# Chapter 1

# Background

### 1.0.1 Natural Language Processing

Natural language processing (NLP) is the area of study in which 'natural' human language is examined via machine. Natural language refers to either spoken or written language, designed to be understandable to a human listener or reader. This language is not explicitly designed to be machine understandable, and machine comprehension of this language is a challenging problem [[CITE: cite: the challenges of NLP]].

NLP is a broad term covering many models and techniques to computationally extracting meaningful information from text, ranging from the simple extraction of individual words, to the extraction of deeper semantic meaning.

Early work in NLP focused around simple grammatical rules and small vocabularies, such as the work of Georgetown-IBM [[CITE: cite: John Hutchins. From first conception to first demonstration: the nascent years of machine translation, 1947–1954. a chronology. Machine Translation, 12(3):195–252, 1997.]] to translate 60 sentences from Russian to English in 1954. With the rapid increase in computational power and digital text corpuses, modern NLP has focused or deeper challenges of extracting meaning from text with tools such as Word2Vec [[CITE: cite: word to vec]] or deep learning methods such as Google's BERT [[CITE: cite: BERT]].

These methods face a daunting challenge, language is not only complex and often duplicitous, but contextual and ever-changing. [[TODO: end better]]

### Tokenisation

## 1.0.2 Information Theory

### Entropy

Entropy is a measure of the uncertainty of a random variable. In the context of information theory, this is defined by **??**, often refereed to as Shannon entropy, named after Claude Shannon for his work in 1948 studying the quantiles of information in transmitted messages [[TODO: cite Claude shannon 1948]]. The definitions hereafter are sourced from Elements of Information Theory by Thomas and Cover [[CITE: elements of information theory]]

**Definition 1.0.1** (Shannon Entropy)**.** Let $X$ be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x) = P(X = x), x \in \mathcal{X}$. The entropy $H(X)$ of the discrete random variable X, measured in bits, is

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \tag{1.1}$$

The entropy of the random variable is measured in bits. A bit can have two states, typically 0 or 1. The entropy of a random variable is the number of bits on average that is required to describe the random variable in question. To measure the entropy in bits, we use a logarithm of base 2, and all logarithms throughout this work are assumed to be in based 2, unless otherwise specified.

To give a typical example of entropy, if a fair coin is tossed there are two equally probable outcomes, giving an entropy of 1 bit. Further, we use the convention of $0 \log 0 = 0$, which sensibly means that adding a state with 0 probably to the random variable does not change it's entropy.

*Remark* (Suprise)*.* The entropy of the random variable X can also be described in terms of the expected surprise, where the surprise of a state is $\log \frac{1}{p(x)}$.

$$H(X) = \mathbb{E}\left[\frac{1}{p(x)}\right] \tag{1.2}$$

[[TODO: add some filler]]

**Lemma 1.0.1.** The entropy of a random variable is strictly non-negative, $H(X) \geq 0$.

*Proof.* $0 \leq p(x) \leq 1$ which implies that $log\frac{1}{p(x)} \geq 0$, hence the sum of products of strictly non-negative terms will always be non-negative. ■

[[TODO: add some filler]]

### Joint Entropy and Conditional Entropy

Above we worked with a single random variable. To extend this we introduce a second discrete random variable $Y$. Using this, we extend the one dimensional entropy to joint entropy.

**Definition 1.0.2** (Joint Entropy)**.** The joint entropy $H(X, Y)$ of a pair of discrete random variables $(X, Y)$ with a joint distribution $p(x, y)$ and state spaces $(\mathcal{X}, \mathcal{Y})$

$$H(X, Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y) \log p(x, y) \tag{1.3}$$

From our definition of entropy and the law of total probability we can create a notion of conditional entropy.

**Definition 1.0.3** (Conditional Entropy)**.** The conditional entropy $H(X|Y)$ of two discrete random variables $X$ and $Y$ is defined as,

$$\begin{aligned} H(X|Y) &= \sum_{y\in\mathcal{Y}} p(y)H(X|Y = y) && (1.4)\\ &= -\sum_{y\in\mathcal{Y}} p(y) \sum_{y\in\mathcal{Y}} p(x|y) \log p(x|y) && (1.5)\\ &= -\sum_{y\in\mathcal{Y}}\sum_{y\in\mathcal{Y}} p(x, y) \log p(x|y) && (1.6)\\ &= -E \log p(X|Y) && (1.7) \end{aligned}$$

Subtly different from the *conditional entropy* is the *cross entropy.* Whereas the *conditional entropy* is the amount of information needed to describe $X$ given the knowledge of $Y$, the *cross entropy* is the amount of information needed to describe $X$ given a optimal coding scheme built from $Y$.

