Text Classification and Feature Selection

Exercise 1.1: Text Classification Tasks

a)

Classification Task	Type
Genre classification	Hierarchical. A genre class can have multiple sub-genres.
Author gender identification	Flat
Language identification	Flat. However, same language can be different in different regions. For example, English language has some differences in UK, US and Australia. They can be treated as sub class of the parent English or each of them can be a class itself.
Sentiment detection	Flat. Though a class such as 'positive' can be sub-classed into how much positive a feedback is!
Categorization of Wikipedia articles	Hierarchical. It is pretty obvious that an article of a certain topic can have multiple sub topics.
Translationese identification	Flat

b)

Classification Task	Category
Genre classification	Multi-category. A novel can be a combination of mystery, drama and thriller.
Author gender identification	Single-category
Language identification	Single-category for web page, sentence and paragraph since they usually belong to same language in the whole instance. In contrast, a document might contain different languages in a single instance. Like E-mails, letters, user manuals, etc. can be in different languages in a same file.
Sentiment detection	Single-category. The class can be divided such as all types of sentiment can be covered. For customer review of products, the class can be positive, negative, mixed and neutral.
Categorization of Wikipedia articles	Multi-category. An article can cover multiple topics of multiple domains.
Translationese identification	Single-category

c) **Genre classification:** For genre classification there are many features that can be considered. Frequency of common words [1] can be used for classifying genre as well as by the count of punctuation marks which are known as stylometric features. In addition, content based features [2] can also be considered in this type of classification task. One particular way to represent content in the form of word distributions are topics by the help of Latent Dirichlet Allocation (LDA) [3] which extracts topics in the form of word distributions. Then each topic can be treated as a feature. Moreover, the most important characters and their interactions [2] in novels can be used for such kind of genre classification task.

Author gender identification: Words used by authors have some degree of dissimilarity according to their gender. For example, in a text containing sport- related words such as 'cricket', 'beat', 'champion'', coach' and 'league', it has been found that the author of a text containing these words will most likely be male rather than a female. On the other hand, for a text containing words such as 'pink' and 'boyfriend', the probability of the writer being female may be higher than the probability of the writer being male. Some other significant difference can be noticed by their word choices. For example, females use weaker and more sweet-sounding words such as 'dear' and 'oh my goodness', while males tend to use stronger words such as 'damn' [4]. There are also significant differences in the use of different signs (‘?’, ‘!!’, “ ”). Women tend to use more multiple question marks than men. Women also write long sentences than men. So Total number of sentences, Total number of characters, Total number of special characters, Vocabulary richness, Total number of question marks, Total number of exclamation marks, Total number of long and short words, etc, these features can be useful while designing such classifier.

Language identification: This task can be achieved by constructing and computing the language models (LM). For each features fitting the features to different LM and comparing their outcome do the job pretty well. For this purpose, we can consider tokens, characters, suffix and Parts of speech (POS) as the features. Moreover, Text length characteristics such as, the number of sentences, number of tokens, the average sentence length, and average token length might be used as features. In addition, Lexical variety like, the number of unique tokens, the ratio between a unique number of tokens and the overall number of tokens can also be used to predict the language of a text.

Sentiment detection: Word-based approach can be handy in this type of classification task where tag or tag cloud are used to link those sentiments. Sentiments then can be easily read from some counting or calculation of those tags. Parts of speech (POS) [6] can be treated as a very important feature since finding adjectives are very important while expressing opinions. Negations [7] can also be very important feature as they are often discarded in stopword list but here in the sentiment analysis these negation terms are so important. Sometimes sentiments towards a product or service are expressed through phrases as well and so phrase itself is a valuable feature in sentiment detection.

Categorization of Wikipedia articles: Collection of the headers and sub-headers of the section of different Wikipedia articles can be used as a very significant feature in this process. Summary of the entire abstract through keywords and other methods can also be useful. Moreover, a list of Wikipefdia’s main categorization system is also necessary in this process to group together the similar subjects.

Translationese identification: The difference of the frequency of pronouns between original and translated English is very significant according to the study of Koppel et al. [8]. They also showed that the frequency of the cohesive markers like 'therefore', 'thus', 'however', etc. are higher in translated text than the original text. However, according to the works of Ella et al. [9] which is basically the extension of works of Koppel et al. (2011), Functional Words(FW) such as, determiners, conjunctions, prepositions, pronouns, auxiliary verbs, modals, qualifiers can be treated as a very crucial feature as it gives around 97% accuracy in four different corpus while used as feature. According to their study, when char tri-grams are used as feature it also gives pretty high accuracy while classifying.

References:

[1] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Text genre detection using common word frequencies," in Proc. 18th conference on Computational linguistics, vol. 2, Association for Computational Linguistics, 2000, pp. 808–814.

