Text Analysis

Inhoudsopgave

Wat is Text analyse?

- Txt = heel vaak ongestructureerde data, met veel waardevolle info in.
- Text Analysis(TA) = automatische proces van txt data te extraheren & classificeren.
- Soorten:
  - Sentiment Analyse(SA).
  - Topic Analysis(TA).
  - Urgency Detection(UD).
  - Intent Categorization(IC).

Soorten

Sentiment Analyse.

- SA = proces om positieve & negatieve sentimenten in txt te onderscheiden.
- Vaak gebruikt in
  - Sociale data.
  - Merk reputatie meten.
  - Klanten beter begrijpen.

Types van Sentiment Analyse

- Focust op polariteit van txt(Positiev, Negatiev, Neutraal).
  - Kan ook verder gaan door te focussen op:
    - Emotie.
    - Dringendheid.
    - Intentie.

Graded Sentiment Analysis

- Geeft verschillende levels van positiviteit/negatieviteit aan:
  - Zeer positief.
  - Positief.
  - Neutraal.
  - Negatief.
  - Zeer negatief.
- Kan vaak geïnterpreteerd worden als 5-sterren rating systeem.

Emotion Detection

- Geeft de emotie achter te txt terug.
- Gebruiken vaan lexicons(woord lijsten met geconeecteerde emotie) of zeer complexe machine learning algoritmes.
- Nadeel aan lexicons = dat vershillende mensen verschillende emoties aan bepaalde woorden zullen hangen.

Aspect-based Sentiment Analysis

- Geeft weer welk specifiek gedeelte in positiev/negatief besproken wordt.

Multilingual sentiment analysis

- = moeilijk <= vraagt veel preprocessing & resources.
- Kunnen ook eerst taal detecteren & dan pipen naar SA in juiste taal.

Waarom is Sentiment Analyse Belangrijk?

- Voordelen:
  - Kunnen grootte # data snel sorteren op sentiment.
  - Kunnen SA toepassen in RT => actie kan onmiddelijk genomen worden.
  - Krijgen consisten antwoord van SA i.v.m. als mensen dit zouden doen(negatief voor 1 persoon $\neq$ negattief voor andere persoon).

Hoe werkt SA?

- = opinion mining.
- Werkt a.d.h.v. *Natural Language Processing(NLP)* & *Machine Learning algorithmes(MLA's)* om automatisch de emotionele toon achter txt te bepalen.

Rule-Based

- Regels houden NLP techinieken in zoals:
  - Stemming.
  - Tokenization.
  - Part-of-speech tagging.
  - Parsing.
  - Lexicon.
- = naïve .
  - <= houden geen rekening met gecominbeerde woorden.
- Kunnen meer geavanceerde technieken gebruiken.
  - Nieuwe regels kunnen voorgaande resultaten impacteren.
  - Hele systeem kan zeer complex worden.
- Moeten constant worden ge-finetuned & onderhouden => regelmatige investeringen = nodig.

Automatisch

- Verwachten geen manuele regels, maar op ML technieken.
- = vaak gemodelleerd als klassiek classificatie probleem.

Training & Prediction Process

- In traininig process leert model *input* aan correcte *output tag* te binden.

- *Feauture extractor transfers* de txt *input* naar *feauture vector*.
- Paren van *feautre vectoren* & *tags* worden in MLA gegeven om model te generen.
- Model geeft voorspelde *tags*.

Classification agloritmes

- Naïve Bayes = familie van probalistische algoritmes, die het theorema van Bayes gebruiken voor categorie van txt te voorspellen.
- Lineaire Regressie gebruikt statistiek om waarde Y te voorspellen voor set van features X.
- Support Vector Machines(SVM) = ¬probalistisch model, die representaties van txt vb gebruikt als punten in een multidimensionele ruimte.
- Deep Learning(DL) = diverse set van algoritmes, die proberen het menselijke brein na te doen door een artificieël Neuraal Netwerk te gebruiken.s

Hybrid manier

- = mogelijk gewenste elementen van regels-gebaseerd en automatische systemen te kiezen en te mengen.
- Voordeel = dat resultaten vaak accurater =.

SA Challenges

- = 1 v/d. moeilijkste taken in NLP <= mensen kunnen het zelf amper.

Subjectiviteit & Toon

- = 2 soorten txt'en:
  - Objectief: ¬ bevatten expliciete sentimenten.
  - Surjectief: bevatten expliciete sentimenten.
- ¬ alle predicaten(adj, ww, zelf.nmw) moeten worden behandeld met zelfde level van respect als het aankomt op sentiment.

Context & Polariteit

- Machines hebben nood aan het expliciet zeggen v/d. context.
- Context kan polariteit veranderen.
- = nood aan veel preprocessing om [deel v/d.] context duidelijk te maken.

Ironie & Sarcasme

- In ironie/sarcasme geven mensen negatief gevoel weer door positieve woorden te gebruiken.
- Machines hebben het zeer moeilijk hiermee.

Vergelijkingen

- Context = cruciaal in achterhalen van sentiment achter vgl <= iets dat positief = in 1 context kan zeer negatief = in andere.

Emojis

- = 2 soorten emojis:
  - Westerse = geëncodeerd als 2 karakters.
  - Oosterse = langere combinaties van karakters met een verticale aard.
- Veel preprocessing = nodig