**Definition 1.0.4** (Cross Entropy)**.** The cross entropy $H_\times(q|p)$ between two probability distributions, defined over the same state space, $p$ nd $q$ is defined as,

$$H(q||p) = -\sum_{x} p(x) \log q(x) \tag{1.8}$$

Although cross entropy has the common notation $H(X, Y)$, in this thesis we will use an alternative $H(X||Y)$, reminiscent if the Kullback–Leibler divergence in Equation 1.9 below, so as to not confuse the cross entropy with the above join entropy Equation 1.3 of the same notation.

*Remark.* Importantly, note that $H(X|Y) \neq H(Y|X)$ and $H(q||p) \neq H(q||p)$, both properties we will exploit later.

### Distances

We can extend these ideas to explore a notion of distance between probability distributions. Kullback–Leibler divergence is a measure of the inefficiency if one were to assume that a distribution is $p$ when the true distribution is $q$.

**Definition 1.0.5** (Kullback–Leibler divergence)**.** Kullback–Leibler divergence (also called relative entropy), $D(p\|q)$, between two probability distributions $p(x)$ and $q(x)$ is,

$$\begin{aligned} D(p\|q) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} && (1.9)\\ &= E_p \log\frac{p(X)}{q(X)} && (1.10)\end{aligned}$$

Again, we use the convention that $0\log\frac{0}{0}=0$ and $p\log\frac{p}{0}=\infty$.

Conveniently, we can also express the Kullback–Leibler divergence in terms of the cross entropy.

**Lemma 1.0.2.**

$$D(p\|q) = H_p(q) - H(p) \tag{1.11}$$

*Proof.*

$$\begin{aligned} D(p\|q) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} && (1.12)\\ &= \sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}} p(x)\log q(x) && (1.13)\\ &= -H(p) + H(q||p) && (1.14)\\ & && (1.15)\end{aligned}$$

■

The Kullback–Leibler divergence has two difficulties; It's not symmetrical and it can return infinite values. Jensen–Shannon divergence builds from the Kullback–Leibler divergence to solve these problems to a symmetric, finite comparison between probability distributions.

**Definition 1.0.6** (Jensen–Shannon divergence)**.** The Jensen–Shannon divergence between two probability distributions $p(x)$ and $q(x)$ is,

$$\mathrm{JSD}(p\|q) = \frac{1}{2}D(p\|m) + \frac{1}{2}D(q\|m) \tag{1.16}$$

using a mixture of the distributions, $m = \frac{1}{2}(p+q)$.

*Remark.* The square root of the Jensen–Shannon divergence provides a metric, often referred to as Jensen–Shannon distance.

[[TS: define metric?]]
[[TS: add mutual information]]
[[TS: add variation of information]]
[[TS: add diagram]]

### Process Entropy

Entropy rate of a stochastic process describes the amount of information required to describe the future state of a process, conditioned on the information in the history of the process.

**Definition 1.0.7** (Entropy Rate)**.** Let $\mathcal{X} = \{X_i\}$ be a stochastic ergodic process with a finite alphabet, where $X_i^j$ denotes a subsequence of the process $(X_i, X_{i+1}, \ldots, X_j)$. The entropy rate can be defined as,

$$H(\mathcal{X}) = \lim_{n\to\infty} H\left(X_n | X_{n-1}, X_{n-2}, \ldots, X_1\right) \tag{1.17}$$

Which, on the assumption of stationary, can be expressed as,

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H\left(X_1, X_2, \ldots, X_n\right) \tag{1.18}$$

While this notion of entropy rate provides a valuable theoretical tool, calculating it for real examples can prove difficult, and often impossible given data. In [[TODO: chapter two]] we will explore a method of estimating a similar quantity.

### Predictability

[[TODO: reword slightly]] Predictability is the probability $\pi$ that an theoretical predictive algorithm could predict the next state of a process correctly, this often be difficult to obtain. However, an upper bound, $\pi \leq \pi^{max}(S, N)$, is possible through the use of Fano's inequality **?**. For a process with $\pi^{max} = 0.3$, at best we could hope to predict this process correctly 30% of the time, no matter how good our predicative algorithm **?**.

