CONTENT

WHY IS THIS IMPORTANT?

- **Main idea:** Investigating the integration of machine learning models in GitHub pull requests.
- **Process:** Analyzing a set of successful GitHub pull requests and identifying patterns and factors that contribute to their acceptance.
- **Final Goal:** Hope to improve the efficiency of software development and its processes.

LASSONDE SCHOOL OF ENGINEERING

EECS 4080 SU 2023

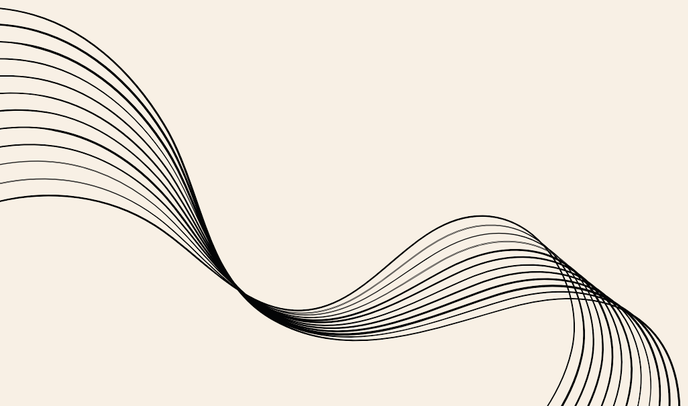
AN EMPIRICAL ANALYSIS OF PULL REQUESTS:

