Academic paper writing tone depending on the author's location

Soyoung Yoon, HyunKyu Jung, Dongmin Lee,
Hyunji Lee, Woojin Jang

1. Introduction

When talking to people from various countries in English, we often feel that the characteristics of English they use are affected by their cultural background. Therefore, we wondered if there would be a difference in English usage in research papers depending on cultural background. We gathered Computer Science papers from different fields, countries, and time to resolve our curiosity. We report our correlation results depending on category and published year. Our code used for analysis is open to the public at https://github.com/amy-hyunji/CS564.

2. Related Work

There has been a similar study about finding the correlation between nationality of authors and some linguistic features found in biomedical articles. Netzel et al. conducted statistical analysis on the effect of author's nationality to the difference in average number of words and verbs per sentence, frequently used words, and choice between interchangeable words. However, our topic is novel in that there has been no research on statistics of linguistic features in Computer Science literature with temporal analysis.

3. Methods

We tried to find the effect of published year, nationality of authors, or number of authors to the usage of singular/plural first person pronouns, the ratio of male/female pronouns, the use of gendered terms, proper noun ratio, verb ratio, and professional word ratio. Also, we extracted 20 keywords from each Computer Science field to find which words are frequently used in which field.

3-1. Dataset processing

We crawl all Computer Science research papers published between 2000 and 2019 from Scopus and extracted their title, abstract text, affiliation, number of authors, and published year.
Then, to test our hypotheses, we utilize some NLP techniques to analyze the extracted abstract texts. Below is the extracted features that we analyze:

First-person pronouns:

We count occurrences of singular/plural first person pronouns.

- singular first person pronouns: I, me, my, mine, myself
- plural first person pronouns: We, us, our, ours, ourselves

Gender-specific words:

We count occurrences of gender-specific based on online search.

ex1) http://perfectyourenglish.com/blog/list-of-masculine-and-feminine-gender/
ex2) https://www.englishbix.com/masculine-and-feminine-gender-words/
ex3) https://en.wikipedia.org/wiki/Gender-neutral_language

We compute the gender index for each paper. Gender index is a number between 1 and -1 and its calculation formula is as follows:

$$\text{gender\_index} = \frac{\#(male\_pronoun) - \#(female\_pronoun\ )}{\#(male\_pronoun) + \#(female\_pronoun)}$$

Gender index close to 1 indicates that the text is using more male-related words and vice versa.

Proper noun ratio:

We also hypothesize that there will be a correlation between the characteristics of sentence structure(linguistic features) with country, subject category, and publication year. To extract linguistic features, we first conduct part-of-speech tagging(pos) for each abstract text using the spacy library. We first load the pretrained english nlp model by spacy.load("en_core_web_sm"). Then, we utilize the "tag"(classified as NNP, VBZ, VBG, IN, NN, IN, CD,

…) information. Proper noun ratio is defined as the proportion of tokens that were classified as "NNP" by the tag information. Often in papers, they name their unique methods and call it by its name. We hope to capture how often they name their own methods and the frequency of other proper noun usages by this proper noun ratio.

Professional ratio:

Professional ratio is defined as unknown words in the nltk corpus. In the nltk library, it has tokens built into the nltk corpus which could be accessed by nltk.corpus.words.words(). There are a total of 236736 vocabularies extracted from various sources. We define words that don't appear in this corpus as "professional". Professional words like MOOCs, 2Ph-HSM, ECG do not appear frequently in casual texts, therefore having a higher probability of not appearing on the nltk corpus. To capture those words, we first tokenize each abstract text. Among the tokens, we count the occurance of tokens that are not inside the set of nltk vocabulary, and report the ratio. By the professional ratio, we aim to capture the frequency of using technical words in relation to CS fields.

TF-IDF analysis:

We extract the top 20 keywords for each category in the Computer-Science department by using TF-IDF. TF-IDF, which is short for Term Frequency - Inverse Document Frequency, is used to score the relative importance of words. We calculate TF by the number of times a word appears in a document divided by the total number of words in the document. We calculate IDF by the log of the number of documents divided by the number of documents that contain the word. This determines the weights of rare words across all documents in the corpus. We use this by removing unnecessary or frequent words such as 'into', 'did', 'is', 'in', etc. Rather than explicitly choosing these words, we use nltk stopwords to filter words. We remove these words because these words change the TF value and we thought that since we are only looking for keywords, these counting will work as a noise.

