Text Length and the Function of Intentionality: A Case Study of Contrastive Subreddits

Emily Ohman
Waseda University
ohman@waseda.jp

Aatu Liimatta
University of Helsinki
aatu.liimatta@helsinki.fi

Abstract

Text length is of central concern in natural language processing (NLP) tasks, yet it is very much under-researched. In this paper, we use social media data, specifically Reddit, to explore the function of text length and intentionality by contrasting subreddits of the same topic where one is considered more serious/professional/academic and the other more relaxed/beginner/layperson. We hypothesize that word choices are more deliberate and intentional in the more in-depth and professional subreddits with texts subsequently becoming longer as a function of this intentionality. We argue that this has deep implications for many applied NLP tasks such as emotion and sentiment analysis, fake news and disinformation detection, and other modeling tasks focused on social media and similar platforms where users interact with each other via the medium of text.

1 Introduction

The relationship between the length, intentionality, register, genre, and emotion-associated word distributions in texts is a complex one. The genre often dictates the length for the sake of convention, such as with academic writing. The register (i.e. language as it is used in a specific situation and for a specific purpose) also affects text length, most obviously when there are platform limitations that impose maximum character counts such as SMS messages or tweets, or based on the general pace of the platform. Both of these effects are linked to and can artificially influence the intentionality and framing of a text. We define intentionality as a careful, self-curated production of text with a deliberate purpose or goal behind the communication.

Different text genres also force text producers to convey their message more succinctly and thus perhaps with more intentionality. Examples of this include works of literature, poetry, and political speeches where the evocation of specific emotions in the reader is a desired effect achieved by carefully choosing the "right" words (see e.g. Lipsitz, 2018; Koljonen et al., 2022). We can see similar evocation tactics on social media too with, for example, "rage bait"[1] posts. Despite the near-ubiquitous presence of rage bait on social media, only a handful of academic papers have explored the topic, and these studies have been from a disinformation perspective rather than from an NLP or linguistic perspective (see e.g. Jennings-Roche, 2023; Clem, 2023; Jagayat and Choma, 2023; Curato, 2021; Johnston, 2024; La Rocca, 2022).

In this study, we explore the functions of length, intentionality, affect, and register in contrastive pairs of corpora on the same topic. As our data, we use subreddits (topic-specific discussion forums on Reddit) where at least two separate subreddits exist for the same topic and one is considered to be more serious or in-depth, and the other more general in nature.

We hypothesize that (1) the average length of a text (post or comment) is longer in the more serious, in-depth subreddits, (2) the language in more serious subreddits has more variability and lexical density, and (3) that positive words carry less information than negative words and therefore the more serious subreddits have less positive words than their general subreddit counterpart showing a different aspect of negativity bias (Kanouse and Hanson Jr, 1987).

This study also contributes to the discussion on best practices in how to work with the computational aspects of texts of varying lengths.

2 Background and Related Work

Many studies make off-handed mentions of the different nature of tweets as compared to other social media texts that are not artificially constrained by

[1] *Rage bait* is when social media content is perceived as having been carefully constructed to induce a maximal negative emotional response in readers for the sake of engagement.

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 1–8
November 16, 2024.

length to the same degree (Öhman, 2021a). Some mention the "informal language and expressive content such as emojis and hashtags" (Demszky et al., 2020), and others discuss the limited length, self-contained nature, and the helpfulness of emojis and hashtags for annotators of tweets in contrast with Reddit comments that tend not to include emojis and are often highly context and conversation dependent (Öhman et al., 2020).

For the most part, when social media message length is studied, the focus is on optimization for marketing purposes (Stephen et al., 2015) or crisis communication strategies (Ma and Yates, 2014). One interesting prior study looked into "perceived partner responsiveness" and found that longer messages were perceived as more intimate whether length was measured by counting tokens, characters, words or non-word characters (Freeman and Brinkley, 2014). In this study we focus on word count because we examine word choice beyond length as a metric.

Perhaps the most famous law in quantitative linguistics, Zipf's law (Zipf, 1935, 1949) describes the distribution of sorted measures. In terms of corpus linguistics, Zipf's law states that when the words in a corpus are ordered by frequency, the value of the nth entry is inversely proportional to n. This frequency can also be used to measure the length of words and more recent studies have extended Zip's law to state that information content causes word length to increase (Piantadosi et al., 2011).

