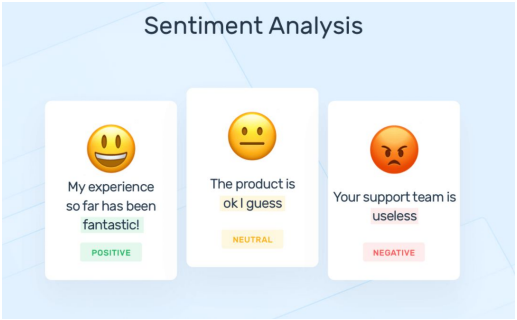
A sub-branch of artificial intelligence that deals with processing of language

Also known as ‘computational linguistics’

NLP spans fundamental linguistic research areas (such as part-of-speech tagging and parsing) to task-oriented research areas (such as sentiment analysis, translation and summarisation)

NLP papers (including NLP applications) are publicly available at: https://aclanthology.org/

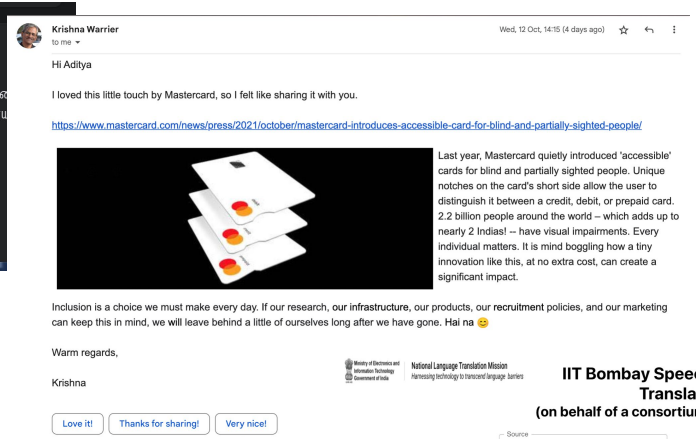
| Why | not | tell | someone | ? |
|---|---|---|---|---|
| adverb | adverb | verb | noun | punctuation mark, sentence closer |



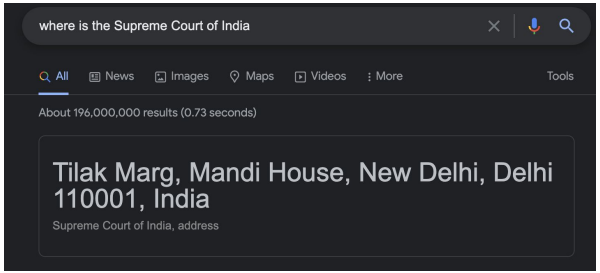

https://towardsdatascience.com/part-of-speech-tagging-for-beginners-3a0754b2ebba
https://vitalflux.com/sentiment-analysis-machine-learning-techniques/

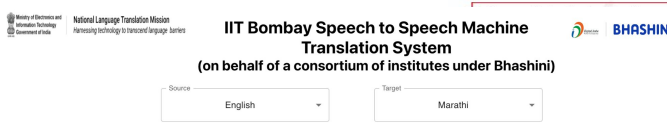
‘Applications of Natural Language Processing’ by Aditya Joshi at Dept of CSE, IIT Bombay

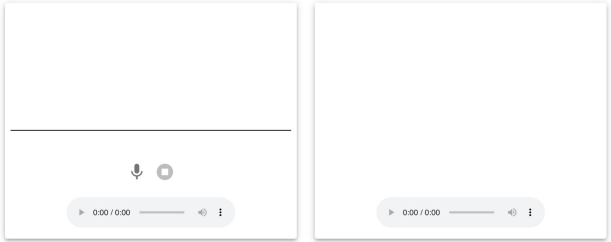
# Where do we see NLP?






## The impact of NLP is in its applications.

‘Applications of Natural Language Processing’ by Aditya Joshi at Dept of CSE, IIT Bombay

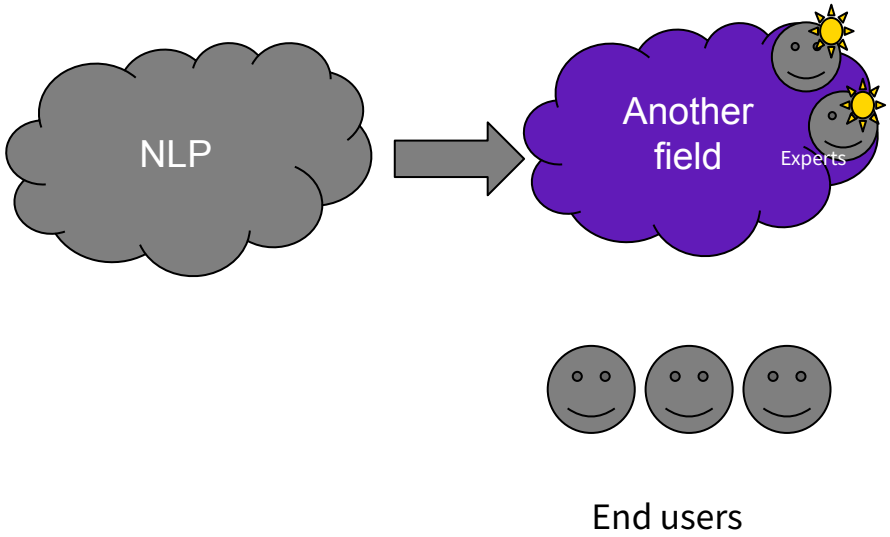
# In this talk, we will..

