Reinterpreting the surprisal calculation of the Topic Context Model

J. Nathanael Philipp

Saxon Academy of Sciences and Humanities in Leipzig

Summary

This paper proposes a novel computation of wordwise surprisal within the information-theoretic Topic Context Model (TCM), replacing the current average surprisal calculation with simple, non-averaged surprisal calculation.

Surprisal Theory

The TCM is based on the Surprisal Theory of Hale (2001) and Levy (2008). It builds on Shannon's information theory (Shannon, 1948; Shannon, 1951) and on Dretske (1981), who distinguished between the quantitative measure of information in Shannon's sense and semantic information, which transfers knowledge. Both types of information have an impact on brain activity. This is the starting point for the Surprisal Theory, which links information to the cognitive processing effort required (see Bentum et al. (2019), Wilcox et al. (2023)). TCM measures surprisal, a quantitative measure which, in addition, due to its context, also incorporates the linguistic property of meaning. The surprisal of a word w depends on its conditional probability in a given context, see Formula 1.

$$\text{surprisal}(w_i) = -\log_2 P(w_i|w_1, \ldots, w_{i-1}, \text{CONTEXT}) \tag{1}$$

Topic Context Model

Currently the TCM calculates the $\overline{\text{surprisal}}$ according to Formula 2. See Philipp (in press) for the evolution of the TCM formulas.

$$\overline{\text{surprisal}}(w_d) = -\phi \log_2 \frac{c_d(w_d)}{|d|} - \frac{1}{|T_d|} \sum_{i \in T_d} \log_2 \left(P(t_i|d) f(w_d, t_i)\right) \tag{2}$$

- $c_d(w)$ is the frequency of a word w given a document d
- $|d|$ is the total number of words in the document d
- $\frac{c_d(w_d)}{|d|}$ denotes the relative word frequency of word w in a document d
- $\phi \in \{0, 1\}$ defines whether to use the relative word frequency or exclude it from the calculations
- T_d is the topic distribution T of the document d
  - let it be indexed such that $P(t_i|d) \geq P(t_j|d)$ if and only if $i \leq j$
  - pick a $\mu \in (0, 1]$, also called threshold
  - let k_μ be the largest integer such that $\sum_{i=1}^{k_\mu} P(t_i|d) \geq \mu$
  - $k_\mu = 0$ we still require $T_d^\mu = \{1\}$ in order to keep the sets non-empty

$$P(w|t_i) = f(w_d, t_i) = \begin{cases} WT_{w_d, t_i} & \text{for } w \in WT \\ \frac{1}{N} & \text{for } w \notin WT \end{cases} \tag{3}$$

- WT is the normalised word topic distribution $WT \in \mathbb{R}^{|T| \times N}$ of the Latent Dirichlet Allocation (LDA, the underlying topic model)
  - the entries of WT are denoted as WT_{w_d, t_i}
  - and the columns in $\mathbb{R}^N$, which correspond to the topics, by WT_{t_i}
- $w \in WT$ means that there exists a row of WT corresponding to the word w
- $w \notin WT$ only occurs when using a TCM that was trained on different words than using to calculate surprisal for, i.e., a trained TCM is used on texts that were not part of the training corpus with new words.

Currently the underlying probability $P(w_d|t_i)$ is calculated as in Formula 4.

$$P(w_d|t_i) = P(t_i|d) f(w_d, t_i) \tag{4}$$

The first change proposed is to reformulate the probability $P(w)$ and calculate it according to Formula 5.

$$P(w_d) = \sum_{i \in T_d} P(t_i|d) f(w_d, t_i) \tag{5}$$

Estimating that $P(w_d)$ has the following bounds: $\min_{t_i} P(w|t_i) \leq P(w) \leq \max_{t_i} P(w|t_i)$. The relative word frequency $\frac{c_d(w_d)}{|d|}$, will no longer be part of the surprisal calculation, as it behaved as a summand (see Formula 2). Thus the surprisal for the TCM can now be calculated according to Formula 6.