Garcia et al. (2012) studied the distribution of positive and negative words and their frequency as they relate to information density and word frequency overall. Using small emotion lexicons (1034 entries for English) they found that word valence and frequency of use are related, and in particular that positive emotional content is more common than negative content. However, because positive words are more common, they point out that the relative rarity of negative words causes them to carry more information.

More recently Singh et al. (2023) used readability and other linguistic complexity metrics to show that the negativity bias holds true for Reddit data too meaning that negative emotions are associated with more complex texts than positive emotions. They continue on to show that current state-of-the-art transformers such as BERT have more trouble with the more complex texts and discuss the implications of this for the evaluation of emotion and sentiment analysis models.

Intentionality is a very understudied concept in NLP. Here we use the term to mean a careful selection of words to achieve a desired affective reaction in the reader, i.e., affective rhetorical devices. Intentionality has mostly been researched as part of the field of rhetoric (see e.g. Bitzer, 1968; Burke, 1969) and in social psychology for example, to examine how other people determine whether an action or message was intentional, but some have looked at the rhetoric and affect of political speech: Teneva (2021) showed that emotional appeals in political discourse aimed at social solidarity, group identification, and shaping public opinion in the Internet news discourse. Their conclusion was that political emotions play an important role in modern argumentation.

In quantitative linguistics, varieties of language that are defined by their situation of use and communicative function are called registers. While there are many approaches to register analysis (e.g. Biber, 1988; Halliday and Matthiessen, 2013), the analysis of registers is generally founded on the idea that certain linguistic features are more or less well-suited to certain situational and functional concerns, and as such, registers tend to prefer and disprefer different sets of linguistic features depending on the situational and functional circumstances of the register. For instance, narrative registers might prefer features such as past-tense verb forms and third-person pronouns.

While text length is commonly recognized as a confounding factor for linguistic analyses, text length itself is rarely the object of study in linguistics. However, Liimatta (2022, 2023) analyzes comment length on Reddit from the point of view of register. These studies show that the length of a text is not determined at random even in contexts where the author can in principle write a text of any length on Reddit and other similar social media platform, when they are not constrained by genre conventions, publisher requirements or limitations of technology. Instead, Liimatta (2022, 2023) demonstrates that text length is closely linked with the idea of register. Just like different communicative functions and situational concerns prefer linguistic features that are well-suited for the situation, so too do different communicative functions prefer text lengths that are similarly well-suited for the situation. Furthermore, Liimatta (2022) goes on to show that the associations between text length and register are not all universal: within Reddit data,

Subreddit name	category	Speakers (users)	Utterances (comments)	Conversations (posts)	Comment:Post Ratio
startrek	general	111,119	2,215,516	110,183	20.11
DaystromInstitute	in-depth	18,811	505,171	17,308	29.19
Aviation	general	93,784	1,171,366	128,447	9.12
Flying	in-depth	38,822	1,261,216	65,024	19.40
AskHistory	general	17,646	78,932	14,441	5.47
AskHistorians	in-depth	193,943	2,065,764	327,340	6.31
wow	general	520,414	14,618,201	809,585	18.10
CompetitiveWoW	in-depth	14,487	126,323	8,861	14.26
wown00bs	beginner	12,452	101,138	14,456	7.00

Table 1: Overview of subreddits

many subreddits can differ in terms of the role that text length plays in them.

Of particular relevance to the present study are the findings by Liimatta (2022, 2023) that longer comments on Reddit have higher frequencies of linguistic features associated with higher information density, such as nominalizations, as well as ones associated with a more complex argument structure, such as infinitives and certain modal verbs; whereas the frequencies of features associated with non-edited, casual, "on-line" production tend to be higher in shorter comments, including features such as contractions, subordinator *that* deletion, first-person singular pronouns, and private verbs. These results suggest that, in general, longer comments tend to be more informationally dense and more carefully edited, as opposed to shorter comments, which tend to be more casual, less informationally dense, and less edited.