- Introduce *some* applications of NLP - via (primarily) research papers
- Understand how to conceptualise a ‘NLP application’ project
- Not necessarily discuss the details of the approach
- **Give a flavour of the positive utility of NLP applications**

# What is an application?

"the action of putting something into operation." (Source: OED)

In the context of this talk:
"Interdisciplinarity" or interdisciplinary studies involves the combination of two or more academic disciplines into one activity (e.g., a research project). (Source: Wikipedia)



# What can NLP applications achieve?

Research papers

Demonstration papers

Startup/business ideas!

## Call for Papers

Commit Your Paper To ACL 2022 Here

### MAIN CONFERENCE

ACL 2022 invites the submission of long and short papers featuring substantial, original, and unpublished research in all aspects of Computational Linguistics and Natural Language Processing. As in recent years, some of the presentations at the conference will be of papers accepted by the Transactions of the ACL (TACL) and by the Computational Linguistics (CL) journals.

### SUBMISSIONS TOPICS

ACL 2022 aims to have a broad technical program. Relevant topics for the conference include, but are not limited to, the following areas (in alphabetical order):

- Computational Social Science and Cultural Analytics
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Ethics and NLP
- Generation
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Language Grounding to Vision, Robotics and Beyond
- Linguistic Theories, Cognitive Modeling, and Psycholinguistics
- Machine Learning for NLP
- Machine Translation and Multilinguality
- NLP Applications
- Phonology, Morphology, and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical
- Semantics: Sentence-level Semantics, Textual Inference, and Other Areas
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech and Multimodality
- Summarization
- Syntax: Tagging, Chunking and Parsing
- Theme: "Language Diversity: from Low-Resource to Endangered Languages"

# Outline

Part I: Overview of NLP applications

Part 2: Typical steps to build an NLP application

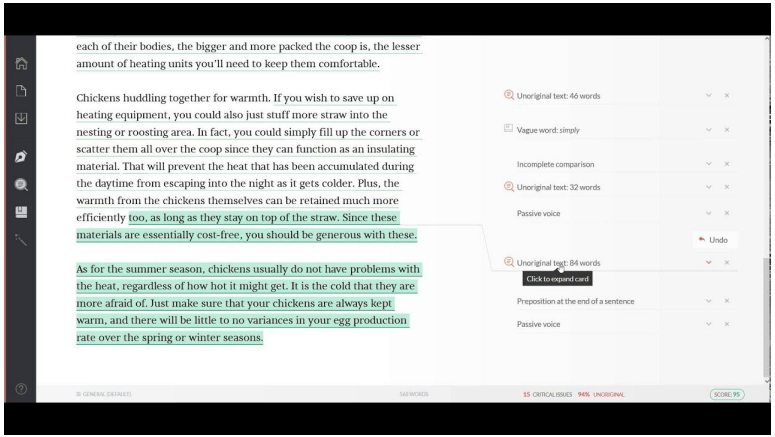
Part 3: Our story of an NLP application (optional)

# Applications of Natural Language Processing

## Part I: Overview of NLP applications

# NLP in Education Technology

## Grammar correction




(Image not from the paper. However, the paper describes an approach) Lichtarge, Jared, Chris Alberti, and Shankar Kumar. "Data weighted training strategies for grammatical error correction." *TACL. 2020.*

## Student Engagement Detection

RoomReader:

- 8 hours of video and audio recordings from 118 participants in 30 gender-balanced sessions.
- The recordings have been edited, synchronised, and fully transcribed. Student participants have been continuously annotated for engagement with a novel continuous scale

Justine Reverdy, Sam O'Connor Russell, Louise Duquenne, Diego Garaialde, Benjamin R. Cowan, and Naomi Harte. 2022. RoomReader: A Multimodal Corpus of Online Multiparty Conversational Interactions. In LREC.

Dataset: https://www.cl.cam.ac.uk/research/nl/bea2019st/

# NLP & Literature

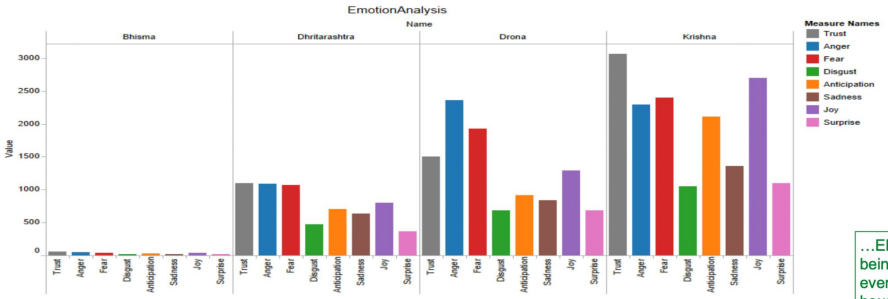



Figure 8: Emotion Analysis of Bhisma, Dhritarashtra, Drona and Krishna

Debarati Das, Bhaskarjyoti Das, and Kavi Mahesh. A Computational Analysis of Mahabharata. ICON 2016.