[[TODO: change into a definition and proof]]

**Definition 1.0.8** (Maximal Predictability)**.** For a process $X$ with entropy $H(X)$, Fano's inequality in the context of our maximal predictability gives,

$$H(X) = H(\pi^{max}) + (1 - \pi^{max}) \log(|\mathcal{X}| - 1) \tag{1.19}$$

In order to find $H(\pi^{max})$ we use the binary entropy function **?**

$$H(\pi^{max}) = -\pi^{max}\log(\pi^{max}) - (1-\pi^{max})\log(1-\pi^{max}) \tag{1.20}$$

Which finally gives us a form that can be solved numerically for the fundamental limit of the process' predictability $\pi^{max}$

$$-H(X) = \pi^{max}\log(\pi^{max}) + (1-\pi^{max})\log(1-\pi^{max}) - (1-\pi^{max})\log(|\mathcal{X}|-1) \tag{1.21}$$

Throughout this thesis, maximal predictabilities will found by solving Equation 1.21 using the Powell's conjugate direction method, implemented in python using SciPy **?**, with a starting estimate for the root at $\pi^{max} = 0.5$.

We extend this notion of maximal predictability of a process, to our idea of cross predictability Definition 1.0.4

# Chapter 2

# The one with data and BOW

The main source of data for analysis is draw from the Twitter accounts of news-media organisations. In the 1950s, the widespread popularity of household television allowed TV broadcasting to become the primary tool for influencing public opinion in developed nations [[TODO: cite: Diggs-Brown, Barbara (2011) Strategic Public Relations: Audience Focused Practice p.48]]. This was a shift from a population that *listened* to radio news, to a population that *watched* news.

The even more rapid rise of mobile internet and social media sites in the last two decades has caused another shift. No longer just a population that watch news at fixed time, or read regularly scheduled newspapers; the conveniences of the modern developed world allow individuals to consume news anytime, anywhere. As of 2019, 55% of US adults get their news from social media either 'often' or 'sometimes' and 88% state that 'social media companies have at least some control over the mix of news people see' '**?**.

Given the importance of a free press and the role of social media sites in the delivery of news, this work aims to study the news on social media. To begin this task, we first define some common terms for clarity.

**Definition 2.0.1** (Social media)**.** The platforms used to consume information by individuals in the public. E.g. Twitter, Facebook, Reddit.

**Definition 2.0.2** (News-media)**.** The organisations that are producing information about a broad range of current events and sharing that information with the public.

**Definition 2.0.3** (News)**.** The *content* produced by news-media organisations.

## 2.1 Data

Using the media analysis source AllSides[1], a collection of news-media organisations was found. The purpose of AllSides is to provide a public analysis of political leanings of news sources [[TODO: cite: https://www.allsides.com/media-bias/media-bias-ratings]], and to aggregate news allowing consumers to view articles from different sides of the political spectrum. Each news source is labelled into one of 5 categories, Left, Lean Left, Center, Lean Right, or Right. Any news source the ratings are determined internally using 'blind surveys of people across the political spectrum, multi-partisan analysis, editorial reviews, third party data, and tens of thousands of user feedback ratings' [[TODO: cite: same as above]]. News sources are only assigned to a single category, but do have an attached confidence rating.

From the website, a list of possible news sources was collected on February 1st, 2019. In this collection was organisation names, political bias', the number of user feedback ratings of the political bias, and, if available, the twitter handles associated with those sources. These collected news sources were broad, containing not just news-media organisations but authors, pundits and think tanks.

To select an appropriate set of news-media organisation an examination and filtering process was undertaken. A source was only considered if it was a organisation (not an individual), that produced news content of a diverse range of topics. Many news sources were connected to think tanks or opinion groups, and only created news of a single topic or campaign. Further, if an news-media organisation has no twitter account or had less than 10,000 followers (a low bar in the social media world), then it was removed from the pool. This mainly removed inactive organisations and news organisation from small rural towns. Finally, a single sources was removed as it was not in English, and a single source was removed as it was the smaller sister site that had all content as a subset of it's larger site. The result of this filtering process is 174 news-media organisations and associated twitter accounts and categorised political bias'. A list of all news-media organisations under analysis can be seen in **?** and all removed sources and the removal justification can be seen in [[TODO: appendix of removals]].