3 Data

We used the convokit Reddit corpus (Chang et al., 2020) to collect data from related but contrastive subreddits. We chose pairs of subreddits that are on the same topic but split into a general and a specific, typically more serious or professional, subreddit. The selection of the subreddits was made based on the authors' knowledge of the topics and the subreddits, with both authors agreeing on each pair of subreddits. The criteria here vary slightly per subreddit; for example, we chose r/startrek as a general subreddit representing Star Trek discussions and r/DaystromInstitute which is a more serious subreddit dedicated to "Serious, in-depth discussion about *Star Trek.*"[2], the pairs r/Aviation and r/flying where aviation is for enthusiasts and flying for pilots, and r/AskHistory and r/AskHistorians where the latter has much more stringent requirements for both posts and comments than the former.

We attempted to pick subreddit pairs of similar relative standing within pairs, however, there are many differences between our chosen pairs. All of the subreddits can be considered to be niche or nerdy in some aspects, including the more general ones and therefore they might not be different enough from each other for some of the pairs. We expect the most significant differences to be shown in the pair r/starttrek and r/DaystromInstitute because the first in the pair is a fairly popular topic of discussion in the mainstream consciousness.

As a TV show, Star Trek has been one of the most long-lasting and successful franchises on TV since the 1960s and it is also easy for more casual viewers to form opinions and ask questions on the subreddits (Weldes, 1999; Pearson and Davies, 2014).

Compared to some of the other pairs, r/aviation and r/flying for example, r/aviation is meant for enthusiasts and r/flying for pilots. However, aviation enthusiasts are a more niche group to begin with compared to TV, and therefore it is unlikely that a large percentage of posts would be from the perspective of a more general public. Similarly, although r/AskHistorians is one of the most strictly moderated subreddits on reddit with stringent guidelines about what top-level posts must contain and r/AskHistory was established as a more lax alternative with fewer citation criteria and the like, the questions and discussions remain rather similar. Finally, we have the World of Warcraft (WoW)-related subreddits, that could be said to be somewhat niche despite their immense popularity because casual players, let alone the general public, are unlikely to be actively discussing the topic. We expect r/wow and r/wownoobs to be somewhat sim-

[2]From the community description of r/startrek

ilar simply because r/wow is so large it is going to catch a large majority of posts that would be more suited for r/wownoobs, but we expect r/wow and r/competitivewow to show more differences on the general to in-depth axis similar to the Star Trek and aviation-related subreddits.

The sizes of the corpora are not uniform and therefore the results were normalized by token count where applicable. Table 1 shows the number of tokens, unique usernames (speakers), conversations (posts), utterances (comments), and the ratio of comments to post to indicate how much engagement is typical for a post on each subreddit.

4 Method

For type-token ratio calculations, we used the same token count as for normalization and lemmatized type counts. We used SpaCy for tokenization and lemmatization. No other preprocessing steps were taken. To calculate the distribution of emotion-associated words, we use the NRC emotion intensity lexicon (Mohammad, 2018) which divides words into Plutchik's 8 core emotion categories of anger, anticipation, disgust, fear, joy, sadness, surprise, and trust and associates a value between 0 and 1 for the intensity of the emotion. Our method can therefore detect the difference between, e.g., annoyance, anger, and rage and be judged more akin to valence (Öhman, 2021b). Although we measure the prevalence of emotion-associated words in all eight Plutchik categories, we focus on Joy and Anger as the proxy emotions we expect to represent positive and negative in this context.

5 Results

The average post/comment length in the less general subreddits are consistently longer than in the more general ones (see table 2). For most pairs, the difference in length is double for the more serious subreddits, but in the case of the History subreddits the average length of posts in the more serious subreddit is almost 15 times longer. For the World of Warcraft subreddits, the general subreddit's average post length is about half of the more serious one as expected, but about the same as the beginner one. Additionally, the same pattern holds with word length as well; while the difference in average word length is not large, it is consistently longer in the more serious subreddits by roughly .2 characters. For example, for *startrek* the median word length was 3 and for the *DaystromInstitute* 4.

Note that the average post length refers to all comments on posts, not just the original post. The difference in length is much larger if going by the original post only. We chose to merge the comments and posts here to show that the entire conversation on the in-depth subreddits is more complex, rather than just the starting post.