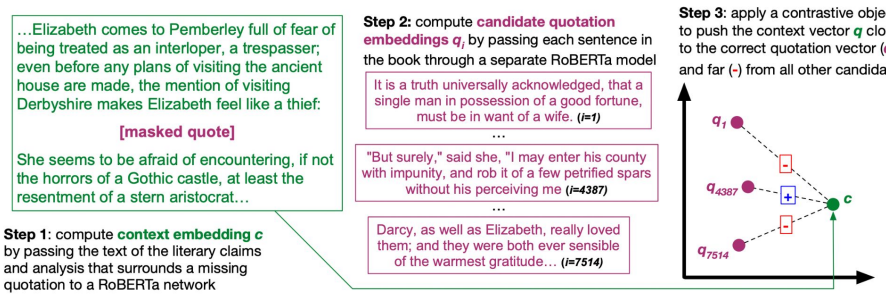
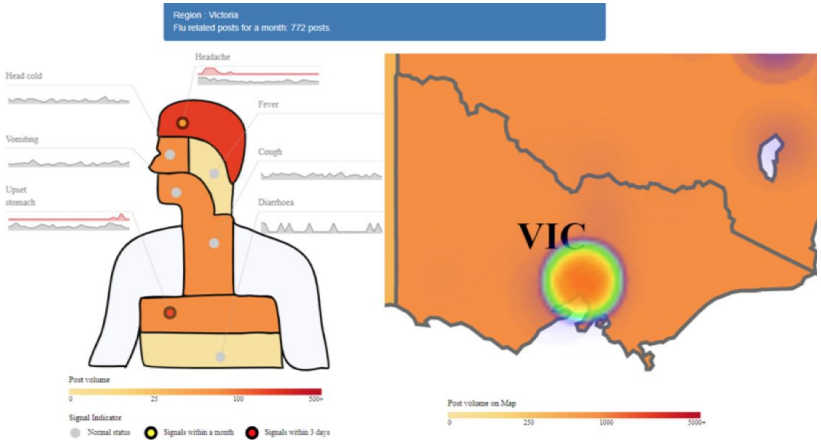



Figure 1: An example of our *literary evidence retrieval* task and the model we built to solve it. The model must retrieve a missing quotation from *Pride and Prejudice* given the literary claims and analysis that surround the quotation. The retrieval candidate set for this example consists of all 7,514 sentences from *Pride and Prejudice*. Our dense-RELiC model is trained with a contrastive loss to push a learned representation of the surrounding context close to a representation of the ground-truth missing quotation (here, the 4,387th sentence from the novel).

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022. RELiC: Retrieving Evidence for Literary Claims. In ACL 2022.

# NLP in the medical domain

## Social media-based epidemic intelligence




Brian Jin*, Aditya Joshi*, Ross Sparks, Stephen Wan, Cecile Paris and C Raina MacIntyre, 'Watch The Flu: A Tweet Monitoring Tool for Epidemic Intelligence of Influenza in Australia', **AAAI 2020.**

## Summarisation of medical papers

**Background:** An individual patient data meta analysis was performed to determine clinical outcomes, and to propose a risk stratification system, related to the comprehensive treatment of patients with oligometastatic nsclc.

**Doc1:** … **We therefore did this phase iii trial to compare concurrent chemotherapy and radiotherapy** followed by resection with st and ard concurrent chemotherapy and definitive radiotherapy without resection … **In an exploratory analysis, os was improved for patients who underwent lobectomy, but not pneumonectomy, versus chemotherapy plus radiotherapy. Chemotherapy plus radiotherapy** with or without resection (preferably lobectomy) are options for patients with stage iiia (n2) non-small-cell lung cancer.

**Doc2:** … Common **adverse events** associated with crizotinib were visual disorder, gastrointestinal side effects, and elevated liver aminotransferase levels, whereas common adverse events with chemotherapy were fatigue, alopecia, and dyspnea. **Patients reported greater reductions in symptoms of lung cancer and greater improvement in global quality of life with crizotinib than with chemotherapy**.

**Doc3:** … First-line gefitinib for patients with advanced non-small-cell lung cancer who were selected on the basis of egfr mutations **improved progression-free survival**, with acceptable toxicity, as **compared with st and ard chemotherapy** …

---

**Ground-truth:** Significant os differences were observed in oligometastatic patients stratified according to type of metastatic presentation, and n status. Long-term survival is common in selected patients with metachronous oligometastases.

**Model:** The pooled **risk stratification** of patients with **oligometastatic nsclc** showed a **significant reduction in the risk of adverse events compared with st and ard chemotherapy, but not radiotherapy.**

Figure 2: A random sampled test set instance. We show how DAMEN selects the information from the background and multiple documents to generate the final summary.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Davide Freddi. Discriminative Marginalized Probabilistic Neural Method for Multi-Document Summarization of Medical Literature. ACL 2022.

