Chapter 1

Background

[[TODO: Introduction to background]] Information theory, what is it, why do we use it Networks, why Natural language processing

1.0.1 Information Theory

Entropy

Entropy is a measure of the uncertainty of a random variable. In the context of information theory, this is defined by Equation 1.1, 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,

$$D(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \tag{1.9}$$

$$= E_p \log \frac{p(X)}{q(X)} \tag{1.10}$$

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.

$$D(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \tag{1.12}$$

$$= \sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}} p(x)\log q(x) \tag{1.13}$$

$$= -H(p) + H(q||p) \tag{1.14}$$

$$\tag{1.15}$$

■

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 π 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,

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

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 π^{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 create a cross predictability using cross entropy Definition 1.0.4. [[TODO: more]]

1.0.2 Networks

1.0.3 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

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.

Definition 2.0.4 (Consumers)**.** Individuals in the public that willingly seek out and consume news from 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 an open analysis of political leanings of news sources [[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' [[CITE: same as above]]. News sources are only assigned to a single category, but do have an attached confidence rating that is provided from users selecting if they agree or disagree with the rating. An example of the ratings can be seen in Figure 2.1.

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 source 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 170 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 us-

[1] www.allsides.com

Figure 2.1: An example collection of News-Media sites that have been classified into biases by AllSides; sourced from [[CITE: https://www.allsides.com/media-bias/media-bias-ratings (as above)]]

ing the Twitter application programming interface (API)[2] and web-scraping tools [[CITE: twint]]. Of interest in this work are the tweets each news organisation tweeted between January 1st, 2019 to January 1st, 2020.

Each major news-media organisation will tweet pieces of news multiple times throughout the day. The manner in which each organisation does this can differ and no standard format is used. The tweets often come in the form of single line description of articles, alongside a link to an article on the news-media organisation's website. The primary purpose of using social media sites to post these stories is to drive traffic to the organisation's website, wherein they can earn revenue from ad impressions. As such, the format of such news tweets is to extract core concepts from articles and frame them in their most essential and appealing way; in essence, they are trying to create so-called 'click-bait' [[CITE: something tot do with clickbait]]. This format is desirable for our work as we want to explore how the language we use in news to appeal to consumers differs between organisations. This simplified format presents a reduced essence of this notion.

Twitter also serves another purpose for news-media organisations as a tool for breaking news. The modern 24 hour news cycle has had many effects on journalism, but chief among them is the need to produce breaking news at a lightning fast pace. The use of social media as a near instant tool for global public communication means that often no time can be wasted in publishing knowledge of a story, often while it is unfolding. Indeed, research has explored the role of Twitter for breaking news in the cases of the 2011 UK summer riots **?**, through activity providing real time updates over the four days, and in the case of the death of Osama Bin Laden in 2011 **?**, where the news was leaked and spread virally through social media before any news-media organisations could fully verify and publish stories on the claim. These breaking news stories are sometimes, but not always, precedes by expressions such as "*Breaking News:*". The inconsistent use of such a preamble can present challenge in our analysis of language moving forward, an [[TODO: will be explored more in this section]].

Using this collection method a total of 3,221,769 tweets were collected from the 170 news-media organisation official Twitter accounts in the 2019 calendar year. This represents an average number of tweets per day of above 50 tweets for each news organisation. In total, this appears a large useful corpus on text data for analysis. However, the activity level and consistency of output variety greatly between organisations. As can be seen in Figure 2.2, some news organisations produce very little content on average. This can be explained through two mechanisms.

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

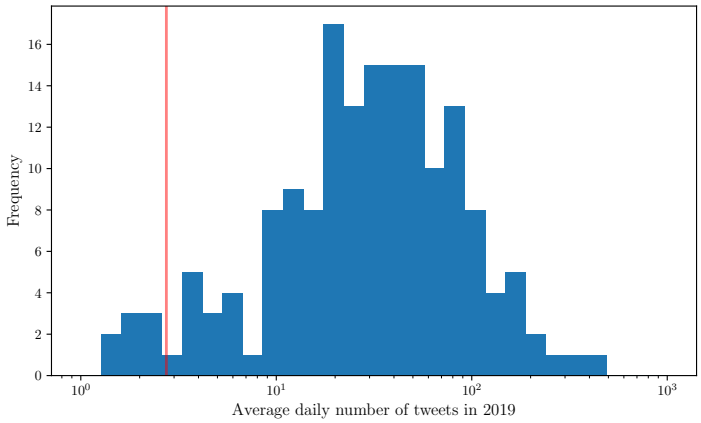


Figure 2.2: The average number of tweets produced each day during the 2019 calendar year for all 171 news-media organisations. The red line is the chosen threshold of 1000 tweets in the year, an average of 2.74 tweets per day.