Data we use contain papers that are tagged to more than one category. In this case, we add the paper to each category list it was added to. Since we are focusing on keywords of each category, we think that adding the same paper multiple times to each category is associated with giving reliable IDF value.

3-2. Analysis

Lastly, we used lm() function in R to analyze regression between the computed indice and some indicators such as number or nationality of authors, published year, and CS fields.

We report factors with meaningful relations on the main paper, and introduce other methods we use but couldn't find significant correlation on the appendix.

4. Analysis Results

4-1. First-person pronouns

Using linear regression analysis with R, we examine the coefficient values for CS papers reporting singular/plural first person pronouns. Figure 1 shows that the current trend is not using singular first person pronouns like I, me, my. The linear regression analysis shows that the country and the number of authors have strong correlations with reporting singular first person pronouns. The United States tends to use singular first person pronouns more than China. The reason seems that the US has a culture that values individuals more than China. Also, as the number of authors increases, the usage of singular first person pronouns decreases.

lm(formula = i ~ year + country + author_num, data = df)

Coefficients:

	Estimate	Standardized	Std. Error	t value	Pr(>|t|)	
(Intercept)	-9.092820	0.000000	5.406111	-1.682	0.0926	.
year	0.005516	0.019932	0.002686	2.054	0.0400	*
con	0.158647	0.049958	0.032622	4.863	1.17e-06	***
author_num	-0.111426	-0.147859	0.007572	-14.715	< 2e-16	***

lm(formula = we ~ year + country + author_num, data = df)

Coefficients:

	Estimate	Standardized	Std. Error	t value	Pr(>|t|)	
(Intercept)	-6.417e+01	0.000e+00	1.279e+00	-50.16	<2e-16	***
year	3.307e-02	8.734e-02	6.355e-04	52.03	<2e-16	***
con	8.401e-01	2.070e-01	6.718e-03	125.05	<2e-16	***
author_num	2.565e-02	2.626e-02	1.556e-03	16.48	<2e-16	***

4-2. Gender Index: male terms & female terms

Using linear regression analysis with R, we examine the coefficient values for CS papers reporting masculine and feminine terms. We defined gender index to compare masculine and feminine terms. The linear regression analysis shows that years paper written has big correlations with gender index. The coefficient value is minus, which means recent papers have been more balanced usage of male and female terms. Figure 3 shows most CS papers mention masculine words which are highly biased, but they're converging to use both gender terms.

glm(formula = gender ~ year + country + author_num, data = df))

Coefficients:

	Estimate	Standardized	Std. Error	t value	Pr(>|t|)	
(Intercept)	21.6546856	0.0000000	1.5549419	13.926	<2e-16	***
year	-0.0104025	-0.0828579	0.0007724	-13.468	<2e-16	***
con	0.0179343	0.0130876	0.0084649	2.119	0.0341	*
author_num	-0.0025619	-0.0091456	0.0016782	-1.527	0.1269	

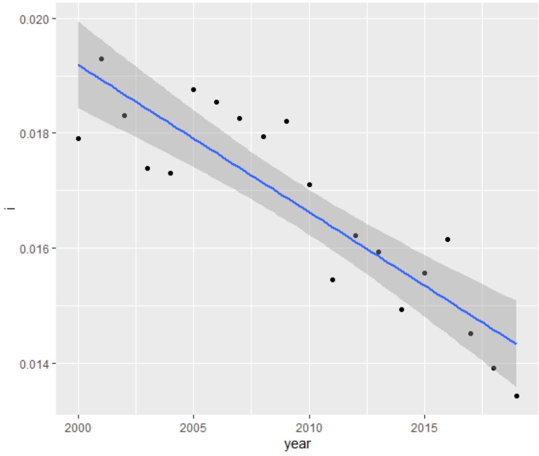
4-3. Gendered Terms

Using linear regression analysis with R, we examine the coefficient values for CS papers reporting gendered terms. There is a big trend to reduce using gendered terms like fireman in English. Therefore, on linear regression results, the paper's publication year has high correlation with using gendered terms. As time flows, people reduce using gendered terms and we can see this result in figure 4. Also, the number of authors has correlation with using gendered terms. It shows that as the number of authors increases, the paper reduces to use gendered terms because authors would recommend themselves not to use gendered terms.