# NLP & COVID-19

## COVID misinformation detection:

**Tweet:** "Coronavirus CV19 was a top secret biological warfare experiment. That is why it is only affecting the poor."
**Misconception:** "Coronavirus is genetically engineered."
**Label:** Agree

**Tweet:** "It looks like we are all going to have to wait much longer for a #COVID19 vaccine."
**Misconception:** "We're very close to a vaccine."
**Label:** Disagree

**Tweet:** "CDC: Coronavirus spreads rapidly in dense populations with public transit and regular social gatherings."
**Misconception:** "Coronavirus cannot live in warm and tropical temperatures."
**Label:** No Stance

Figure 1: **CovidLies Dataset.** Given a *tweet*, we annotate whether any of the known *misconceptions* are expressed in the tweet, in particular, if the tweet spreads the misconception (e.g., they Agree), combats the spread of the misconception (e.g., they Disagree), or takes No Stance towards the misconception.

Hossain, Tamanna, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. "COVIDLies: Detecting COVID-19 Misinformation on Social Media." In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. 2020.

## Analysing emotions during the COVID-19 pandemic:

"First, we instructed them to "write in a few sentences how you feel about the Corona situation at this very moment. This text should express your feelings at this moment" (min. 500 characters).

The second part asked them to express their feelings in Tweet form (max. 240 characters) with otherwise identical instructions."

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. 2020. Measuring Emotions in the COVID-19 Real World Worry Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*