Firstly, some organisations are not very active on social media. In particular, smaller organisations, which are typically less well resourced, place a lower priority on social media posting. This lowered tweet volume, presents a challenge for this work. In particular the use of bag of words tools and the non-parametric entropy estimator in chapter 3 require a substantial amount of text to reach meaningful results. As such, organisations that produced less that 1000 tweets in 2019 were removed from further consideration. This removed a total of 11 news-media organisations, listed in Table 2.1.

Secondly, an issue was identified in long periods of inactivity of a few organisations. Five news-media organisations, for reasons unknown had large periods of time in which they did not post any tweets. These periods of time, spanning a few months, present key issues to our investigation. Moving forward we will consider time an important aspect of news, especially in the context of breaking news, as such these organisations are not only at a disadvantage in this space but present an anomaly in our data that hinders our ability to extract meaningful results from them. As such, these five organisations, listed in Table 2.2 were removed from further consideration and analysis.

To further confirm the validity of the sources in terms of their activity level over time, we examine the isolated daily activity of each news-media organisation. An activity curve for the New York Times can be seen in Figure 2.3. A clear weekly trend, wherein tweet activity is decreased, but not zero, during the weekends can bee seen in the New York Time activity, but is emblematic of a general trend seen in most news-media organisation.

News-media Organisation	Bias	Number of tweets in 2019
RealClearPolitics	Center	532
IJR	Lean Right	777
WND News	Right	709
PRI	Center	346
EurekAlert!	Center	610
FAIR	Center	697
Crowdpac	Center	521
Inside Philanthropy	Center	781
Diplomatic Courier	Center	750
Peacock Panache	Left	198
Independent Voter	Center	303

Table 2.1: [[TODO:]]

News-media Organisation	Bias
American Thinker	Right
Pacific Standard	Lean Left
Philly.com	Lean Left
Splinter	Left
ThinkProgress	Left

Table 2.2: [[TODO:]]

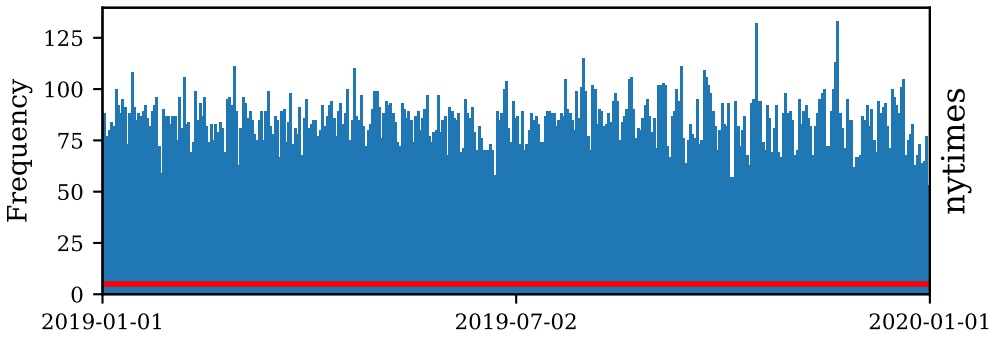


Figure 2.3: Twitter activity over 2019 for 'The New York Times'. Twitter handle is 'nytimes' with 44800317 followers and 31029 total tweets in 2019. A reference value of 5 tweets per day is shown in red. This is only one news-media organisation and all other activity figures for other organisations are available in **??**

Further, many news-media organisations have distinct spikes that occur a key points during the year. These spikes indicate an extreme news day, wherein an organisation is covering a rapidly evolving breaking news story, or responding to major changes in discourse through the day. These are interesting and important features in the data, an worth keeping in mind during further analysis. The full collection of figures containing the daily activity levels of all included organisations can be seen in **??**.

With this now activity-level cleaned data, our news-media organisations have a slightly higher average number of tweets per day of 52.97. With the total number of remaining tweets at 2,977,980.

As a result of this filtering we are left with 154 news-media organisations with complete data for the 2019 calendar year. Of these organisations the bias distribution is still somewhat representative of social media. [[TODO: find a source that discusses why left wing is more popular on social media.]] In total there are 73 organisations in the left half of the bias spectrum, 44 in the centre and 37 on the right; expanded on in Table 2.3. This distribution, although shifted towards the left, still provides ample sources for the effect of bias to be explored further in this work, with keen attention to the potential impact of the skewed distribution.