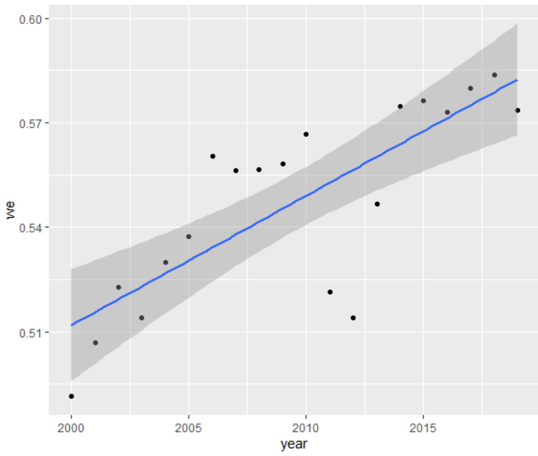
lm(formula = gender ~ year + country + author_num, data = df)

Coefficients:

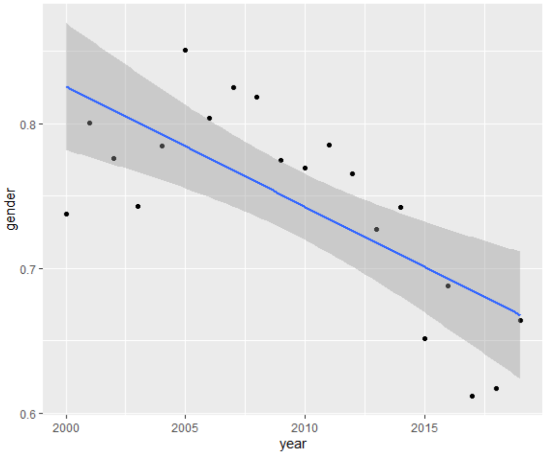
	Estimate	Standardized	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.515e-02	0.000e+00	3.113e-02	2.735	0.00624	**
year	-4.081e-05	-3.367e-03	1.547e-05	-2.638	0.00834	**
con	-2.686e-04	-2.061e-03	1.648e-04	-1.630	0.10305	
author_num	-2.158e-04	-6.790e-03	3.853e-05	-5.602	2.12e-08	***



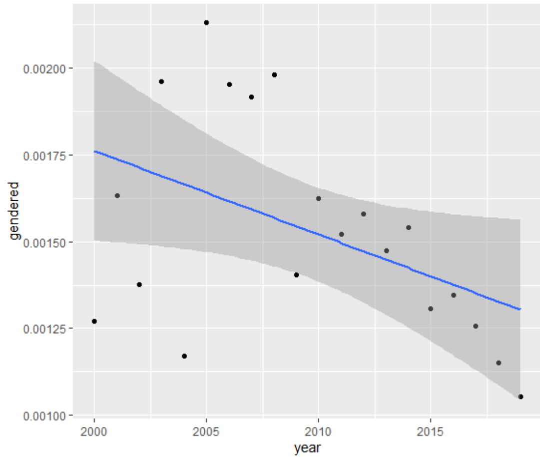
[Figure 1]



[Figure 2]



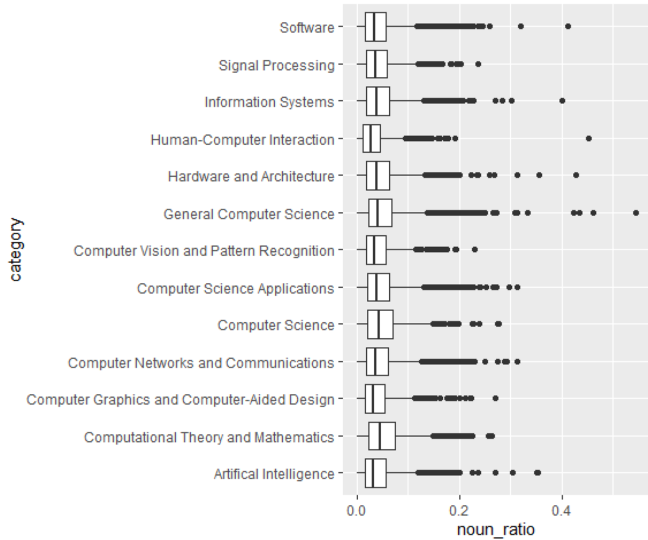
[Figure 3]

