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,

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

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)**.** The 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]]

Entropy Rates

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.

"The entropy rate is almost surely an asymptotic lower bound on the per-symbol description length when the process is losslessly encoded" [[CITE: Some asymptotic properties of entropy of a stationary ergodic data source with applications to data compression,]]

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 chapter 3 we will explore a method of estimating a similar quantity.

Predictability

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 [4]. 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 [17].

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,

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

The entropy of the maximal predictability $H(\pi^{max})$ is substituted with the binary entropy function [17],

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

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 [18], 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

[[TODO: go through Newman networks and state a bunch of definitions]]

1.0.3 Natural Language Processing

[[TS: introduce in here the notion that we're interested in text and NLP is a fundamental building block of understanding it]]

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 [3]. 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' '[13].

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 [6], 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' [6]. 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 **??**.

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] and web-scraping

[1] www.allsides.com

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

Figure 2.1: An example collection of News-Media sites that have been classified into biases; sourced from Allsides website [6]

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 [19], through activity providing real time updates over the four days, and in the case of the death of Osama Bin Laden in 2011 [8], 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.

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

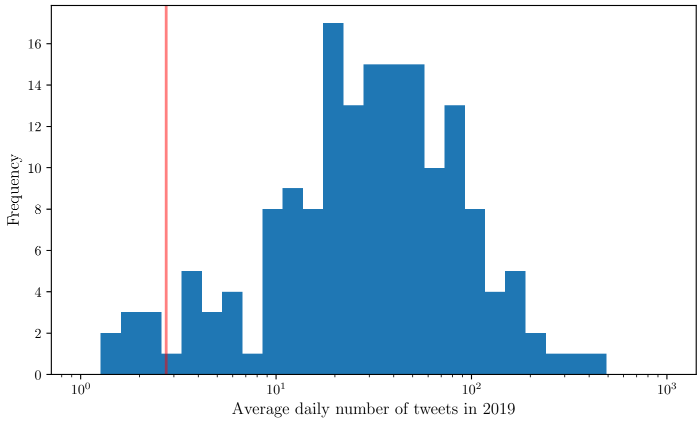


Figure 2.2: The average number of tweets produced each day during the 2019 calendar year for all 171 news-media organisations. The red line is the chosen threshold of 1000 tweets in the year, an average of 2.74 tweets per day.

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. 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,

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: Table of news-media organisations that were removed from data due to a low number of tweets in the 2019 calendar year.

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

Table 2.2: Table of news-media organisations that were removed from data due to long periods of inactivity.

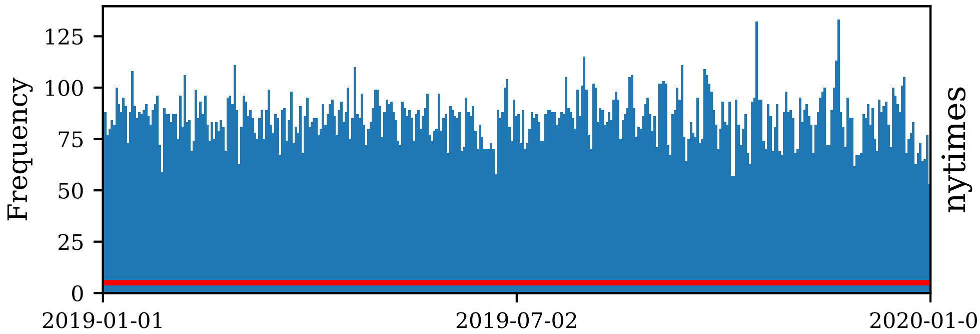


Figure 2.3: Twitter activity over 2019 for 'The New York Times'. Twitter handle is 'nytimes' with 44800317 followers and 31029 total tweets in 2019. A reference value of 5 tweets per day is shown in red. This is only one news-media organisation and all other activity figures for other organisations are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

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. [[CITE: 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.

On the Twiiter account of each news-media organisation they can elect to provide a text-based 'location'. In some cases this option can be used for other purposes, such for self promotion (e.g. the *New York Daily* states it's location as '`New York City / fb.com/nydailynews`') and many organisa-

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

Table 2.3: The number of news-media organisations in each political bias classification in 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: The aggregated self-defined locations of news-media organisation according to their Twitter account metadata.

tions have elected to leave the field blank. Of the organisation with locations, these can be difficult to disambiguate and compare. In situations were multiple cities or locations are defined, the largest possible inclusion was taken. For example '`New York and the World`' would become 'Worldwide' in our classification, as would '`NYC, London, Paris, Hong Kong`'. These classifications were done manually due to the complexity of possible locations, and is summarised in Table 2.4.