Because a higher type-token ratio (TTR) has been associated with registers with a higher information density (e.g. Biber, 1988), and as such we might expect the less general subreddits to have a higher TTR, we calculated the moving average type-token ratios (MATTR) for the subreddits. We chose MATTR over TTR since the lengths of the posts between the subreddit pairs were just so drastically different that for example, for r/startrek a typical post would only have about 70 words of which between 55 and 60 were unique making the TTR values very high (.75 to .8), whereas r/DaystromInstitute had lower TTR values typically between .35 and .45 because a typical post was between 800 and 1400 words of which 400-500 were unique. However, the MATTR values for all the subreddits in question was between .80 and .81 suggesting that MATTR was no better than TTR as a measure for information density or intentionality in our texts.

In table 2 we present the log-likelihood significances between *joy* and *anger*. The full log-likelihood (see table 3) and emotion word distributions (see table 4) are presented in the appendix.

6 Analysis and Concluding Discussion

We found that the average length of a text in posts and comments were longer in the more serious, in-depth subreddits at almost twice the length on average, with some significant deviations for the two History subreddits where the posts on the more serious subreddit were almost 15 times longer than on the less serious one. This finding supports the earlier results on the relationship between text length and register, which associated longer Reddit comments with linguistic features related to information density and more carefully edited content. The World of Warcraft subreddits differed here slightly with the post lengths on the general subreddit still about half the length of the more serious one (33.4 vs. 59.7), but with little difference between the competitive and beginner subreddits (57.4 vs. 59.7). This might be because the general subreddit is less specific and less moderated in terms of allowed

Subreddit name	category	Tokens	Avg. comment length	Avg. word length	Emotion (log-likelihood)
startrek	general	77,286,488	34.88	3.88	Joy+
DaystromInstitute	in-depth	38,072,122	75.37	4.08	Anger+
Aviation	general	27,775,314	23.71	3.99	Joy+, Anger+
Flying	in-depth	49,372,833	39.15	3.78	
AskHistory	general	4,319,093	23.90	4.18	Joy+, Anger+
AskHistorians	in-depth	150,197,577	351.89	4.26	
wow	general	515,071,857	33.39	4.01	Joy++
CompetitiveWoW	in-depth	7,541,880	59.70	4.30	Anger++
wown00bs	beginner	5,781,910	57.17	3.97	Joy-, Anger-

Table 2: Results

content, or that by asking a question on a beginner subreddit posters feel less self-conscious about asking what might be considered "stupid questions" on the main subreddit. Perhaps the niche nature of the discussion is not so much about expertise but about specificity. Interestingly the average post length on the main WoW subreddit and the main Star Trek subreddit was about the same (33.4 vs. 34.9) and on the general r/aviation and r/AskHistorians (23.7 vs. 23.9).

Our finding that the serious subreddits have both longer average comment lengths and features associated with higher information content, such as higher average word length, is also in line with the findings by Liimatta (2022, 2023) correlating longer comment length on Reddit with higher frequencies of linguistic features associated with densification of information, such as nominalizations, and shorter comments with features of more casual and personal, less carefully considered language, as well as Piantadosi et al. (2011) that information content causes word length to increase.

Another difference became apparent when sorting the posts on the subreddits by all-time top posts; the top posts in the general subreddits tended to contain images or videos rather than the long essay-like posts in the more serious subreddits, further highlighting the differences between the general and niche.

For the emotions, we did find that the posts in the more general subreddits tended to contain more joy-related words at higher intensities than their more in-depth counterparts. However, the case for anger-related words was not as clear-cut. The results indicate that r/DaystromInstitute posts tend to contain more anger than r/startrek and the same can be said for r/wow when compared to r/CompetitiveWoW, but for the others, it seemed that the general subreddit posts contained more emotion-related words at higher intensities in general.

In conclusion, we have shown that the length of a post differs based on what we consider intentionality, that is, the in-depth nature of the text with more niche conversation significantly longer than more general conversation on the same topic. We were unable to show a difference in lexical variation due to the short messages in the general subreddits, but the results support the idea that general discussions are more positive than the more complex, longer texts in the in-depth subreddits.

7 Future Work

The complex relationships between the length, intentionality, register, genre, and emotion-associated word distributions in texts remains and interesting and useful area of study. For instance, while we have demonstrated an indirect relationship between emotion-associated word distributions and certain types of register variation, insomuch as both of them are linked with text length, it would be fruitful to perform a deeper analysis of the connections between different register dimensions, emotion-associated words, and text length, to better understand this multifaceted problem.

