How to tackle unstructured data with data science?
Showcase of GenIDA example

Erik Novak
Jožef Stefan Institute
Ljubljana, Slovenia
Motivation

• Patient reported data can contain unstructured information
  • Text written by the writers own words

• Unstructured text contains valuable information which can help understand the patient’s disease

• Manual evaluation and analysis of patient reported data takes a lot of time and effort – especially if the data is in different languages

• Data science can speed up the analysis of patient reported data and find new patterns within
Outline

- GenIDA dataset
- Analysing unstructured data
  - Text representations
  - Text clustering
  - Cluster summarization
- Results
- Limitations
GenIDA Dataset

- Kleefstra syndrome subset
  - 155 unique users
  - 46 questions – 5 open questions, 41 categorical

<table>
<thead>
<tr>
<th>Question</th>
<th>Number of answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major Problem – Quality of life</td>
<td>147</td>
</tr>
<tr>
<td>Major Behavioural/Cognitive Problem</td>
<td>165</td>
</tr>
<tr>
<td>Problem of discomfort following medication</td>
<td>164</td>
</tr>
<tr>
<td>Major medical problems</td>
<td>152</td>
</tr>
<tr>
<td>Relative’s strengths</td>
<td>150</td>
</tr>
</tbody>
</table>
Analysing unstructured data

- Using machine learning to extract information from
- Analysis process
  - Text representation
  - Text clustering
  - Cluster summarization
Overview | Text Representations
Analysing unstructured data

TASK:
• Mapping text into a common semantic space where texts that have similar information are closer to each other
• Transforming text into something a computer can easily process – numbers, vectors, tensors, etc.

APPROACH:
Using language models to transform text into computer readable objects
Language Models | Text Representations
Analysing unstructured data

• In its essence, language models (LMs) are just predicting words in a blank. More formally, giving a context, an LM predicts the probability of a word occurring in the context

• The predictions are done by first transforming the text into vectors (mathematical objects), and then calculating the prediction

• The models we use predict the probability that two texts are similar or not - must support cross-lingual texts

• MODEL: sentence-transformers/distiluse-base-multilingual-cased-v1
Overview | Clustering
Analysing unstructured data

TASK:
• Grouping texts such that those that contain similar information are in the same group

APPROACH:
Using clustering algorithms - a subset of unsupervised learning models that enable grouping of data points
Kmeans | Clustering
Analysing unstructured data

• Different clustering algorithms were tested (DBSCAN, OPTICS, etc.)

• Ended up using Kmeans clustering – an algorithm that partitions data points into $k$ clusters, where the number $k$ is predetermined

• Simulating dataset bisection ($k=2$)
Results | Clustering
Analysing unstructured data

Stop cluster bisection, if:

1) Number of answers in cluster: less or equal 5 OR
2) Distance between answers in cluster: less of equal 0.18

<table>
<thead>
<tr>
<th>Question</th>
<th>Number of answers</th>
<th>Number of clusters</th>
<th>Average number of answers per cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major Problem – Quality of life</td>
<td>147</td>
<td>38</td>
<td>3.9</td>
</tr>
<tr>
<td>Major Behavioural/Cognitive Problem</td>
<td>165</td>
<td>52</td>
<td>3.2</td>
</tr>
<tr>
<td>Problem of discomfort following medication</td>
<td>164</td>
<td>22</td>
<td>7.4</td>
</tr>
<tr>
<td>Major medical problems</td>
<td>152</td>
<td>45</td>
<td>3.4</td>
</tr>
<tr>
<td>Relative’s strengths</td>
<td>150</td>
<td>41</td>
<td>3.7</td>
</tr>
</tbody>
</table>
Results (cont.) | Clustering

Analysing unstructured data

EXAMPLE: ANSWERS IN CLUSTER

• L'impossibilité de communiquer verbalement- La communication n'est pas facile, il semble tres souvent absent, difficulté pour imiter, deficience intellectuelle

• La communication est difficile, il semble tres souvent absent, il imite peu, que ce soit sons, gestes ou signes. Deficience intellectuelle

• I would have to say the lack of verbal communication effects her the most on a day to day basis.