Using the Twitter user handles associated with each of the news-media organisations, the history of all tweets for each account was collected using the Twitter application programming interface (API)[2]. Of interest in this work are the tweets each news organisation tweeted between January 1st,

[1] www.allsides.com

[2] https://developer.twitter.com/en/docs

2019 to January 1st, 2020.

Using this collection method a total of 3,221,769 tweets were collected from the 174 news-media organisation official Twitter accounts.

60054638 words

# Chapter 3

# The one with cross entropy

We can calculate...

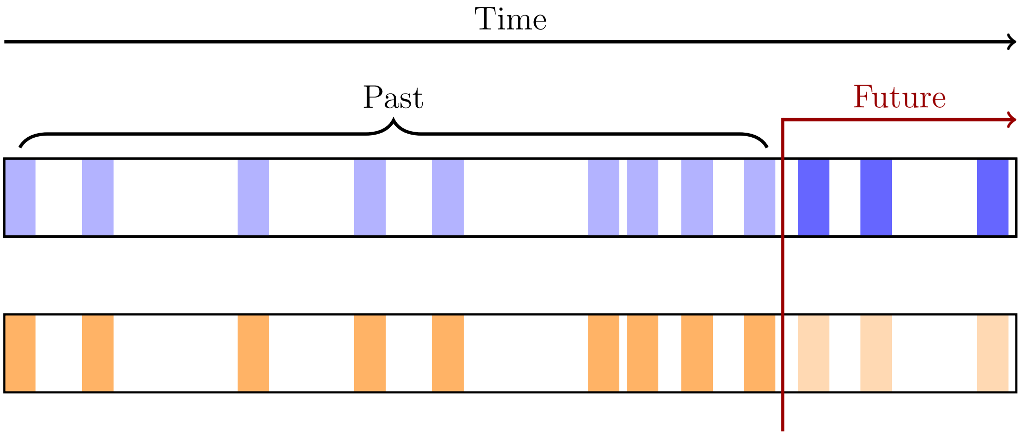
Recall Equation 1.17, which while a useful theoretical tool, can be very difficult to compute.

**Definition 3.0.1.** For a stochastic process $\mathcal{X} = \{X_i\}$, with a realisation of $n$ states and a finite alphabet, the entropy rate can be estimated using,

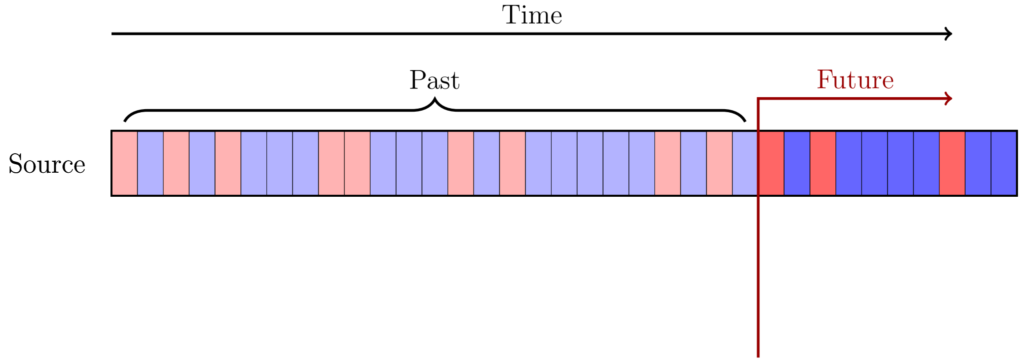
$$H(\mathcal{X}) \approx \frac{|\mathcal{X}| \log |\mathcal{X}|}{\sum_{i=0}^{n} \Lambda_i} \tag{3.1}$$

Where $|\mathcal{X}|$ is the size of the alphabet and $\Lambda_i$ is the length of the shortest subsequence starting at position $i$ that does not appear as a contiguous subsequence in the previous $i$ symbols $X_0^i$. This can also be obtained by adding 1 to the longest match-length,

$$\Lambda_i = 1 + \max \left\{ l : X_i^{i+l} = X_j^{j+l}, 0 \leq j \leq N - i, 0 \leq l \leq N - i - j \right\} \tag{3.2}$$



old



# Chapter 4

# What is this chapter about?

# Bibliography

# Bibliography