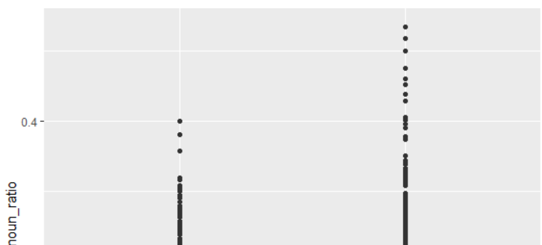


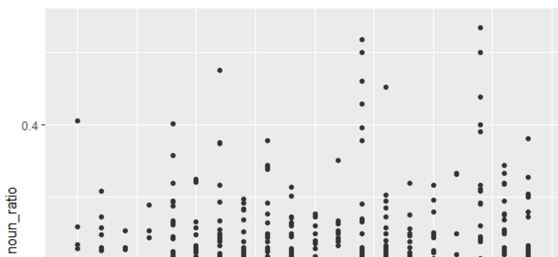
[Figure 4]

4-4. Correlation of proper nouns

We analyze the correlation between CS categories, year, and country on the ratio of proper nouns. (In the plot, we denote the proper noun ratio as 'noun_ratio') The first plot represents the result for analyzing the dataset distribution of proper noun ratio(x-axis) and CS categories(y-axis). The following fields have the lowest ratio in proper nouns: "Computer Graphics and Computer-Aided Design", "Human-Computer Interaction", and "Computer Vision and Pattern Recognition". Among the three of them, two of them relate to visual images. Those topics seem to use less naming. In contrast, "Information systems", "Computational theory and Mathematics", and "Hardware and architecture" categories show more ratio of proper nouns, meaning they do more naming. Maybe there would be more proportion of unique symbols, and named objects. For the distribution compared over countries, authors from China seemed to write more in favor of proper nouns. If we look at the boxplot by year and noun_ratio, another interesting part was that the usage of proper nouns was significantly higher during 2005 and 2014, compared with other years.