The number of followers a news-media organisation has on Twitter is an important metric, as it underpins the default mechanism through which people consume the news content produced. Although other algorithm newsfeed behaviours provide an important effect, the follower count plays a role in such a feed and is an effective metric of 'success' in the eyes of the news-media organisations.

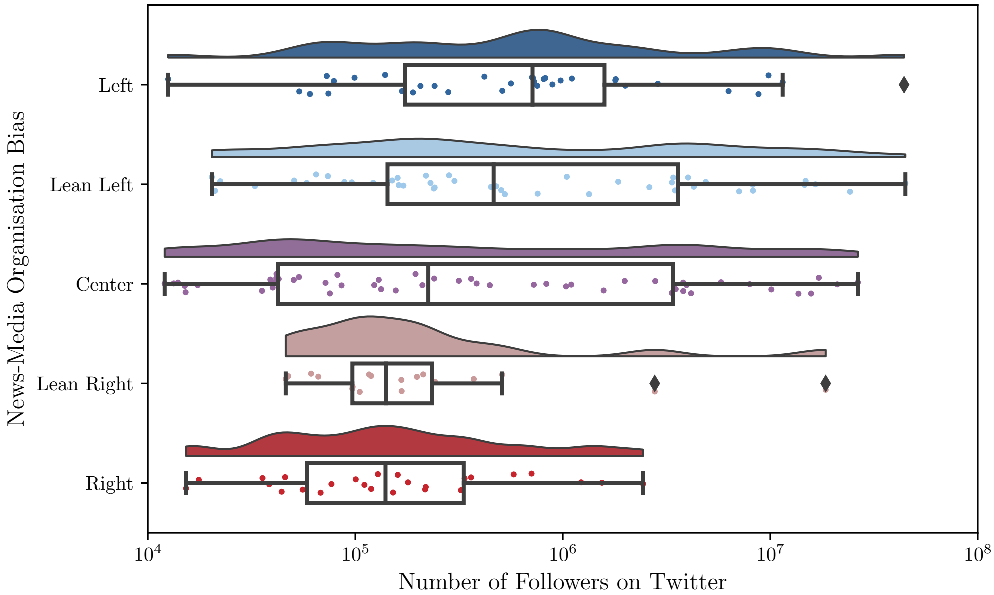


Figure 2.4: Number of followers on Twitter of news-media organisations included in the data. Grouping is according to Allsides bias.

The most followed organisation in the data is *The New York Times* with 44800317 accounts following at the time of collection in [[TODO: time of collection]]. The least followed account included in the data is [[TODO: least followed account]]. Interestingly, the follower counts are slightly higher for left biased organisations than for the right, in addition to be more numerous. This is shown via the followers distributions for each bias in Figure 2.4, which is indicative of the larger trend in social media of slightly left leaning demographics [**?**].

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 [11].

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,

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

This idea of using matched subsequences of text draws from the original work by Lempel and Ziv [21] in compression algorithms based on coding schemes. These algorithms attempt to compress a sequence down into the smallest possible representation, which at perfect efficiency would be the entropy, H. However these universal coding algorithms have no universal rate of convergence [16, 14] and in practice other approaches are often employed, tailored to the specific application at hand.

The idea of an entropy estimator based on match lengths was originally put forward by Grassberger [7] and proved consistent for independent and identically distributed (i.i.d.) processes and mixing Markov chains [15], stationary processes [10] and more generally to random fields [12].

Wyner and Ziv [20] showed that for every ergodic process the match length Λ_n grows like $\frac{\log n}{H}$ in probability.

Extending from this notion Kontoyianni et al. showed the convergence of Equation 3.1 in stationary ergodic processes using the match-length Λ_i. This match-length in Equation 3.2 can be seen as the length of the next phrase to be encoded in the sliding-window Lempel–Ziv algorithm.