In future work we hope to include part-of-speech and syntactic structures as features related to register and intentionality. We also hope to study additional subreddits and better utilize both fine-tuned LLMs and qualitative analysis.

Limitations

The emotions lexicons used for the analysis are not specifically designed for social media which could influence how well the labeled emotions in the lexicon correspond to the emotions in our data.

Furthermore, in the selection process of the subreddits, several options were excluded due to computational limitations. Future studies should include a more varied selection of subreddits and more manual validation of the results.

Ethics Statement

All data used in this study were obtained from public forums using convokit. We recognize that user-generated content can be subject to ethical concerns regarding privacy and consent. However, given the public and anonymous nature of Reddit posts and the fact that we limited our analysis to text-level features (length, register variation, and intentionality), without making inferences about individual users or their personal identities, any negative impact on any specific users or communities is mitigated.

The study was conducted with respect for the autonomy of online community members, and we acknowledge that online discourse can contain sensitive content. To mitigate any risks of harm, we avoided analyzing subreddits that could contain vulnerable populations or sensitive topics.

Finally, the potential biases in our study, including selection bias or platform-specific biases, were considered in our methodology. We took care to transparently report these limitations in our findings, ensuring that our conclusions are contextualized within the broader ethical and social considerations of using online data.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 24K21058.

References

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge.

Lloyd F Bitzer. 1968. The rhetorical situation. *Philosophy & rhetoric*, pages 1–14.

Kenneth Burke. 1969. A rhetoric of motives. *U of California P*.

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Dany Clem. 2023. *Accessible and Activist Rhetorics: TikTok as a Learning Tool*. Ph.D. thesis, Arkansas State University.

Nicole Curato. 2021. Interruptive protests in dysfunctional deliberative systems. *Politics*, 41(3):388–403.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Linda Kramer Freeman and Jason Brinkley. 2014. Length matters: Message metrics that result in higher levels of perceived partner responsiveness and changes in intimacy as friends communicate through social network sites. *The Journal of Social Media in Society*, 3(1).

David Garcia, Antonios Garas, and Frank Schweitzer. 2012. Positive words carry less information than negative words. *EPJ Data Science*, 1(1):1–12.

M. A. K. Halliday and Christian M. I. M. Matthiessen. 2013. *Halliday's Introduction to functional grammar*, 4 edition. Routledge, London.

Arvin Jagayat and Becky L Choma. 2023. A primer on open-source, experimental social media simulation software: Opportunities for misinformation research and beyond. *Current Opinion in Psychology*, page 101726.

Allison Jennings-Roche. 2023. Delegitimizing censorship: Contending with the rhetoric of an anti-democratic movement. *The Political Librarian*, 6(1).

Hank Johnston. 2024. The maga movement's big umbrella. *Mobilization: An International Quarterly*, 28(4):409–433.

David E Kanouse and L Reid Hanson Jr. 1987. Negativity in evaluations. In *Attribution: Perceiving the cause of behavior*. Lawrence Erlbaum Associates, Inc.

Juha Koljonen, Emily Öhman, Pertti Ahonen, and Mikko Mattila. 2022. Strategic sentiments and emotions in post-second world war party manifestos in finland. *Journal of computational social science*, 5(2):1529–1554.

Gevisa La Rocca. 2022. The mediatization of disinformation as a social problem: The role of platforms and digital media ecology. *Information Disorder. Learning to Recognize Fake News, Peter Lang, Berlin*, pages 43–62.

Aatu Liimatta. 2022. Do registers have different functions for text length? A case study of *Reddit*. *Register Studies*, 4(2):263–287.

Aatu Liimatta. 2023. Register variation across text lengths: Evidence from social media. *International Journal of Corpus Linguistics*, 28(2):202–231. Publisher: John Benjamins.

Keena Lipsitz. 2018. Playing with emotions: The effect of moral appeals in elite rhetoric. *Political Behavior*, 40(1):57–78.

Xin Ma and Justin Yates. 2014. Optimizing social media message dissemination problem for emergency communication. *Computers & Industrial Engineering*, 78:107–126.