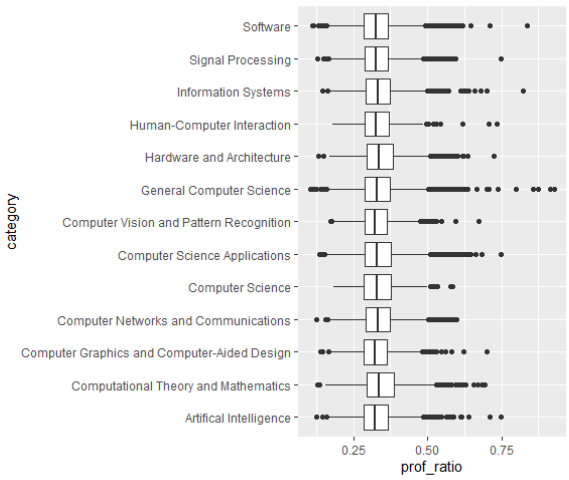


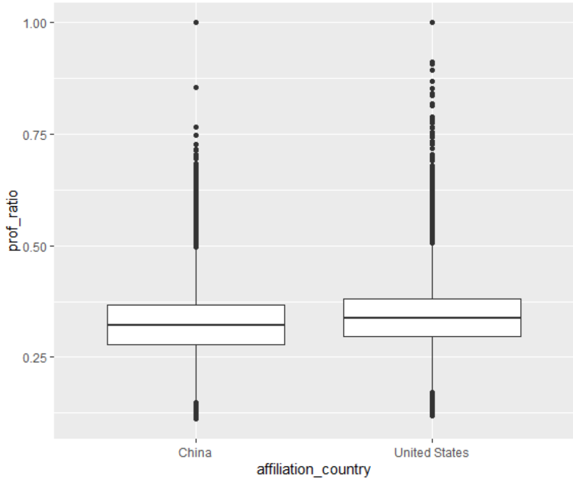




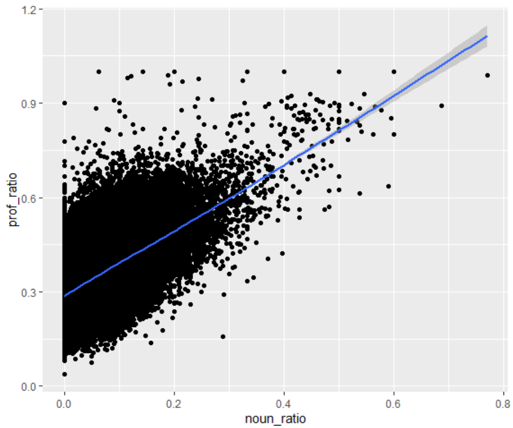
4-5. Correlation of professional words

For the professional word ratio, “Hardware and architecture” and “Computational theory and mathematics” fields have the highest distribution. This is fairly reasonable in the fact that those two need to write many mathematical equations, symbols, and proper nouns. The plot which compares between countries shows that the United States slightly used more proportion of professional words when writing papers.





4-6. Correlation between word ratio and proper noun ratio



After the analysis, there is also a significant correlation between the professional word ratio and proper noun ratio. We plot the correlation between the noun ratio and professional ratio for all datasets, and conduct regression analysis using geom_smooth() from R package. We find that there is a linear correlation between the two features. This means the more proper nouns an abstract has, the more professional ratio an abstract has. The result is understandable in that professional words, or out-of-vocabulary words tend to be proper nouns, and proper nouns are frequently used as technical, or professional words.

4-7. TF-IDF analysis by keywords on category

We report the number of papers for each category at Table 1. Table 2 shows the top 10 keywords for each category. (Table 1 and 2 are in Appendix)
We can see that there are many intersections between the tables. For example,

- The most frequent word was `method` which came out 11 times out of 12 papers.
- `data`, `network`, ‘based’ and `paper` came out 9 times
- ‘using’ came out 8 times
- ‘information’ and ‘system’ came out 7 times
- ‘learning’ came out 6 times
- ‘algorithm’, ‘propose’, ‘energy’ and ‘performance’ came out 5 times

All these words seemed likely to come out frequently in CS papers.
Also, we could see some differences in each category by looking at their unique keywords. For example,

- ‘vision’ and ‘vr’ in computer graphics and computer-aided design
- ‘hardware’ in hardware and architecture
- ‘convolution’ in artificial intelligence
- ‘fluid’ and ‘lagrangian’ in computational theory and architecture.

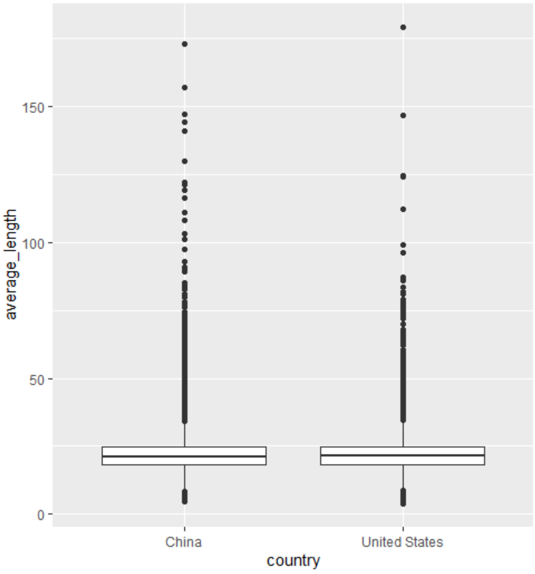
These unique words seemed likely to be keywords for each category.

5. **Appendix**

During the experiment, we make some hypotheses related to the tone of writing and the countries especially USA and China.