[[TODO: talk more about how this match length works (diagram)]]

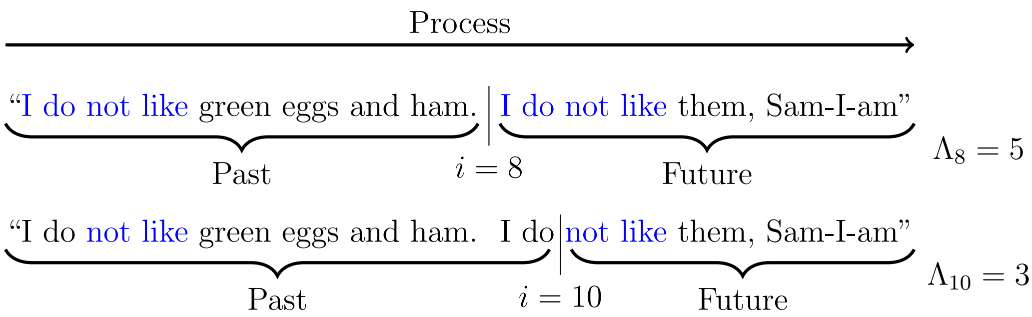


Figure 3.1: An example calculation of the match-length based Λ_i applied to a line from Green Eggs and Ham by Doctor Seuss. Blue text is that which has been matched from past to the future.

Even before it's formalisation by Kontoyianni et al, similar estimators had appeared in the literature applied to experimental data to determine the entropy rates of processes [1, 2, 5, 9].

Moving forward we will assume any any discussion of the *entropy rate* of a single process is assumed to be the *Kontoyianni entropy rate* of that process, unless otherwise stated.

3.2 Assumptions 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

Similar to the extension of entropy to cross entropy in Definition 1.0.4, we can generalise our notion of Kontoyianni entropy rate in Definition 3.1.1 to a *cross* entropy rate which we will call the Kontoyianni cross entropy rate.

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

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

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 $\mathcal{S}$.

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

where T_a^b and S_a^b are continuous subsequences starting from index a to index b of the target, $\mathcal{T}$, and source, $\mathcal{S}$, processes respectively.

This approach is matching segments of text in the target to segments of text anywhere in the source in the same manner that the Kontoyianni entropy rate matched segments of text in the future of a process from a index to the history before the index. In essence, the entropy rate was asking how much information was needed on average to describe the future of a source from it's past. Comparatively, the full cross entropy rate above is taking asking how much information was needed on average to describe the target given knowledge of the entire lifetime of the source. [[TS: Not sure how long I should keep up the colouring for.]]

In the context of our problem, this estimator is cheating by viewing the future of news through the lens of the source process. As [[TODO: ref]] illustrates, the text subsequence match in the target could be drawn from future time points in the source. This approach is interesting as it asks a simpler but very relevant question; how different are the two sources from an full information perspective.

[[TODO: tikz of how the full entropy works]]

While this question is interesting, and will be explored further, we need to refine this definition to help answer our question of information flow.

Rather than looking at the entire lifetime of the source during a matching calculations, we can reduce our search space to the text that occurred in the *past* of the source. To achieve this we use an important piece of our data, the time that tweets occurred. For each word in the target process, T_i has an associated time with it, $t(T_i)$. When matching the future of $\mathcal{T}$, starting from an index i, we can reduce the source process, $\mathcal{S}$ to only the words that were themselves tweeted before time $t(T_i)$.

Put simply, we can alter the Kontoyianni full cross entropy rate to a time-synced cross entropy rate by replace the full process, $\mathcal{S}$, with a time reduce source process $\mathcal{S}_{\leq t(T_i)}$. This can be seen visually in [[TODO: ref tikz]] and is formally defined as follows.

Definition 3.3.2 (Kontoyianni Time-synced Cross Entropy Rate)**.** The time-synced cross entropy rate of a target process $\mathcal{T}$ coded from a source process $\mathcal{S}$ can be estimated via,

$$\hat{H}(\mathcal{T}||\mathcal{S}) = \frac{N_{\mathcal{T}} \log_2 N_{\mathcal{S}}}{\sum_{i=1}^{N_{\mathcal{T}}} \Lambda_i(\mathcal{T}|\mathcal{S}_{\leq t(T_i)})} \tag{3.5}$$

Where $\Lambda_i(\mathcal{T}|\mathcal{S}_{\leq t(T_i)})$ is given by the shortest subsequence starting at position i in target $\mathcal{T}$ that does not appear as a contiguous subsequence in the time reduced source $\mathcal{S}_{\leq t(T_i)}$ where,

$$\mathcal{S}_{\leq t(T_i)} = \{S_j | t(S_j) \leq t(T_i) \forall i\}. \tag{3.6}$$

Which gives,

$$\begin{aligned}\Lambda_i(\mathcal{T}|\mathcal{S}_{\leq t(T_i)}) = \max\{l : T_i^{i+l} = S_j^{j+l}, 0 \leq j \leq N_{\mathcal{S}},\\ 0 \leq l \leq \min(N_{\mathcal{S}} - j, N_{\mathcal{T}} - i)\},\end{aligned}$$

where T_a^b and S_a^b are continuous subsequences starting from index a to index b of the target, $\mathcal{T}$ process, and the time reduced source, $\mathcal{S}_{\leq t(T_i)}$, respectively.