# NLP in the legal domain

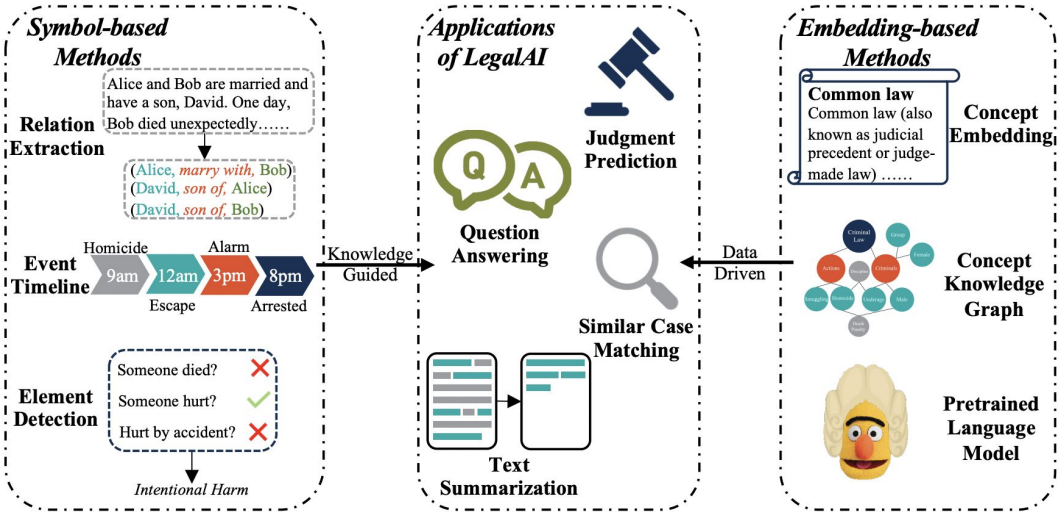


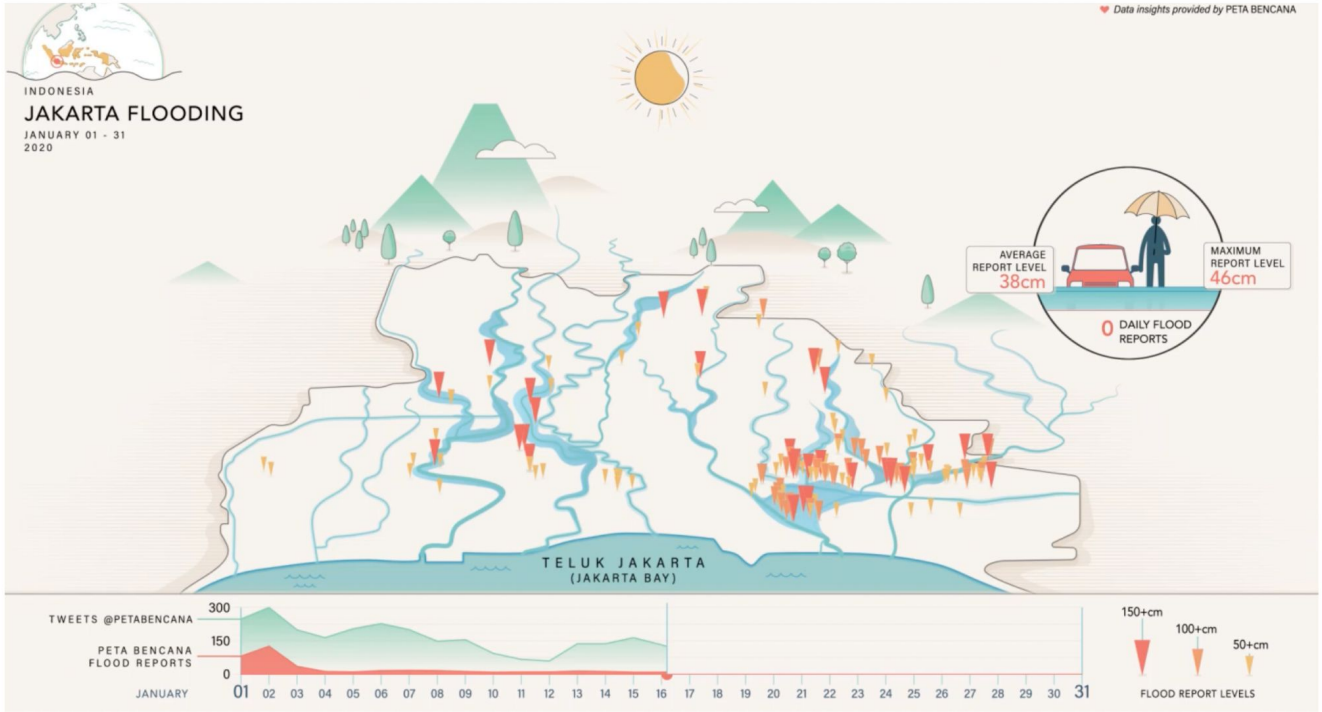

Figure 1: An overview of tasks in LegalAI.

Ethical considerations

Multi-lingual datasets

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In ACL 2020.

# NLP for social good (1/3)




https://developer.twitter.com/en/use-cases/build-for-good/extreme-weather/jakarta-flood

# NLP for social good (2/3)

## Fake news detection

| LEGITIMATE | FAKE |
|---|---|
| **Nintendo Switch game console to launch in March for $299** The Nintendo Switch video game console will sell for about $260 in Japan, starting March 3, the same date as its global rollout in the U.S. and Europe. The Japanese company promises the device will be packed with fun features of all its past machines and more. Nintendo is promising a more immersive, interactive experience with the Switch, including online playing and using the remote controller in games that don't require players to be constantly staring at a display. | **New Nintendo Switch game console to launch in March for $99** Nintendo plans a promotional roll out of it's new Nintendo switch game console. For a limited time, the console will roll out for an introductory price of $99. Nintendo promises to pack the new console with fun features not present in past machines. The new console contains new features such as motion detectors and immerse and interactive gaming. The new introductory price will be available for two months to show the public the new advances in gaming. |

Table 2: Sample legitimate and crowdsourced fake news in the Technology domain

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *COLING*.