5-1. Country and average length of the sentence

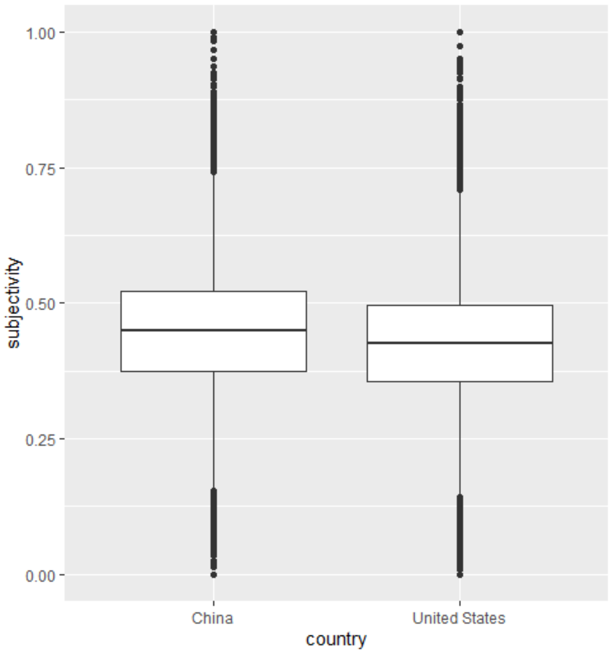
First hypothesis is "Average length of the sentences will differ depending on the native-tongue language". In other words, in the case of the USA, since the native language is English, the sentences will be short because the people can easily explain. On the other hand, in the case of China, people might have difficulty explaining the non-natvie language which can lead to long sentences.



Above graph shows the average length of sentences depending on countries. Beyond the expectation, there were not big differences on the average length. Therefore we can say that native language does not matter for the average length. For the failure of the hypothesis, we can infer that the development of the translation programs made the part.

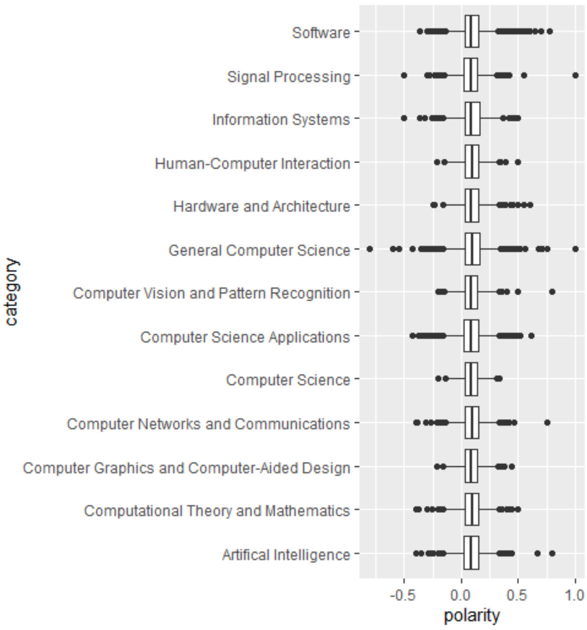
5-1. Subjective and Country

We extracted the "subjectivity" feature using NLP techniques. When the value of the subjectivity is close to 1, the sentence contains more opinion of the authors. If the value is closer to 0, the sentence contains factual information. We made the hypothesis that the USA will use more subjective words in the sentence than China due to the type of the country. However, when we drew the plot using R, we found that the difference was insignificant and even China had a higher subjectivity score than the USA. We conclude that the type of the country does not give the effect on the subjectivity of the sentence.