From the news-media Twitter accounts we can also access metadata about the organisation. Two useful such pieces of metadata are the geographic location, and the number of followers on twitter.

[[TODO: The geographic location can be posted byt he user blah blah

Bias	Number of Organisations
Left	31
Lean Left	42
Center	44
Lean Right	16
Right	21

Table 2.3: The number of news-media organisation in each political bias classification within our data.

Location	Counts
New York	34
Washington, D.C.	20
Cafilforna	11
Other US City	44
General US	8
Worldwide	6
United Kingdom	5
Qatar	1
Pakistan	1
Korea	1
Unspecified or Unclear	43

Table 2.4: [[TODO:]]

not always accurate,]]

[[TODO: Followers are important as they are the default mechanism through which people consumer the content]] Figure 2.4

390.0pt

60054638 words

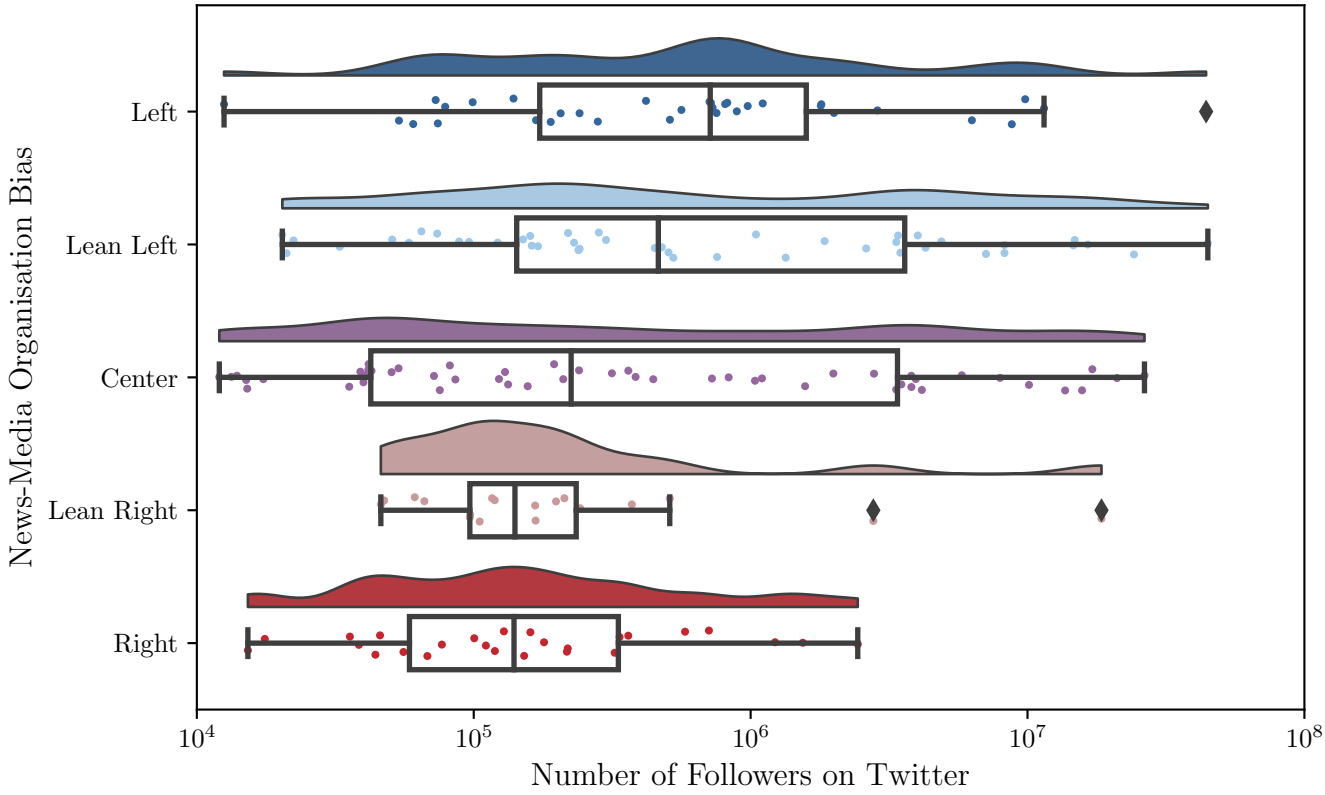


Figure 2.4: [[TODO:]]

Chapter 3

The one with cross entropy

We can calculate...