$$\text{surprisal}(w_d) = -\log_2 \sum_{i \in T_d} P(t_i|d) f(w_d, t_i) \tag{6}$$

Results

Tests were performed on five different corpora, four from the Leipzig Corpora Collection and one consisting of Tweets and YouTube comments, all in German. For each corpus four different TCM were trained, with 10, 20, 50 and 100 topics respectively. The threshold was $\mu = 1$. Figure 1 shows the average surprisal distribution based on Formula 2 with $\mu = 1$ for all topic-corpus-distributions and Figure 2 shows the same distributions for suprisal based on Formula 6.

Conclusion

In this paper, we have demonstrated that replacing the average surprisal with a unified, non-averaged surprisal measure in the Topic Context Model (TCM) leads to substantially lower surprisal values while simultaneously increasing the spread of surprisal per word. One should note that increasing the number of topics has no significant effect on the overall distribution and spread of the surprisal values across the corpus.

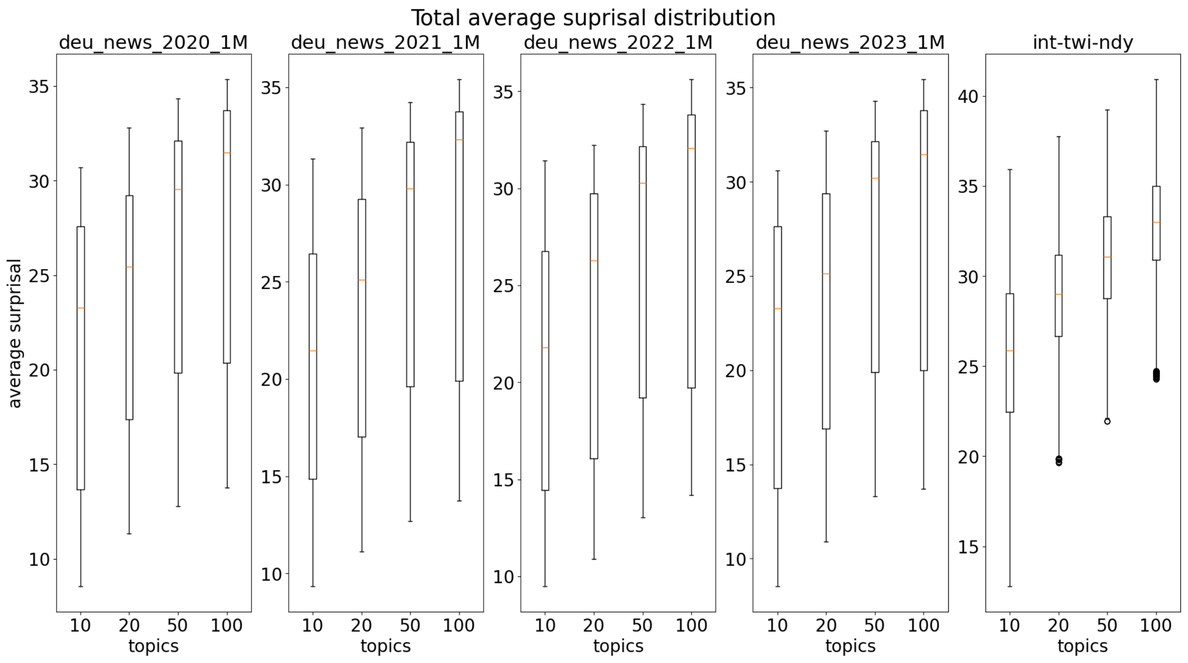

Figure 1: Average surprisal distribution based on Formula 2, from Philipp (in press).