# NLP for social good (3/3)

## Hate speech detection in Hindi-English code-mixed tweets

Yeh idiot log hamesha aise hi nuisance create karte rehte hain. → Hate

Maine aaj breakfast khayaa. → Non-hate

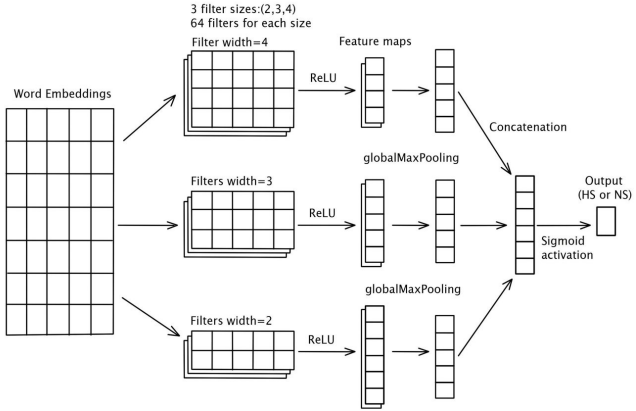
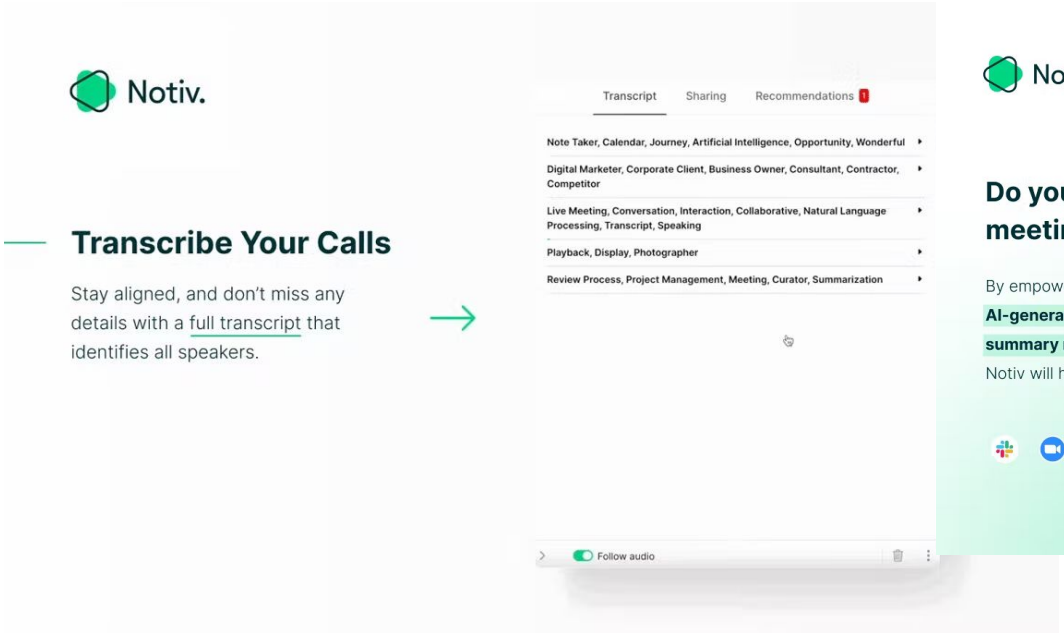
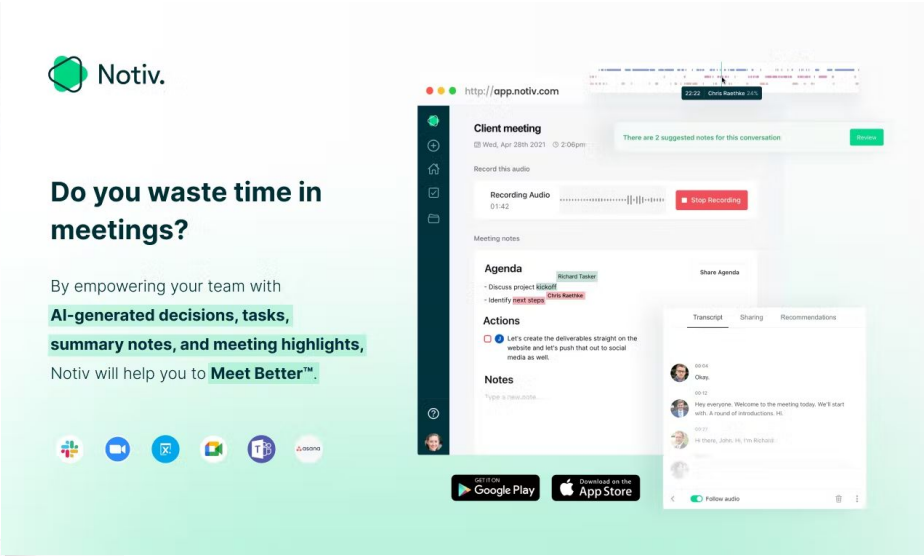



Figure 1: CNN model for hate speech detection.

Kamble, Satyajit, and Aditya Joshi. "Hate speech detection from code-mixed Hindi-english tweets using deep learning models." In ICON 2018.