EXPLORING THE INTEGRATION OF MACHINE LEARNING MODELS IN SOFTWARE DEVELOPMENT

CLAUDIA FARKAS
PROFESSOR MALEKNAZ NAYEBI

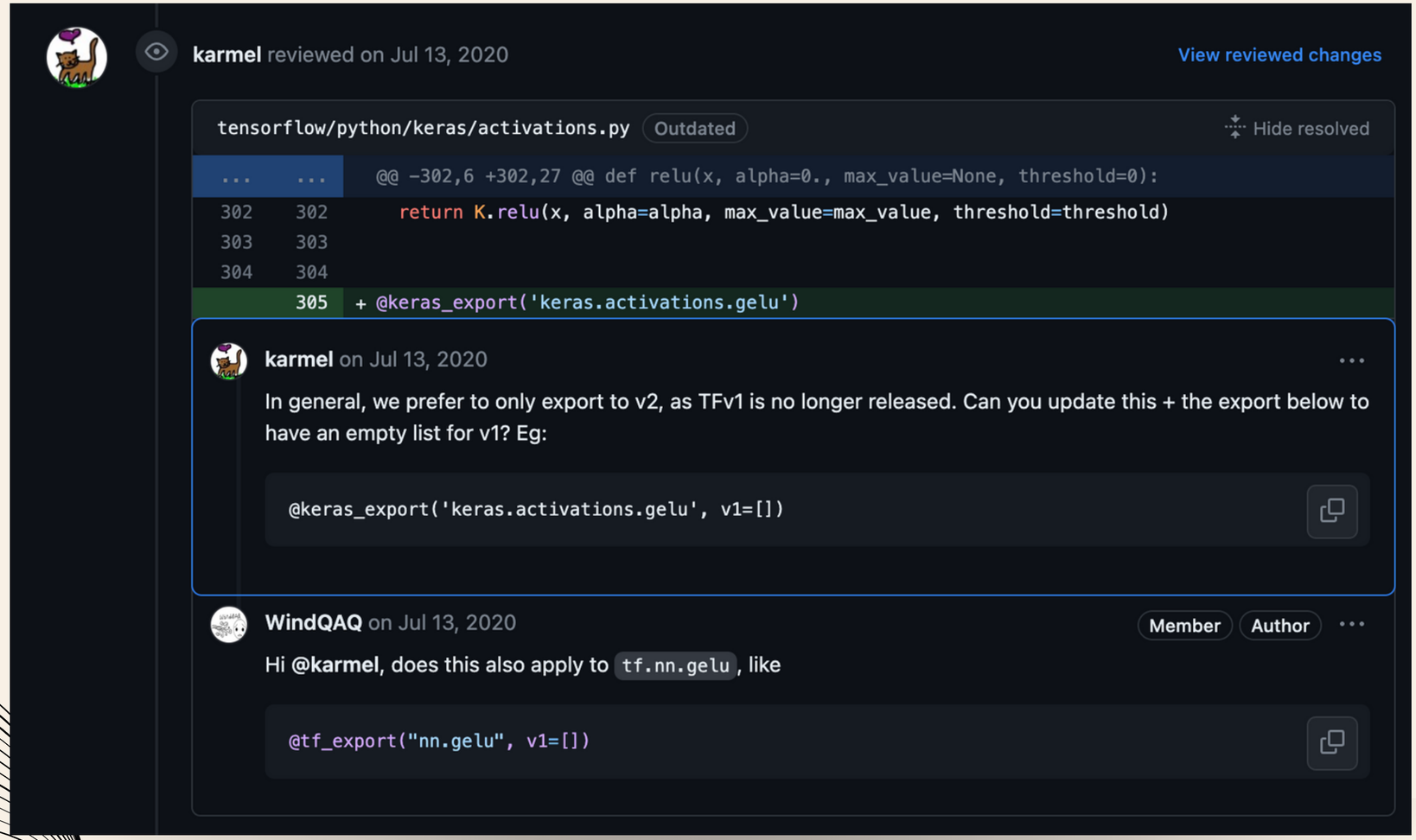
THE RESEARCH QUESTION

How Can We Identify and Classify Messages Contained in a Pull Request for Machine Learning Initiatives?



SOME BACKGROUND

https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453906256



What is a PullRequest?

"Pull requests let you tell others about changes you've pushed to a branch in a repository on GitHub. Once a pull request is opened, you can discuss and review the potential changes with collaborators and add follow-up commits before your changes are merged into the base branch." -

https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/proposing-changes-to-your-work-with-pull-requests/about-pull-requests

