Introduction to Text Analytics for Social Media Chatter

Nabanita Roy
Data Scientist @ ACI Worldwide
Education Lead @ Women in AI Ireland

PyCon SK 2022

AGENDA

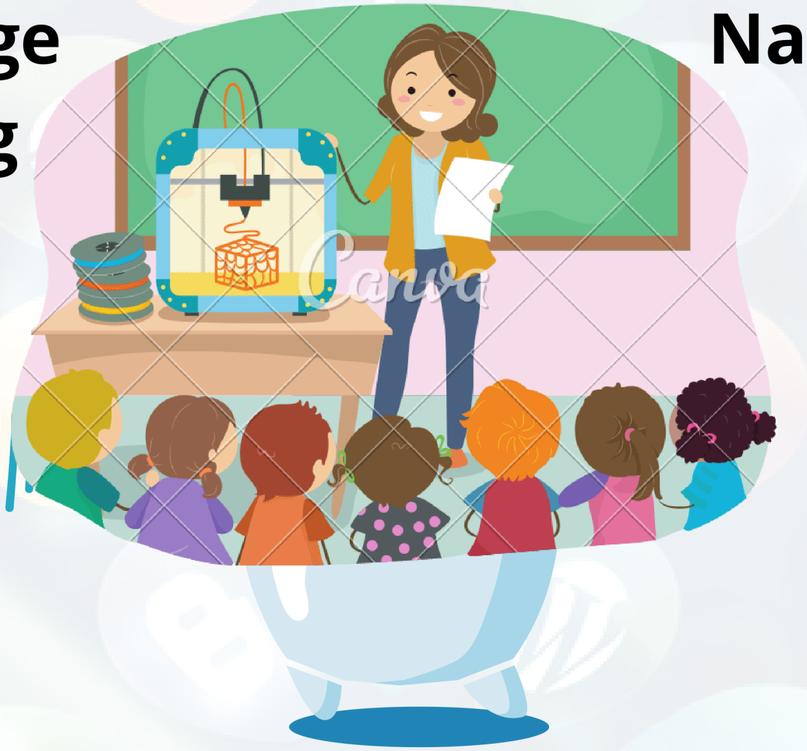
- Natural Language Processing
- The Social Media (SM)
- Applications of Social Media Analytics
- Traditional NLP vs NLP for SM
- NLP techniques for Social Media

What is Natural Language Processing?

- Process information in the form of natural language or texts
- Combination of Linguistics, Cognitive Science, Statistics, and many more...
- Enables machines to interact with humans like humans!

Training Machines to Listen, Comprehend and Speak

Natural Language Understanding



Natural Language Generation

Training Machines to Listen and Comprehend

Data Cleaning and Preprocessing

- Tokenization
- Normalization
- Stopwords Removal
- Letter Case
- Punctuations

Feature/Information Extraction

- Parts of Speech
- Entity Recognition
- Topic Modeling

Natural Language Understanding

Why Social Media Analysis?

- A plethora of data, increasing every second
- Individuals and organizations - conversations contain usable information
- It is a two-way street! Bi-directional communication is facilitated by Social Media

Types of Social Media Platforms

Follow

*not limited to...

Social networks

Blogs / Microblogs

Discussion Forums

Social bookmarks

Wiki Projects

Social News

Media Sharing

Application of Social Media Analysis

*not limited to

- All Industries
  - Business Intelligence
  - Recruitment
- Healthcare
  - Mental Health
  - Suicide Prevention
- Defense and security
- Media, Journalism, and Social News
- Marketing
- Finance
- Politics
- Environmental, Social & Governance (ESG)
  - Disaster Management and Recovery
  - Policy Analysis

NLP is Hard

- Segmentation
- Idioms and phrases
  - *Happiness is a piece of cake*
  - *I called her and gave her a piece of my mind*
- Synonyms
  - *Ecstasy is a piece of cake.*
- Homonyms
  - *A haunting piece of music.*
- Homophones
  - *Happiness is a peace of cake.*
- World Knowledge
- Multiple languages with diverse syntax
- Contractions and more...