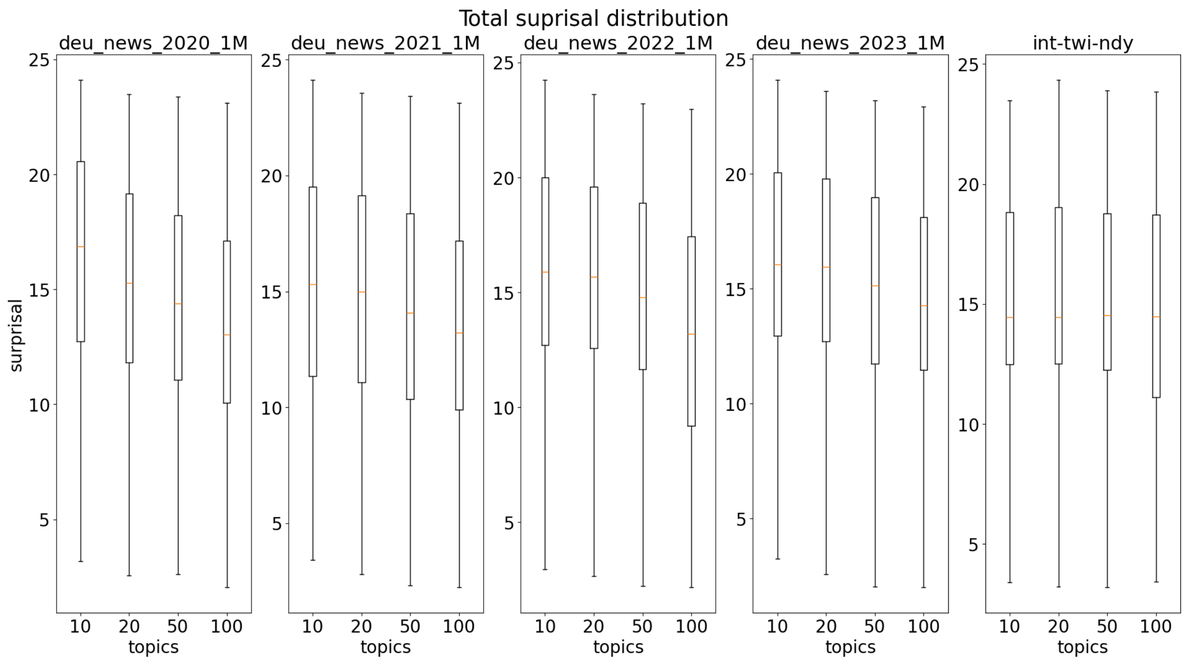

Figure 2: Surprisal distribution based on Formula 6.