GitHub

TensorFlow

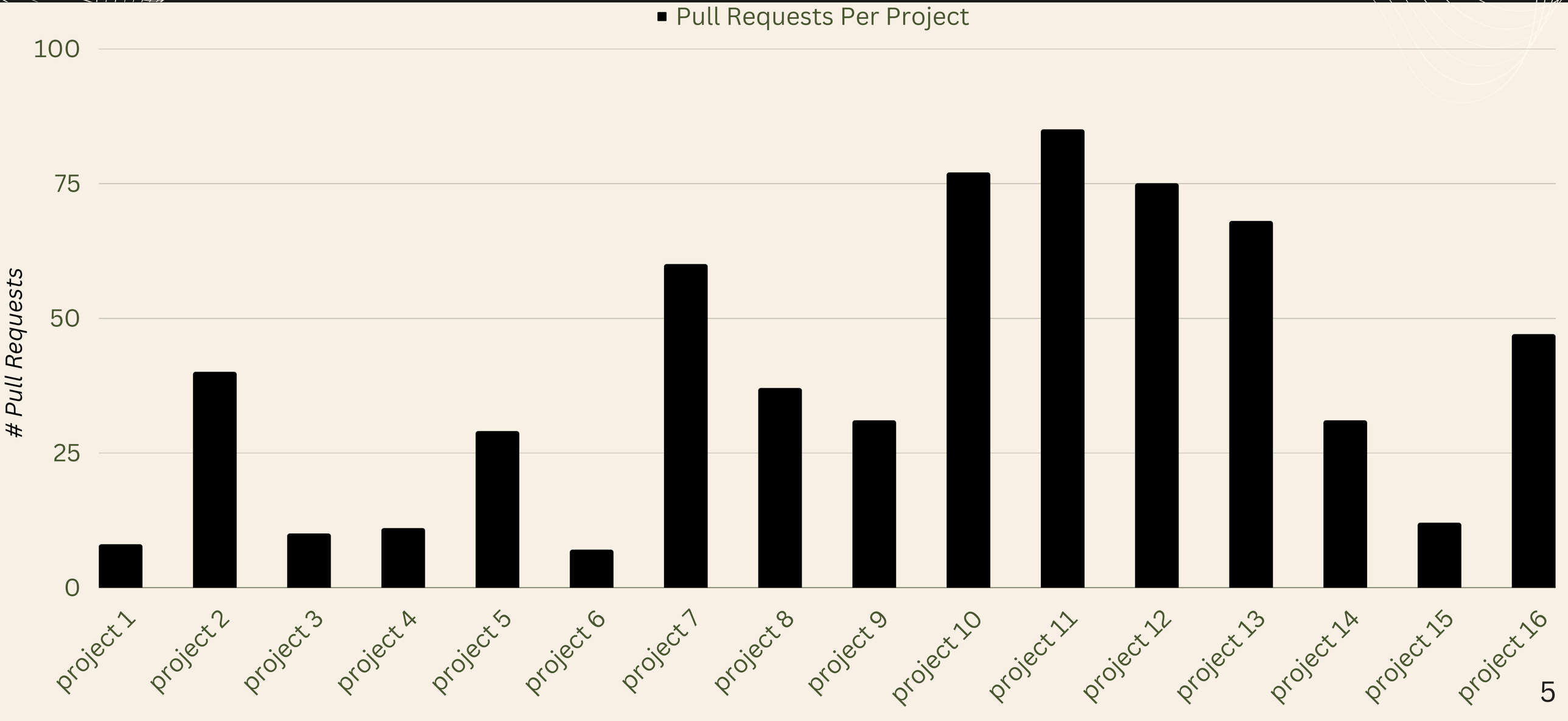
THE DATA

(sample data)

Date	Type	ID	Name	Body	URL	Comment Type
2020-07-08T01:38:31Z	Commit	8a917da9fad79e74772c8f3bef4ff5f825b88d5c	WindQAQ	COMMENTED In general we prefer to only export to v2 as TFv1 is no longer released. Can you update this + the export below to have an empty list for v1? Eg: ``` @keras_export('keras.activations.gelu' v1=[]) ```	https://github.com/tensorflow/tensorflow/commit/8a917da9fad79e74772c8f3bef4ff5f825b88d5c	Conventional Review