[2] Lena Hettinger, Martin Becker, Isabella Reger, Fotis Jannidis, Andreas Hotho, " Genre Classification on German Novels," in 2015 26th International Workshop on Database and Expert Systems Applications (DEXA).

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," in the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.

[4] Adelaide Haas, "Male and Female Spoken Language Differences: Stereotypes and Evidence," in Psychological Bulletin 1979, vol. 86, No. 3, 615-626.

[5] Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow, "Native tongues, lost and found: Resources and empirical evaluations in native language identification," in Proceedings of COLING, The COLING 2012 Organizing Committee, 2012, pages 2585–2602.

[6] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey," in Ain Shams Engineering Journal vol. 5, December 2014, Pages 1093-1113.

[7] Yelena Mejova, Padmini Srinivasan, "Exploring Feature Definition and Selection for Sentiment Classifiers," in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[8] Moshe Koppel and Noam Ordan, "Translationese and its dialects," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 1318–1326, Portland, Oregon, USA, June 2011.

[9] Ella Rabinovich, Shuly Wintner, "Unsupervised Identification of Translationese," in Transactions of the Association for Computational Linguistics, vol. 3, pp. 419–432, 2015.

1.2 PMI for Topic Categorization

Using the point-wise mutual information (PMI) measure, we made feature selection in text classification.

d) Using the functions that compute and return the expected PMI value and maximum PMI value, we ranked words based on these two criteria.

A table of the top 20 features according to the expected PMI:

	Word	Expected PMI	Counts by categories
1	dangerously	0.0	[3, 3, 3, 3]
2	handshake	0.0	[3, 3, 3, 3]
3	binge	0.0	[2, 2, 2, 2]
4	bombshell	0.0	[2, 2, 2, 2]
5	landmarks	0.0	[2, 2, 2, 2]
6	leighton	0.0	[2, 2, 2, 2]
7	sly	0.0	[2, 2, 2, 2]
8	spurn	0.0	[1, 1, 1, 1]
9	culpa	0.0	[1, 1, 1, 1]
10	alerted	0.0	[1, 1, 1, 1]
11	expanse	0.0	[1, 1, 1, 1]
12	decks	0.0	[1, 1, 1, 1]
13	reigned	0.0	[1, 1, 1, 1]
14	leonardo	0.0	[1, 1, 1, 1]
15	paragraphs	0.0	[1, 1, 1, 1]
16	hamlet	0.0	[1, 1, 1, 1]
17	malaise	0.0	[1, 1, 1, 1]
18	innocuous	0.0	[1, 1, 1, 1]
19	swiped	0.0	[1, 1, 1, 1]
20	squaring	0.0	[1, 1, 1, 1]

A table of the top 20 features according to the expected PMI:

	Word	Maximum PMI	Category	Counts by categories
1	fullquote	2.0	Business	[0, 0, 3626, 0]
2	aspx	2.0	Business	[0, 0, 1813, 0]
3	quickinfo	2.0	Business	[0, 0, 1813, 0]
4	sadr	2.0	World	[467, 0, 0, 0]
5	hamas	2.0	World	[439, 0, 0, 0]

6	falluja	2.0	World	[364, 0, 0, 0]
7	mozilla	2.0	Sci/Tech	[0, 0, 0, 362]
8	ziff	2.0	Sci/Tech	[0, 0, 0, 278]
9	zarqawi	2.0	World	[246, 0, 0, 0]
10	jhtml	2.0	Sci/Tech	[0, 0, 0, 244]
11	financequotelookup	2.0	Sci/Tech	[0, 0, 0, 242]
12	qtype	2.0	Sci/Tech	[0, 0, 0, 242]
13	sym	2.0	Sci/Tech	[0, 0, 0, 242]
14	infotype	2.0	Sci/Tech	[0, 0, 0, 242]
15	qcat	2.0	Sci/Tech	[0, 0, 0, 242]
16	qaida	2.0	World	[222, 0, 0, 0]
17	preseason	2.0	Sports	[0, 218, 0, 0]
18	mosul	2.0	World	[217, 0, 0, 0]
19	sunni	2.0	World	[213, 0, 0, 0]
20	treasuries	2.0	Business	[0, 0, 213, 0]

e) We can notice that all the words with highest PMI have the same value of the metric in both tables. In the table of the top 20 features according to *the expected PMI* we can see that the highest value of the metric is related to those words occurring equally frequent over all categories. Therefore, these words are characteristic to all four corpora. In the table of the top 20 features according to *the maximum PMI* we can see that the highest value of the metric is related to those words occurring frequently just in single category. Therefore, they are inherent to the specific corpus.
In general, the words from the first table are more commonly used in English speech.