This time-synced entropy rate is testing not just the differences in the language processes of the source and target, but also measuring what information in the target is present in the source's history. This is an important distinction, as it allows us to probe a very important aspect of our data, namely, the time in which news is created.

If a piece of information appears earlier in the source than in the target, it will be detected during the match length search, resulting in a lower entropy. This is to say, in the context of news, if the source breaks a story first, *less* information is required to describe the subsequent news output from the target.

Conversely, if a target produces a piece of information before the source, then that information will not appear in the history of the time-synced source during the match-length search. This will result in lower values of Λ_i for that piece of information, which raises the cross entropy rate.

From this, we can find that, on average, if a source produces information earlier than a target, the cross entropy rate, $\hat{H}(\mathcal{T}||\mathcal{S})$, will be lower than if the target produces information earlier than the source. This method of examining who produces information first can be extended into a notion of *information flow*, a discussion we will leave for subsection 3.3.2.

[[TODO: tikz of how the time synced entropy works]]

[[TODO: we can extend this with predictability]]

3.3.1 Validating the Assumptions Cross Entropy Estimation

legth vs cross entropy

3.3.2 Information Flow

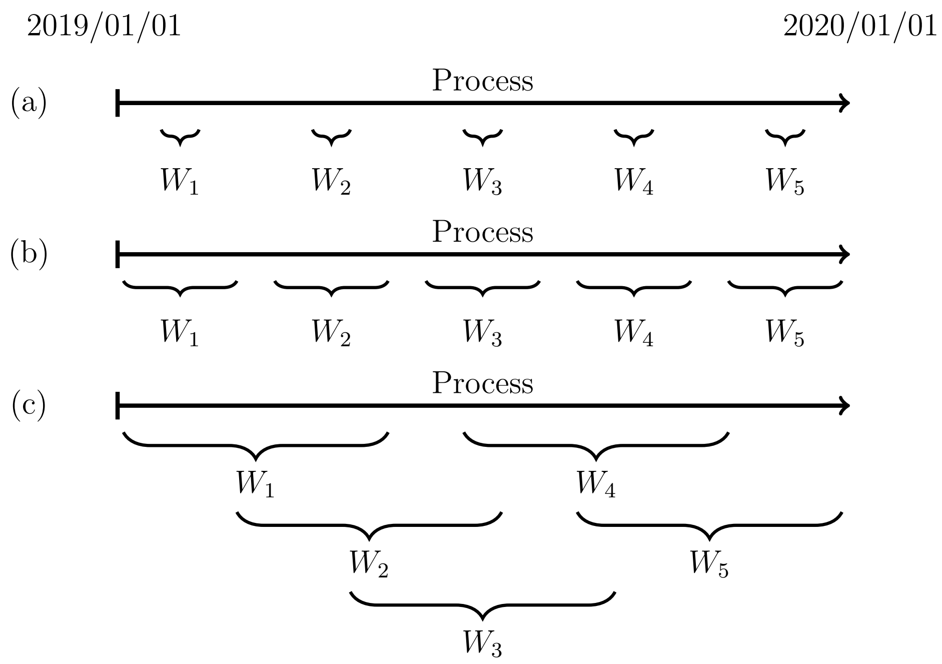
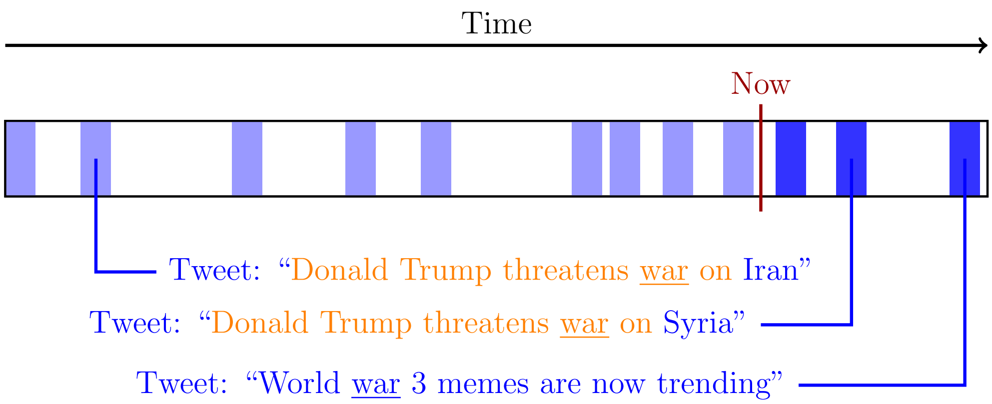


