1. **NLP**
   * Pattern Discovery (Unsupervised Learning)
   * Prediction (Supervised Learning)
   * Benefits of text mining are (or will be) obvious especially in text-rich data environments
     + Email and spam filtering
     + Law (Court-orders)
     + Financial Disclosures
     + Medicine (discharge summaries, doctor notes)
     + Marketing (Customer comments and reviews)
     + Customer Support (In-bound help, problems with ordering, FAQs, etc.)
     + Survey research and analysis of open-ended results
     + Content Management
   * Stemming - Big, bigger, biggest = big
   * Lemmatization - drive, drove, driven = drive
2. **Text processing and Text Mining**

* Text -> format used for machine learning
* Document term matrix – count representation
* TFIDF – weighted count representation. Weighting scheme is to find words that occur in subset of document is a key word
  + More in subset document -> key word
  + In All document -> not important
  + As a result, it down weights those words, subset of more word gets higher weight.
* CV – counts the word and makes DTM. Row is document and column are word
  + Tokenize by space, removes single word
* Column represents vocabulary
  + Out of vocabulary
    - Sckitlearn removes it
    - Spacy – is all 0
* Threshold
  + max number of tokens to represent my vocabulary
  + how often those this word occurs
    - eg: 5 times and not more than 1000 times
* n-gram sequence is protected
  + retain context of these words

1. **Sentiment Analysis -** Feeling inside a document

* Dictionary/Lookup approaches
  + Tokenized representation lookup with score then sums - **Affin**
* Lookup with rules-based modifier
* Text Blob- normalized score between -1 and 1 – **SpacyTextBlob**
* Data annotation and hand-labeling is generally best for domain needs – annotations
* Vader learned on social media.

1. **Name Entity Recognition**
   * Start to extract information from text
     + Knowledge graph
   * Span of text to become part of category
   * Spacy has a generalized model so for domain specific we need domain specific annotations.
   * When tuned for specific problem we can help extract knowledge quickly vs human in the loop.
     + Healthcare to detect symptoms
     + Lawyer
     + Hiring
   * Custom NER model in spacy for domain specific model.
   * Extract knowledge from our core objects in corpus
2. **Embeddings**
   * New learned representation
   * Word2Vec
3. Topic Modelling
   * Top2Vec
   * Gensim – Doc2Vec and Tok2Vec
     + Train word and document vectors jointly
     + Doc2Vec – learn together
4. Rasa – Conversational AI
   * Chatbox
   * HR – on boarding
5. Universal Sentence encoder
6. Hugging Face
7. Transformer
   * fine tune Bert
   * CNN

**CLINICAL TRIAL MATCHING**

**Business Problem:**

One of the barriers faced in the development of a drug is the identification and recruitment of suitable candidates who meet the research criteria. This makes the recruitment process one of the most time-consuming and expensive steps of a trial.

Traditionally, clinical research coordinators had to manually look through medical records to identify an eligible candidate who fit the criteria for the study. The challenge to this approach is that it is time-consuming and requires a person with some background knowledge or context to perform this task. As a result, this manual form of recruitment and managing trials process slows the development of drugs and need require a large amount of capital.

NLP can make this process of clinical trial matching a seamless and automated process. It can be used to extract clinical data points to match complex clinical trial criteria from vast amounts of fragmented medical data that are stored in structured and unstructured form. Cutting the lengthy recruitment process will lead to potentially life-saving treatments being approved at a rate far quicker than would have been possible without the adoption of such techniques. Moreover, we can also create a recommendation engine that could help the patient enroll in the clinical trial themselves instead of relying on their doctors to inform them about suitable studies.

**Customer Reviews Analysis**

**Business Problem:**

In the era of digitalization, most companies have various sources of customer feedback, social media, call logs, mobile apps, to name a few. Customer feedback are important factor affecting other peoples’ purchases and booking. It’s one of the key indicators if the business can sustain in market for long period and greatest source for learning and improving a business. Therefore, analyzing such feedback to come up with actionable insights, is becoming essential for any business with an online presence.

NLP could play an important role analyzing and understanding customer reviews could help make their product better and increase their profit.

* Sentiment analysis could help cluster customers reviews in positive, negative and neutral categories.
* After that applying topic modelling could help understand the reasons behind those sentiments.

This will provide insight to the business on what to improve and what is going well in their product.

**HR – Resume Evaluation**

**Business Problem:**

One of the barriers faced in the recruitment process is to review hundreds or maybe thousands of job application for a single position. This makes the recruitment process slow and time consuming.

Traditionally, a recruiter or hiring manager will manually go through each resume and can spend hours filtering at the top of the funnel. The challenge to this approach is that it is time-consuming and requires a person with some background knowledge or context to perform this task.

NLP can make this process of recruitment a seamless and automated process. It can help to extract qualifications from the resume to match the job posting. Cutting the time-consuming resume scanning process will lead to potentially faster hiring and could save company a lot of money. Moreover, if it is ethically programmed and trained with unbiased data it could help to remove subconscious bias and increase diversity in candidates. Also, we could implement a recommendation engine which could automatically reach out to previous applicants.

Dataset has only 1000 records which is less for machine learning model. We can see this from the Junior Analyst results. Moreover, junior analyst didn’t do any train and test spilt so the result could actually be more optimistic.

To further explore I used spacy and UMAP to embed the reviews. Then used RandomForest and XGBoost with train test split of 75-25 percent (with stratify) and hyperparameter training. I got the same result as the junior analyst.

Looking at all these results we can’t deploy the model even with ~70% accuracy as it is only good for predicting rating 5 which is due to huge imbalance in data with ~70% of data is for rating 5. Recall and precision for rating 1,2,3 and 4 are 0 meaning our model is predicting every review to be rating 5. So, it is not good for predicting any other rating except 5.

I believe we need to collect more data for other categories before we can deploy any model and get rid of ratings from the platform.

**Junior Analyst Decision Tree:**

Table

Description automatically generated

**Random Forrest:**

Table

Description automatically generated

**XGBoost**:

Table

Description automatically generated