NLP project risk tool

Thomas Wood, Director: Fast Data Science Ltd

Guidance

This checklist can be used to apply your NLP or data science project to a rigorous and standardised review and provoke discussion. An NLP or data science project is considered a success if it completes and the insights gained from it, or the model developed in the project, result in a policy change, or a cost saving for the business. The reason for a risk assessment is to enhance the probability of the project being successful.

High risk indicators

Indicator	
The project is self-funded by a private individual or individuals rather than a company.	
The project is unrelated to the principal daily activities of the main stakeholder – in other words, the project is either a side project or a 'passion project'.	
The stakeholder in the project does not have executive authority in the organisation.	
The stakeholders are more than one organisation.	
The project is not connected to the main purpose of a department.	
There is scepticism or conflict in the organisation about the need for this project.	
There is not an abundance of data available (e.g., a daily influx of data from customer interactions), or data must be hand tagged.	
The project objective is open and not well-defined: just to explore the data.	
The stakeholder wants to disrupt an industry or field in which they have no experience.	
The client is a large organisation with a complex process of procurements, purchase orders, and approvals.	
The impetus for the project is to find a use case for a new and very over-hyped technology, rather than a business need.	
A small amount (<100) of text samples is available.	
Data needs to be classified into large numbers (100s) of categories.	
Remarks	

Medium risk indicators

Indicator	
The model needs to be retrained regularly.	
The NLP model must extract multiple values from text, such as dosages, addresses, drug names, chemical names.	
The text data must first be extracted from PDFs or similar.	
The text data is multilingual.	
The text data is poorly formatted (tweets, TTS transcripts, social media).	
There is a risk of AI bias.	
Remarks	