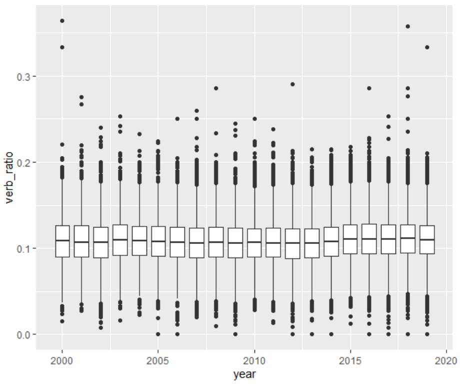


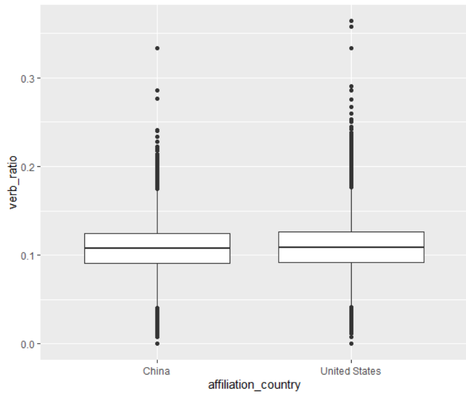
5-2. Category and polarity

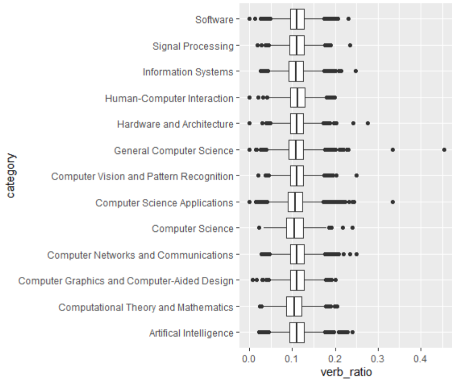
We also tried to analyze the polarity with NLP. Polarity shows the positiveness and negativeness of the sentence. The higher the value, the sentence is more likely to be positive. We thought that the polarity value would be different among the category of computing science. However we found that the polarity value does not differ. Instead, we find that although the difference is insignificant, most sentences tend to be more positive than negative.



5-3. Analysis on the verb ratio
When doing linguistic analysis, we also analyzed the verb ratio in relation to subject category, year, and country, but we couldn't find significant correlation.







5-4. Analysis on sentiment features with respective to continents

We also analyzed on sentiment features(passive_active / sentiment / average length, subjectivity mean) for each "continents", not by their country. After analysis, it turned out that although there was a slight difference in the mean, there was no significant difference on the dataset distributions. The results were printed out by running the following github code: https://github.com/amy-hyunji/CS564/blob/main/sentiment_by_continent.py

Analyzing passive_active mean for each continent ...
africa: 0.7614852108244179
oceania: 0.7629911280101395
europe: 0.7752298457564015
asia: 0.7831195982153585
america: 0.786590076925661
etc: 0.8174702133555002

Analyzing sentiment mean for each continent ...
etc: 0.08836420792330771
america: 0.08913689826934176
europe: 0.0920463522760198
oceania: 0.09357843377427345
asia: 0.09572871149430463
africa: 0.10438220028909116

Analyzing average length mean for each continent ...
asia: 23.023404172984222
africa: 23.311195235335223
europe: 23.855943520037886
oceania: 23.91424790112259
america: 24.22609688306706
etc: 24.88632866314361

Analyzing subjectivity mean for each continent ...
etc: 0.41757890266237396

europe: 0.4292510158123804
america: 0.4302195144287949
africa: 0.43160777092986774
oceania: 0.4375912199908536
asia: 0.447075388143239

5-4. Number of papers for each category for TF-IDF analysis

	Category	Number of papers
1	Software	43903
2	Signal Processing	17140
3	Hardware and Architecture	17807
4	Computer Science Applications	75429
5	Artificial Intelligence	23065
6	Information Systems	22998
7	Computational Theory and Mathematics	11865
8	General Computer Science	85746
9	Computer Networks and Communications	35611
10	Human-Computer Interaction	8545
11	Computer Vision and Pattern Recognition	12882
12	Computer Graphics and Computer-Aided Design	6232

[Table 1]

Category	**Software**	**Computer Science Applications**	**Signal Processing**	**Hardware and Architecture**	**Artificial Intelligence**	**Information Systems**
1	data	data	data	algorithm	subgraphs	data
2	power	network	power	proposed	segmentation	network
3	network	based	network	data	significant	based
4	using	cloud	using	based	network	cloud
5	paper	proposed	paper	performance	small	proposed
6	model	paper	based	decoding	probabilistic	paper
7	time	using	algorithm	polar	methods	using
8	used	algorithm	used	learning	primary	algorithm
9	based	iot	learning	method	using	iot

10	systems	used	systems	hardware	networks	energy

Category	**Computational Theory and Mathematics**	**General Computer Science**	**Computer Networks and Communications**	**Human-Computer Interaction**	**Computer Vision and Pattern Recognition**	**Computer Graphics and Computer-Aided Design**
1	point	data	point	research	research	vision
2	order	model	order	information	information	low
3	free	csp	free	data	data	tools
4	second	based	second	analysis	analysis	vr
5	vertical	implementation	vertical	model	model	support
6	solution	method	solution	tourism	network	people
7	upper	use	upper	paper	tourism	complete
8	outer	rights	outer	development	paper	seeingvr
9	fluid	reserved	fluid	network	development	14
10	plate	used	plate	china	china	application

[Table 2]