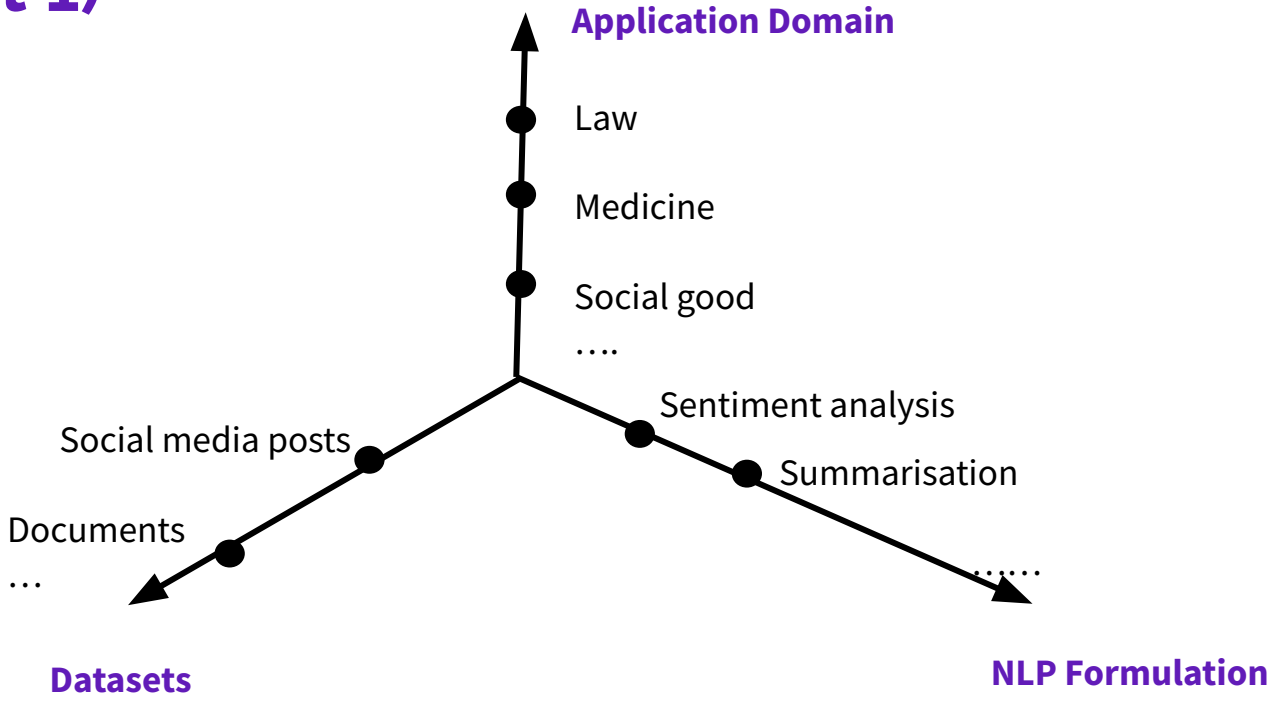
# NLP for productivity




https://www.producthunt.com/products/notiv

# Dimensions of NLP applications (i.e., summary of part 1)



# Applications of Natural Language Processing

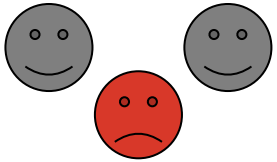
# Part II: Typical steps to build an NLP application

# Step 1: What is your idea?

Who will it help?

Which non-NLP experts do I need to consult?

Who can it harm?



End users

# Step 2: Where will the data come from?

Are labeled datasets available?

- Shared tasks, benchmarks

Can I create my own dataset?

- Download or create? (e.g. Twitter API)
- How will I label it?
- Can I make it publicly available for research?

# Step 3: Approach

What approaches exist?

**Modeling it as a NLP problem:**

- Classification/generation
- Identify related NLP problems (sentiment detection, information extraction)

Use the appropriate NLP perspective:

- Examine existing LMs to identify problems
- Language model (LM)-based training

# Step 3.1: NLP Perspective

## Examine pre-trained LMs

Examine how well existing LMs perform on biomedical summarisation.

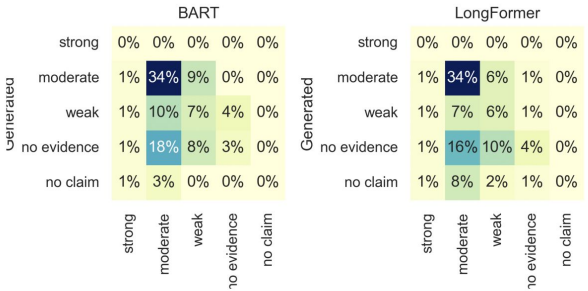



Figure 2: Modality of generated vs target summaries.

Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. 2022. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *ACL 2022*.

## LM-based training

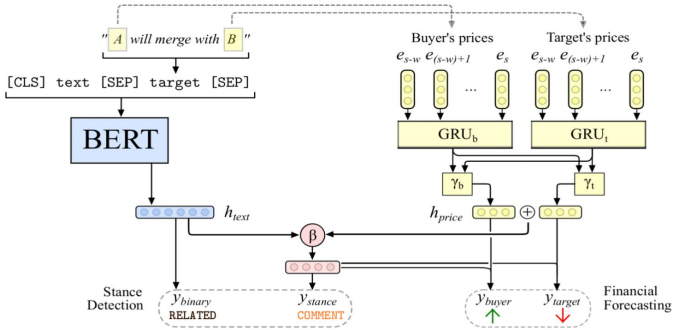



Figure 2: Overview of the proposed multi-task SDTF architecture. Price embeddings are not shown. Right, middle, and left components represent resp. textual, blended and financial signals. $\gamma$ is a multi-head attention mechanism, and $\beta$ is a bilinear transformation (Subsection 5.3).

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2022. Incorporating Stock Market Signals for Twitter Stance Detection. In ACL 2022.

# Step 4: Evaluation and Deployment

Dashboard

Qualitative & Quantitative Evaluation

Monitoring

# A typical workflow for an NLP application project (i.e., summary of part 2)


What is the idea?
Who will it help?
Who can it **harm**?
Experts from which discipline can be helpful?

Structured data
Unstructured data
Privacy concerns
Data labeling

Making a dataset available is a good research contribution!

Input/output view
Simple baseline
Incremental approach
Multi-perspective evaluation

# Applications of Natural Language Processing

## Part III: Our story of an NLP application

(Optional; will take 10 minutes)

# Automatic Prediction of Drunk-Texting

Aditya Joshi, Abhijit Mishra, A. R. Balamurali, Pushpak Bhattacharyya, and Mark Carman. "A Computational Approach to Automatic Prediction of Drunk-Texting." In ACL 2015.

# Goal

Concept: What are we trying to solve?

We wish to *automatically identify if it was written under the influence of alcohol.*

Our work presents the **first quantitative evidence that text contains signals that can be exploited to detect drunk-texting.**

# Motivation

Concept: Who can it help?

Alcohol abuse may lead to unsociable behaviour such as aggression (Bushman and Cooper, 1990), crime (Carpenter, 2007), suicide attempts (Merrill et al., 1992), drunk driving (Loomis and West, 1958), and risky sexual behaviour (Bryan et al., 2005) and privacy leaks.