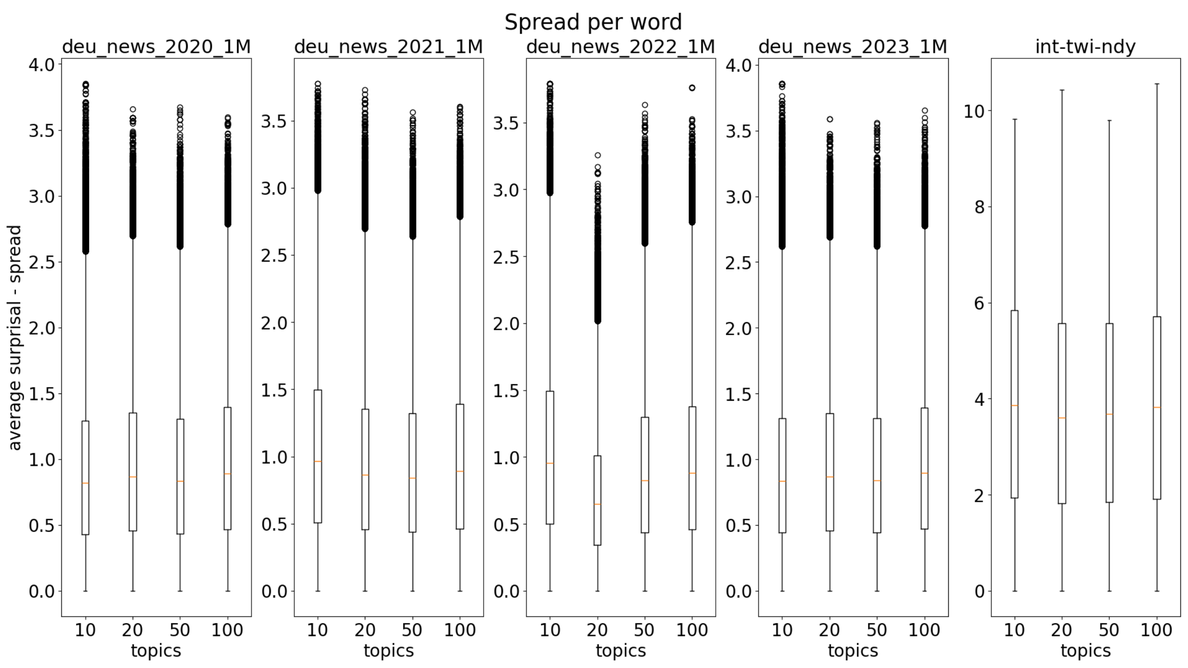

Figure 3: Average surprisal spread distribution based on Formula 2, from Philipp (in press).

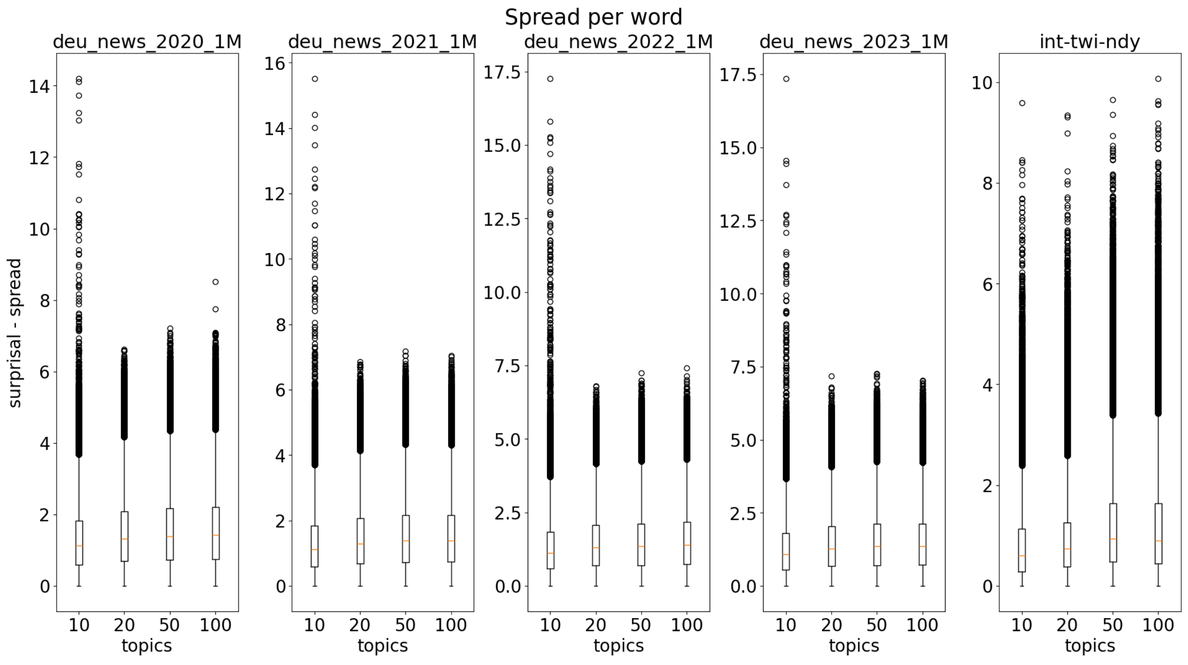

Figure 4: Surprisal spread distribution based on Formula 6.