Traditionally, NLP techniques assumed the texts are:

- Grammatically correct
- Standard syntax
- Monolingual
- Correct Spellings
- Used formal language

NLP for Social Media is Harder

- **Informal language**
  - Tokenization won't work
  - Inconsistent punctuations and Cases
  - Spell Correctors needed
- **Diverse vocabulary**
  - Stemming/ Lemmatization won't work
  - Neologism
  - Phonetic spellings
- **A new way of expressions**
  - Hashtags, entity names, emoticons
  - text-embedded media
- **Transliteration**
- **Different Platform**
  - Different Parser
  - Different ways of expressions

Traditional NLP

- Any other language
- Any other symbol

- English / target language
- Spaces and common punctuations for segmentation

Noise Content

Social Media vs Formal Texts

NLP for Social Media is Harder

- **Informal language**
  - Tokenization won't work
  - Inconsistent punctuations and Cases
  - Spell Correctors needed
- **Diverse vocabulary**
  - Stemming/ Lemmatization won't work
  - Neologism
  - Phonetic spellings
- **A new way of expressions**
  - Hashtags, entity names, emoticons
  - text-embedded media
- **Transliteration**
- **Different Platform**
  - Different Parser
  - Different ways of expressions

Conversational Language

Social Interaction and networks

Demographic Information

Inexpensive and Readily Available

Training Machines to Listen and Comprehend

Data Cleaning and Preprocessing

- Tokenization
- Normalization
- Stopwords Removal
- Letter Case
- Punctuations

Feature/Information Extraction

- Parts of Speech
- Entity Recognition
- Topic Modeling

Natural Language Understanding

Training Machines to Listen and Comprehend Social Media Data

Data Cleaning and Preprocessing

- Appropriate tokenization
- Stopwords
- Noise
  - Accents
  - Emoticons
  - HTML Tags
- Contractions
- Spell Check
- Language Detection
  - Translation, if required
- Letter Case
- Normalization
  - Stemming
  - Lemmatization
- Punctuations
- Handle tags and new words

Natural Language Understanding

NLP Tasks for Social Media

***not limited to**

- Classification and Clustering
  - Behavior analysis
  - Customer Segmentation
  - Topic Modeling
  - Opinion mining - Sentiment, Emotion, Sarcasm
  - Event detection
  - User Modeling
  - Trends Identification
- Information Extraction
  - Named Entity Recognition
  - Parts of Speech Tagging
  - Relationship extraction
- Language Identification and Translation
- Text Summarization
- Content filtering
  - Hate speech, violence, harassment
  - Spams and scams
  - Inappropriate content
  - Rumour Detection

Summary / Conclusion

- NLP for Social Media has gained a lot of focus due to the availability of rich, real-time kaleidoscopic data
- Data availability Annotations 😨 Labels 😞 Relevant Content 😫
- A great proxy for mental health studies and behavioral analytics
- Not everything is an ML model, some are rule-based / syntax-based operations
- No standard way of analyzing text data - You need to understand what's there and design your pipeline accordingly

Resources for this Talk

- **Slides**: https://github.com/royn5618/Talks_Resources/tree/main/PyConSlovakia2022
- **Sample Code**: https://github.com/royn5618/Medium_Blog_Codes/tree/master/EarthDayBlog
- **A Beginner-friendly Blog on SM Data Preparation & Analysis**: https://nroy0110.medium.com/earthday2022-tweet-analysis-and-visualization-using-pattern-hugging-face-and-plotly-1a0f54aa0b59

References

- http://spark-public.s3.amazonaws.com/nlp/slides/textprocessingboth.pdf
- https://www.morganclaypool.com/doi/10.2200/S00809ED2V01Y201710HLT038
- https://arxiv.org/pdf/2201.09451.pdf
- https://media.dlib.indiana.edu/media_objects/hh63sw18c
- https://www.youtube.com/watch?v=vFu7EFfC-Xc
- https://cse.iitkgp.ac.in/~pawang/courses/SC15/Lecture1_mc.pdf

Useful Python Libraries

- NLTK - https://tweetnlp.org/get-started/
- SpaCy - https://spacy.io/
- Gensim - https://pypi.org/project/gensim/
- Pattern - https://github.com/clips/paattern
- TweetNLP - https://tweetnlp.org/get-started/

Thank You

Q/A

PyCon SK 2022