16 projects in TensorFlow , each varying from 7 to 85 messages

THE DATA


Pull Requests Per Project
100
75
50
25
0
Pull Requests
project 1
project 2
project 3
project 4
project 5
project 6
project 7
project 8
project 9
project 10
project 11
project 12
project 13
project 14
project 15
project 16

WHAT NOW?

1) Manual Classification

Utilizing excel to manually represent the results of the classification.

2) Supervised Learning

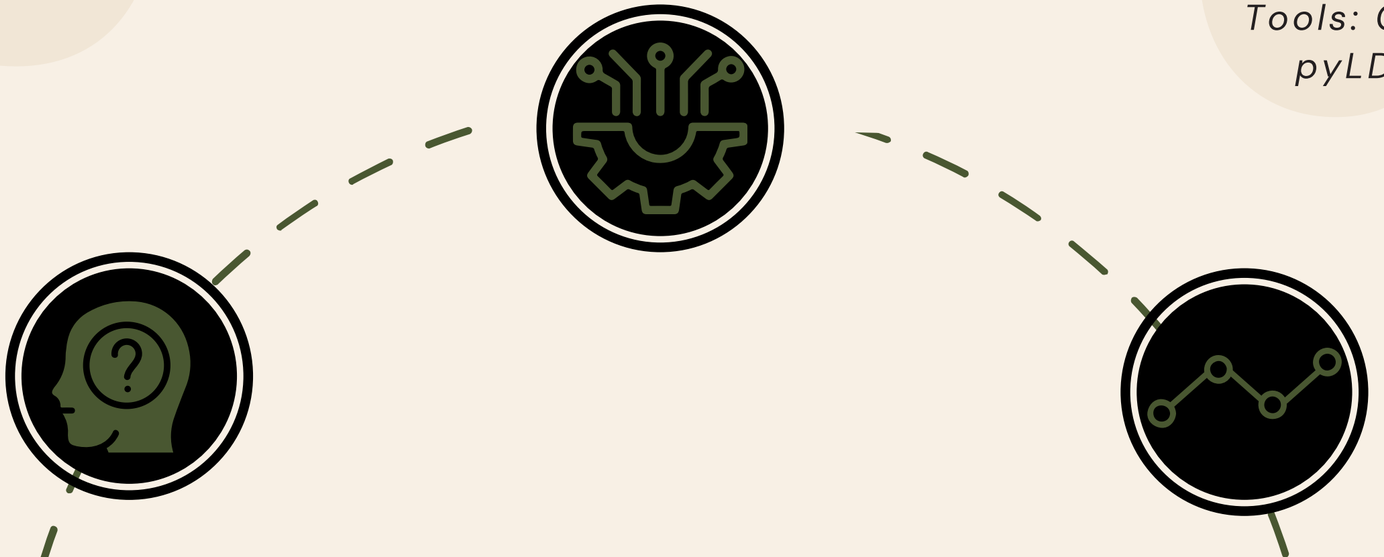
Training a ML model with labeled data, guiding it to make accurate predictions on unseen data.