3.1 Entropy Rate

Recall Definition 1.0.7 of the entropy rate of a stochastic process, which while a useful theoretical tool, can be very difficult to compute. [[TODO: why is it hard to compute]]

To overcome this, we seek a way to estimate the entropy of the process from a known sequence of data. In 1998 Kontoyianni et al. proved the convergence of a non-parametric entropy estimator in stationary processes **?**.

Definition 3.1.1 (Kontoyianni Entropy Rate)**.** 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 Λ_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,

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

[[TODO: Outline of proof of convergence of this estimator]]

3.2 Assumption of Entropy Rate Estimation

[[TODO: A discussion and plots of entropy rate convergence]]
[[TODO: A comparison with other entropy rates]]

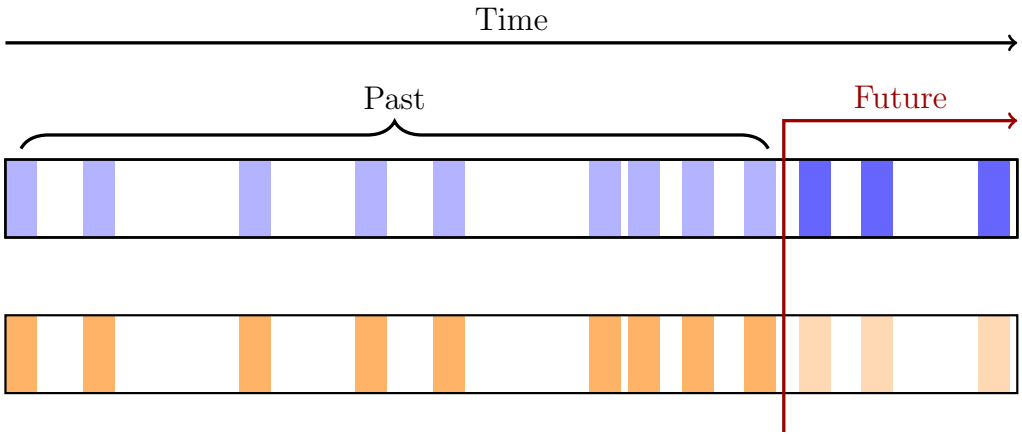
3.3 Cross Entropy Rate

Definition 3.3.1 (Cross Entropy Rate)**.** The cross entropy of a target process $\mathcal{T}$ coded from a source process $\mathcal{S}$ can be estimated,

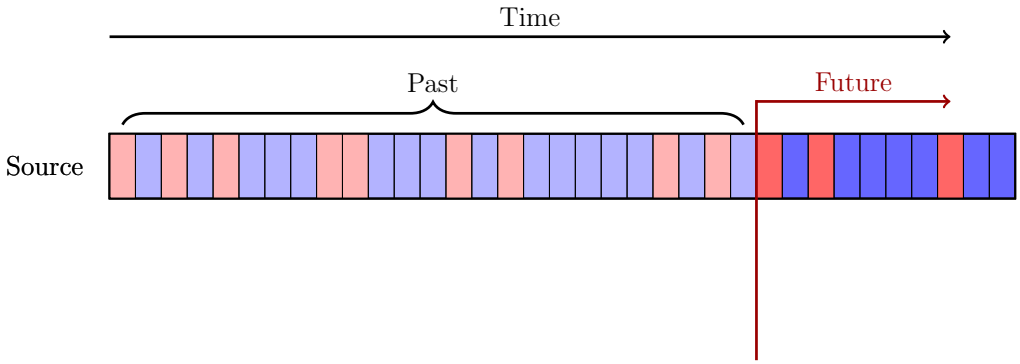
$$\hat{H}(\mathcal{T}||\mathcal{S}) = \frac{N_{\mathcal{T}} \log_2 N_{\mathcal{S}}}{\sum_{i=1}^{N_{\mathcal{T}}} \Lambda_i(\mathcal{T}|\mathcal{S})} \tag{3.3}$$

Where $\Lambda_i(\mathcal{T}|\mathcal{S})$ is given by the shortest subsequence starting at position i in target $\mathcal{T}$ that does not appear as a contiguous subsequence in the source $\mathcal{S}$.

$$\Lambda_i(\mathcal{T}|\mathcal{S}) = \max\left\{l : T_i^{i+l} = S_j^{j+l}, 0 \leq j \leq N_{\mathcal{S}} \leq l \leq \min(N_{\mathcal{S}} - j, N_{\mathcal{T}} - i)\right\} \tag{3.4}$$



old


Time
Past
Future
Source

Chapter 4

What is this chapter about?

Bibliography

Bibliography