Applications:

- Investigation following an alcohol abuse incident
- NLP-based techniques to prevent regretful drunk-texting

# Experiment Setup

Data: How can we get the data? .. and annotations?

as a challenge. We use **distant supervision using hashtags.**

- Two datasets were created:
  - Dataset 1 (2435 drunk, 762 sober): Tweets containing hashtags #drunk, etc, and those with #notdrunk, etc.
  - Dataset 2 (2435 drunk, 5644 sober): The drunk tweets are downloaded using drunk hashtags, as above. Other tweets by same users are sober tweets.

# Methodology

Approach: What is the corresponding NLP problem? What are the *current* techniques available?

a SVM-based classifier

| Feature | Description |
|---|---|
| **N-gram Features** | |
| Unigram & Bigram (Presence) | Boolean features indicating unigrams and bigrams |
| Unigram & Bigram (Count) | Real-valued features indicating unigrams and bigrams |
| **Stylistic Features** | |
| LDA unigrams (Presence/Count) | Boolean & real-valued features indicating unigrams from LDA |
| POS Ratio | Ratios of nouns, adjectives, adverbs in the tweet |
| #Named Entity Mentions | Number of named entity mentions |
| #Discourse Connectors | Number of discourse connectors |
| Spelling errors | Boolean feature indicating presence of spelling mistakes |
| Repeated characters | Boolean feature indicating whether a character is repeated three times consecutively |
| Capitalisation | Number of capital letters in the tweet |
| Length | Number of words |
| Emoticon (Presence/Count) | Boolean & real-valued features indicating unigrams |
| Sentiment Ratio | Positive and negative word ratios |

# Results: Which features work best?

Evaluation: How well is it performing?

| | (%) | (%) | (%) | NR (%) | PR (%) |
|---|---|---|---|---|---|
| **Dataset 1** | | | | | |
| N-gram | **85.5** | 72.8 | 88.8 | 63.4 | 92.5 |
| Stylistic | 75.6 | 32.5 | 76.2 | 3.2 | 98.6 |
| All | 85.4 | 71.9 | 89.1 | 64.6 | 91.9 |
| **Dataset 2** | | | | | |
| N-gram | 77.9 | 82.3 | 65.5 | 87.2 | 56.5 |
| Stylistic | 70.3 | 70.8 | 56.7 | 97.9 | 6.01 |
| All | **78.1** | 82.6 | 65.3 | 86.9 | 57.5 |

# A human evaluation

Evaluation: How well is it performing.. against humans?

nk texts on a held-out dataset.

Humans do it with an accuracy of 68.8%, our (automatic) classifiers do the same with 64% accuracy!

| | A (%) | NP (%) | PP (%) | NR (%) | PR (%) |
|---|---|---|---|---|---|
| Annotators | 68.8 | 71.7 | 61.7 | 83.9 | 43.5 |
| **Training Dataset** | **Our classifiers** | | | | |
| Dataset 1 | 47.3 | 70 | 40 | 26 | 81 |
| Dataset 2 | 64 | 70 | 53 | 72 | 50 |

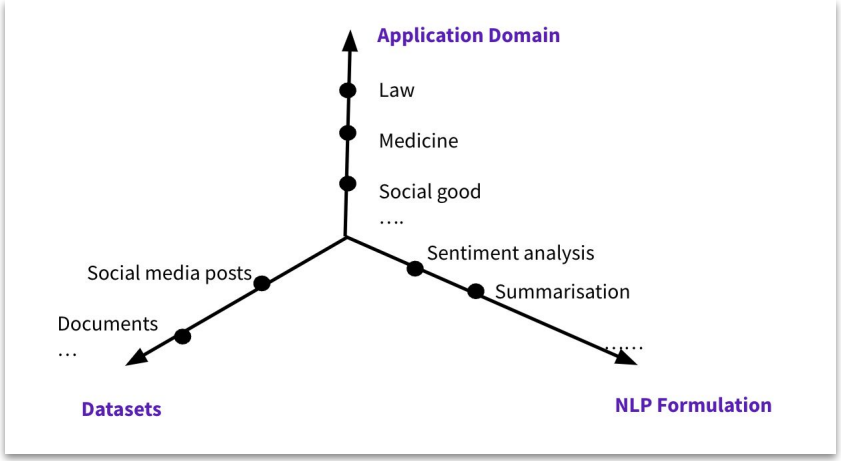
# Summary: Applications of NLP

Part I: Application Areas ***(Education, Literature, Medicine, Law, Social good, Productivity)***

Part 2: Typical steps to build an NLP application ***(Concept, Data, Approach & Evaluation)***

Part 3: Our story of an NLP application ***(Drunk texting prediction, Distant supervision, Human evaluation)***

# Thank you!




| Concept | Data | Approach & Evaluation |
| --- | --- | --- |
| Who will it help/**harm**? Experts from which discipline can be helpful? .. | Structured/ Unstructured data Privacy Data labeling .. | Input/output view Multi-perspective evaluation .. |

Aditya Joshi
aditya.m.joshi@gmail.com