Figure 3.2: [[TODO:]]



Bibliography

[1] S. Chen and J. H. Reif. Using difficulty of prediction to decrease computation: Fast sort, priority queue and convex hull on entropy bounded inputs. In *Proceedings of 1993 IEEE 34th Annual Foundations of Computer Science*, pages 104–112. IEEE, 1993.

[2] S. Chen and J. H. Reif. Fast pattern matching for entropy bounded text. In *Proceedings DCC'95 Data Compression Conference*, pages 282–291. IEEE, 1995.

[3] B. Diggs-Brown. *Strategic Public Relations: An Audience-Focused Approach*. Cengage Learning, p. 48, 2011.

[4] R. M. Fano and D. Hawkins. Transmission of Information: A Statistical Theory of Communications. *American Journal of Physics*, 29:793–794, Nov. 1961. ISSN 0002-9505. doi: 10.1119/1.1937609.

[5] M. Farach, M. Noordewier, S. Savari, L. Shepp, A. Wyner, and J. Ziv. On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. In *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 48–57, 1995.

[6] J. Gable, S. McDonald, and J. Blades. Media Bias Ratings AllSides. https://www.allsides.com/media-bias/media-bias-ratings, 2019.

[7] P. Grassberger. Estimating the information content of symbol sequences and efficient codes. *IEEE Transactions on Information Theory*, 35(3): 669–675, May 1989. ISSN 00189448. doi: 10.1109/18.30993.

[8] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma. Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2751–2754, Austin, Texas, USA, May 2012. Association for Computing Machinery. ISBN 978-1-4503-1015-4. doi: 10.1145/2207676.2208672.

[9] P. Juola. What can we do with small corpora? Document categorization via cross-entropy. In *Proceedings of an Interdisciplinary Workshop on Similarity and Categorization, Department of Artificial Intelligence, University of Edinburgh, Edinburgh, UK*, 1997.

[10] I. Kontoyiannis and Y. M. Suhov. Prefixes and the entropy rate for long-range sources. In *IEEE International Symposium on Information Theory*, pages 194–194. INSTITUTE OF ELECTRICAL ENGINEERS INC (IEEE), 1994.

[11] I. Kontoyiannis, P. Algoet, Y. Suhov, and A. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Transactions on Information Theory*, 44(3):1319–1327, May 1998. ISSN 00189448. doi: 10.1109/18.669425.

[12] A. N. Quas. An entropy estimator for a class of infinite alphabet processes. *Theory of Probability & Its Applications*, 43(3):496–507, 1999.

[13] E. Shearer and E. Grieco. Americans Are Wary of the Role Social Media Sites Play in Delivering the News, Oct. 2019.

[14] P. Shields and B. Weiss. Universal redundancy rates for the class of B-processes do not exist. *IEEE transactions on information theory*, 41 (2):508–512, 1995.

[15] P. C. Shields. Entropy and prefixes. *The Annals of Probability*, pages 403–409, 1992.

[16] P. C. Shields. Universal redundancy rates do not exist. *IEEE transactions on information theory*, 39(2):520–524, 1993.

[17] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, Feb. 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1177170.

[18] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. *arXiv:1907.10121 [physics]*, July 2019.

[19] F. Vis. Twitter as a Reporting Tool for Breaking News. *Digital Journalism*, 1(1):27–47, Feb. 2013. ISSN 2167-0811. doi: 10.1080/21670811.2012.741316.

[20] A. Wyner and J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*, 35(6):1250–1258, Nov./1989. ISSN 00189448. doi: 10.1109/18.45281.

[21] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, May 1977. ISSN 0018-9448. doi: 10.1109/TIT.1977.1055714.