Saif Mohammad. 2018. Word affect intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Emily Öhman. 2021a. *The language of emotions: Building and applying computational methods for emotion detection for English and beyond*. UniGrafia: University of Helsinki.

Emily Öhman. 2021b. The validity of lexicon-based sentiment analysis in interdisciplinary research. In *Proceedings of the workshop on natural language processing for digital humanities*, pages 7–12.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. Xed: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552.

Roberta Pearson and Máire Messenger Davies. 2014. *Star Trek and American Television*. Univ of California Press.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Pranaydeep Singh, Luna De Bruyne, Orphee De Clercq, and Els Lefever. 2023. Misery loves complexity: Exploring linguistic complexity in the context of emotion detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12871–12880.

Andrew T Stephen, Michael Sciandra, and Jeffrey Inman. 2015. Is it what you say or how you say it? how content characteristics affect consumer engagement with brands on facebook. *How Content Characteristics Affect Consumer Engagement with Brands on Facebook (October 1, 2015). Saïd Business School WP*, 19.

Ekaterina V Teneva. 2021. The rhetoric of political emotions in the internet news discourse. *Galactica Media: Journal of Media Studies*, 3(1):125–145.

Jutta Weldes. 1999. Going cultural: Star trek, state action, and popular culture. *Millennium*, 28(1):117–134.

George Kingsley Zipf. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Routledge.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.

Appendices

Log-likelihood

	subreddit	O1	1%	O2	2%	LL	%DIFF	Bayes	ELL	RRisk	LogRatio	OddsRatio
JOY	st/DI	1171246	1.52	421554	1.11	31983.22	36.87	31964.65	0.00002	1.37	0.45	1.37
ANG		619212	0.8	316266	0.83	272.86	-3.55	254.3	0	0.96	-0.05	0.96
JOY	av/fly	297945	1.07	559895	1.13	604.19	-5.41	586.02	0	0.95	-0.08	0.95
ANG		162804	0.59	225974	0.46	5700.07	28.07	5681.91	0.00001	1.28	0.36	1.28
JOY	askhist/hists	44201	1.02	1466122	0.98	94.65	4.84	75.8	0	1.05	0.07	1.05
ANG		42566	0.99	1221812	0.81	1427.85	21.15	1409	0	1.21	0.28	1.21
ANG		83855	0.51	594570	0.54	303.83	-6.18	285.17	0	0.94	-0.09	0.94
JOY	wow/comp	7329450	1.42	87558	1.16	3834.93	22.57	3814.86	0	1.23	0.29	1.23
ANG		5859108	1.14	112940	1.5	7683.63	-24.04	7663.55	0	0.76	-0.4	0.76
JOY	wow/noob	7329450	1.42	79103	1.37	123.34	4.01	103.27	0	1.04	0.06	1.04
ANG		5859108	1.14	49113	0.85	4581.69	33.92	4561.62	0	1.34	0.42	1.34

Table 3: Log-likelihood calculations

Emotion word distributions

	Nanger	Nanticipation	Ndisgust	Nfear	Njoy	Nsadness	Nsurprise	NTrust
startrek	80.11904	137.4671	46.7845575	97.41404086	151.546	77.83572	40.96666	236.8187
DaystromInstitute	83.07029	114.2822	36.260473	117.9215713	110.7251	75.66423	36.94725	220.891
aviation	58.61455	104.6397	32.8963086	97.75651717	107.2698	65.30503	45.19683	187.3386
flying	45.76897	126.2415	28.0944731	79.7299766	113.4014	58.57647	39.32689	206.2676
AskHistory	98.554	100.7872	48.1698912	157.3611404	102.3396	89.99936	38.91132	204.3497
AskHistorians	81.34697	95.62274	37.160274	129.5403759	97.61289	74.19905	34.24224	205.5677
wow	113.7532	126.5027	53.5203207	139.9568105	142.2995	94.2153	62.13591	211.5859
CompetitiveWoW	149.7507	120.5169	53.4918681	166.0155359	116.0964	111.2234	71.68543	198.8933
wownoob	84.94327	137.9898	33.5502715	116.2166153	136.812	72.26706	54.0252	226.2408

Table 4: Normalized Emotion-word distributions by intensity measures