• Defizite in der Kommunikation (Sprechen und Verstehen), daher auch kaum Beteiligung an Konversationen Schlechte Koordination von Bewegungen/motorische Unsicherheiten Offener Mund und ständiger Speichelfluss

• A compreensão das coisas, e não conseguir falar.

• Frustration with not being able to communicate clearly, and be understood.
Results (cont.) | Clustering
Analysing unstructured data

EXAMPLE: ANSWERS IN CLUSTER
• Intellectual disability
• Intellectual disability
• Cognitive disability - not understanding, language difficulties and inability to make friends
• Intellectual disability and speech impairment
• Kleefstra syndrome and moderat intellectual disability
• Mild intellectual disability
• Mild intellectual disability. Also, hypotonicity makes it difficult to partake in all sports.
• Learning disability, very poor speech.
Overview | Summarization
Analysing unstructured data

TASK:
Understanding what is are the commonalities of the answers found in the same cluster

APPROACH:
Using ChatGPT to generate the commonalities and the reasoning of the generation
ChatGPT | Summarization

Analysing unstructured data

ChatGPT is a generative language model, developed by OpenAI, with whom the user can “have a conversation”.

The model can perform various task (summarization, question answering, text formatting, etc.), but it is also prone to hallucinating – making facts up.
ChatGPT Prompt | Summarization
Analysing unstructured data

The “role” of ChatGPT – giving context
Take on the role of a medical expert.

The description of the data + the task that ChatGPT has to perform
From the list of user reported data found between ``, please provide what do the list items have in common in label format. Each list item starts with a minus sign (-). Each list item can be in a different language.

Additional tasks and restrictions ChatGPT has to follow
Additionally, provide an explanation for the label. Be specific with the explanation. The explanation must not be more than one sentence long. Both the label and explanation must be in English.

The format in which the generated answer has to be provided
Provide the answer in the following format:

COMMONALITY: {label}
EXPLANATION: {explanation}
```
```

Where the data is located
copy the answers here (each item in a separate line)
```
```

Jožef Stefan Institute
Department for Artificial Intelligence
Results | Summarization
Analysing unstructured data

EXAMPLE: ANSWERS IN CLUSTER

• L'impossibilité de communiquer verbalement- La communication n’est pas facile, il semble tres souvent absent, difficulté pour imiter, deficience intellectuelle

• La communication est difficile, il semble tres souvent absent, il imite peu, que ce soit sons, gestes ou signes. Deficience intellectuelle

• I would have to say the lack of verbal communication effects her the most on a day to day basis.

• Defizite in der Kommunikation (Sprechen und Verstehen), daher auch kaum Beteiligung an Konversationen Schlechte Koordination von Bewegungen/motorische Unsicherheiten Offener Mund und ständiger Speichelfluss

• A compreensão das coisas, e não conseguir falar.

• Frustration with not being able to communicate clearly, and be understood.

COMMONALITY:
Communication difficulties

EXPLANATION:
The individuals described in the list exhibit challenges in verbal communication, including the inability to speak or communicate clearly, limited use of gestures or signs, and frustration resulting from the lack of effective communication.
EXAMPLE: ANSWERS IN CLUSTER

• Intellectual disability
• Intellectual disability
• Cognitive disability - not understanding, language difficulties and inability to make friends
• Intellectual disability and speech impairment
• Kleefstra syndrome and moderate intellectual disability
• Mild intellectual disability
• Mild intellectual disability. Also, hypotonicity makes it difficult to partake in all sports.
• Learning disability, very poor speech.

COMMONALITY:

Intellectual disability

EXPLANATION:

All the listed items mention intellectual disability as a common symptom or characteristic of Kleefstra syndrome.
Limitations

• **Each answer can be only in one cluster.** Some answers contain multiple symptoms connected with different problems

• **Clustering algorithm stopping condition is based on heuristics**

• **Summarization is highly dependant on the cluster quality**

• **ChatGPT summarization can be too generic.** In cases where an answer contains multiple symptoms, the summarization can return an “umbrella” term covering all of the answers
Questions?

A question of the audience: Are we solving the right task?

Thank you!

Erik Novak
erik.novak@ijs.si