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)
- NLP for Social Media
- Traditional NLP vs NLP for SM
- Accessing SM Data
- Data Preprocessing
- Mining and NLU for SM Data
- Examples

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!

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

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

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 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

Getting Data from Social Media

- Twitter API and wrappers around it (TwitterNLP, Pattern)
- YouTube Data API
- Web Scraping

- Required - Login or API Key
- Some of them are paid APIs

```
from pattern.web import Twitter
import pandas as pd
import json
import time
import os
import datetime
```

Download Tweets containing keywords: earthday or earth day

```
TWITTER = Twitter(language='en')
```

```
RANGE_X = 1
RANGE_Y = 4
COUNT = 100
DATA_FOLDER_NAME = 'data_' + datetime.datetime.now().strftime("%d_%m_%Y_%H_%M_%S")
DATA_FOLDER_NAME
```

```
os.makedirs(DATA_FOLDER_NAME, exist_ok=True)
```

```
for i in range(RANGE_X, RANGE_Y):
    print("Scraping {}-th data".format(i))
    list_tweets = [tweet for tweet in TWITTER.search('"earthday" OR "earth day"', start=i, count=COUNT)]
    print("Number of Tweets Scraped: {}".format(len(list_tweets)))
    list_text = [(each_tweet.text, each_tweet.date) for each_tweet in list_tweets]
    df = pd.DataFrame(list_text)
    df.to_pickle(DATA_FOLDER_NAME + '/data_twitter_' + str(i) + '.pkl')
    time.sleep(10)
    print("Finished Scraping {}-th Data".format(i))
```

NLP Techniques for Social Media Data

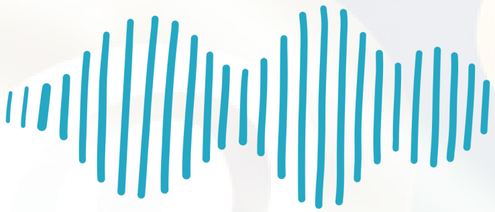
Preprocessing

- Appropriate tokenization
- Stopwords
- Noise
  - Accents
  - Emoticons
  - HTML Tags
- Contractions
- Language Detection
  - Translation, if required
- Letter Case
- Normalization
  - Stemming
  - Lemmatization
- Punctuations

NLP Techniques for Social Media Data

Natural Language Understanding

Natural Language Understanding

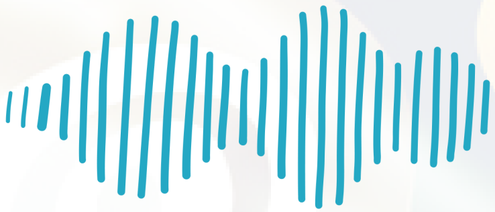


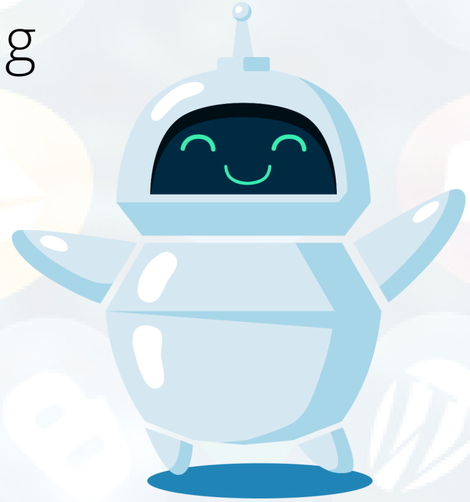
Natural Language Generation

NLP Techniques for Social Media Data

Natural Language Understanding

Natural Language Understanding





- Sentiment or Opinion
- Parts of Speech
- Entity Recognition
- Relationship Extraction
- Topic Modeling

Summary