1 Text Classification and Feature Selection

1.3 χ^2 -Feature selection

$$\chi^2(\text{mathematics,composition}) = \frac{370 * (30 * 340 - 120 * 60)^2}{(30 + 120) * (30 + 60) * (120 + 340) * (60 + 340)} = 13.405797101$$

$$\chi^2(\text{mathematics,gravity}) = \frac{370 * (3 * 367 - 87 * 87)^2}{(3 + 87) * (3 + 87) * (367 + 87) * (376 + 87)} = 9.27139367$$

$$\chi^2(\text{mathematics,differential}) = \frac{370 * (50 * 320 - 40 * 50)^2}{(50 + 40) * (50 + 50) * (320 + 40) * (320 + 50)} = 60.49382716$$

$$\chi^2(\text{mathematics,theory}) = \frac{370 * (7 * 363 - 83 * 23)^2}{(7 + 83) * (7 + 23) * (363 + 83) * (363 + 23)} = 0,317943502$$

$$\chi^2(\text{chemistry,composition}) = \frac{370(43 * 337 - 107 * 7)^2}{(43 + 107) * (43 + 7) * (337 + 107) * (337 + 7)} = 60.995660207$$

$$\chi^2(\text{chemistry,gravity}) = \frac{370 * (-50 * 90)^2}{(370 + 50) * (370 + 90)} = 387810.55900621$$

$$\chi^2(\text{chemestry,differential}) = \frac{370 * (2 * 368 - 48 * 98)^2}{(2 + 48) * (2 + 98) * (368 + 48) * (368 + 98)} = 6.010295147$$

$$\chi^2(\text{chemistry,theory}) = \frac{370 * (5 * 365 - 45 * 25)^2}{(5 + 45) * (5 + 25) * (365 + 45) * (365 + 25)} = 0.755889097$$

$$\chi^2(\text{astronomy,composition}) = \frac{370 * (47 * 323 - 103 * 73)^2}{(47 + 103) * (47 + 73) * (323 + 103) * (323 + 73)} = 7.15333772$$

$$\chi^2(\text{astronomy,gravity}) = \frac{370 * (53 * 317 - 77 * 37)^2}{(53 + 77) * (53 + 37) * (317 + 77) * (317 + 37)} = 44,135628321$$

$$\chi^2(\text{astronomy,differential}) = \frac{370 * (19 * 351 - 111 * 81)}{(19 + 111) * (19 + 81) * (351 + 111) * (351 + 81)} = 0.768877373$$

$$\chi^2(\text{astronomy,theory}) = \frac{370 * (11 * 359 - 19 * 119)^2}{(11 + 19) * (11 + 119) * (359 + 19) * (359 + 119)} = 714.906766087$$

$$\chi^2(\text{physics,composition}) = \frac{370 * (30 * 340 - 120 * 70)}{(30 + 120) * (30 + 70) * (340 + 120) * (340 + 70)} = 0.412301166$$

$$\chi^2(\text{physics,gravity}) = \frac{370 * (34 * 336 - 56 * 56)^2}{(34 + 56) * (34 + 56) * (336 + 56) * (336 + 56)} = 20.419450743$$

$$\chi^2(\text{physics,differential}) = \frac{370 * (29 * 341 - 71 * 71)^2}{(29 + 71) * (29 + 71) * (341 + 71) * (341 + 71)} = 5.123096239$$

$$\chi^2(\text{physics,theory}) = \frac{370 * (7 * 363 - 23 * 93)^2}{(7 + 23) * (7 + 93) * (363 + 23) * (363 + 93)} = 0.113234933$$

$$\chi^2 avg(\text{composition}) = \frac{90}{370} * 13.41 + \frac{50}{370} * 61 + \frac{130}{370} * 7.15 + \frac{100}{370} * 0.41 = 15,221891892$$

This means compostion would be a good feature for chemisry

$$\chi^2 avg(\text{gravity}) = \frac{90}{370} * 9.27 + \frac{50}{370} * 387810 + \frac{130}{370} * 44.1 + \frac{100}{370} * 20.42 = 2311152,353918919$$

This means gravity would be a good feature for chemistry

$$\chi^2 avg(\text{differential}) = \frac{90}{370} * 60.5 + \frac{50}{370} * 6.01 + \frac{130}{370} * 0.77 + \frac{100}{370} * 5.12 = 17,182702703$$

This means differential would be a good feature for mathematics

$$\chi^2 avg(\text{theory}) = \frac{90}{370} * 0.32 + \frac{50}{370} * 0.76 + \frac{130}{370} * 714.9 + \frac{100}{370} * 0.11 = 251,391351351$$

This means theory would be a good feature for physics
Since we have no feature above the threshhold for astronomy the next best feature would be gravity. This is caused by the realy high value caused by 0 cases of gravity in the class chemistry.