Tools: Classification Report, Heat Map

3) Unsupervised Learning

Training a ML model without labeled data, allowing it to identify patterns or structures on its own.

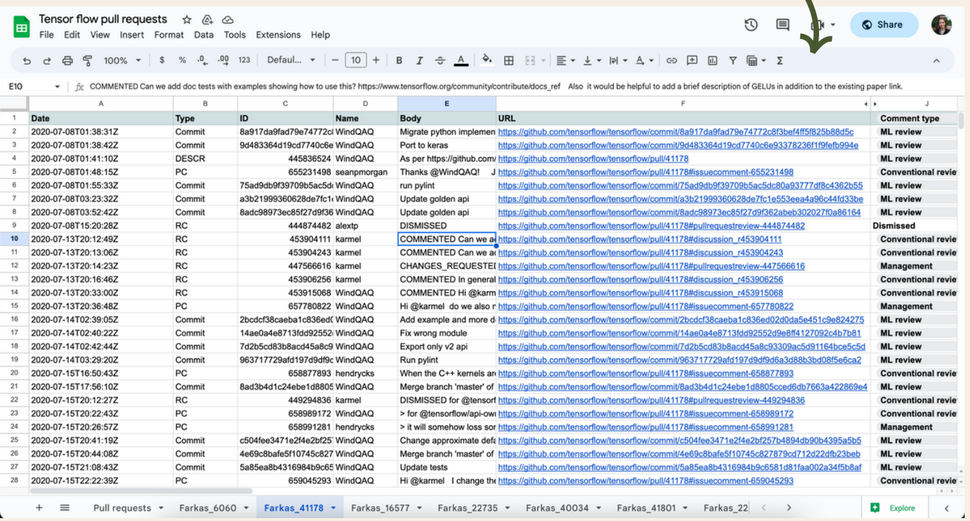
Tools: Gensim and the pyLDAvis library



1) MANUAL CLASSIFICATION

CATEGORIZATION PROCESS

1.Data Collection



Date	Type	ID	Name	Body	URL	Comment type
2020-07-08T01:38:31Z	Commit	8a917da9fad79e74772c	WindQAQ	Migrate python implemen	https://github.com/tensorflow/tensorflow/commit/8a917da9fad79e74772c8f3bef4ff5f825b88d5c	ML review
2020-07-08T01:38:42Z	Commit	9d483364d19cd7740c6e	WindQAQ	Port to keras	https://github.com/tensorflow/tensorflow/commit/9d483364d19cd7740c6e93378236f1f9fefb994e	ML review
2020-07-08T01:41:10Z	DESCR	445836524	WindQAQ	As per https://github.com/	https://github.com/tensorflow/tensorflow/pull/41178	ML review
2020-07-08T01:48:15Z	PC	655231498	seanpmorgan	Thanks @WindQAQ! :)	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-655231498	Conventional revie
2020-07-08T01:55:33Z	Commit	75ad9db9f39709b5ac5dc	WindQAQ	run pylint	https://github.com/tensorflow/tensorflow/commit/75ad9db9f39709b5ac5dc80a93777df8c4362b55	ML review
2020-07-08T03:23:32Z	Commit	a3b21999360628de7fc1e	WindQAQ	Update golden api	https://github.com/tensorflow/tensorflow/commit/a3b21999360628de7fc1e553eea4a96c44fd33be	ML review
2020-07-08T03:52:42Z	Commit	8adc98973ec85f27d9f36	WindQAQ	Update golden api	https://github.com/tensorflow/tensorflow/commit/8adc98973ec85f27d9f362abeb3020271f0a86164	ML review
2020-07-08T15:20:28Z	RC	444874482	alextp	DISMISSED	https://github.com/tensorflow/tensorflow/pull/41178#pullrequestreview-444874482	Dismissed
2020-07-13T20:12:49Z	RC	453904111	karmel	COMMENTED Can we a	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453904111	Conventional revie
2020-07-13T20:13:06Z	RC	453904243	karmel	COMMENTED Can we a	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453904243	Conventional revie
2020-07-13T20:14:23Z	RC	447566616	karmel	CHANGES_REQUESTE	https://github.com/tensorflow/tensorflow/pull/41178#pullrequestreview-447566616	Management
2020-07-13T20:16:46Z	RC	453906256	karmel	COMMENTED In general	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453906256	Conventional revie
2020-07-13T20:33:00Z	RC	453915068	WindQAQ	COMMENTED Hi @karm	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453915068	Conventional revie
2020-07-13T20:36:48Z	PC	657780822	WindQAQ	Hi @karmel do we also r	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-657780822	Management
2020-07-14T02:39:05Z	Commit	2bcdcf38caeba1c836ed(	WindQAQ	Add example and more d	https://github.com/tensorflow/tensorflow/commit/2bcdcf38caeba1c836ed02d00da5e451c9e824275	ML review
2020-07-14T02:40:22Z	Commit	14ae0a4e8713fdd92552	WindQAQ	Fix wrong module	https://github.com/tensorflow/tensorflow/commit/14ae0a4e8713fdd92552d9e8ff4127092c4b7b81	ML review
2020-07-14T02:42:44Z	Commit	7d2b5cd83b8acd45a8c9	WindQAQ	Export only v2 api	https://github.com/tensorflow/tensorflow/commit/7d2b5cd83b8acd45a8c93309ac5d91164bce5c5d	ML review
2020-07-14T03:29:20Z	Commit	963717729afd197d9df9c	WindQAQ	Run pylint	https://github.com/tensorflow/tensorflow/commit/963717729afd197d9df9d6a3d88b3bd08f5e6ca2	ML review
2020-07-15T16:50:43Z	PC	658877893	hendrycks	When the C++ kernels an	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-658877893	Conventional revie
2020-07-15T17:56:10Z	Commit	8ad3b4d1c24ebe1d8805	WindQAQ	Merge branch 'master' of	https://github.com/tensorflow/tensorflow/commit/8ad3b4d1c24ebe1d8805cced6db7663a422869e4	ML review
2020-07-15T20:12:27Z	RC	449294836	karmel	DISMISSED for @tensorf	https://github.com/tensorflow/tensorflow/pull/41178#pullrequestreview-449294836	Conventional revie
2020-07-15T20:22:43Z	PC	658989172	WindQAQ	> for @tensorflow/api-ow	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-658989172	Conventional revie
2020-07-15T20:26:57Z	PC	658991281	hendrycks	> it will somehow loss scr	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-658991281	Management
2020-07-15T20:41:19Z	Commit	c504fee3471e2f4e2bf25	WindQAQ	Change approximate defa	https://github.com/tensorflow/tensorflow/commit/c504fee3471e2f4e2bf257b4894db90b4395a5b5	ML review
2020-07-15T20:44:08Z	Commit	4e69c8bafe5f10745c827	WindQAQ	Merge branch 'master' of	https://github.com/tensorflow/tensorflow/commit/4e69c8bafe5f10745c827879cd712d22dfb23beb	ML review
2020-07-15T21:08:43Z	Commit	5a85ea8b4316984b9c65	WindQAQ	Update tests	https://github.com/tensorflow/tensorflow/commit/5a85ea8b4316984b9c6581d81faa002a34f5b8af	ML review
2020-07-15T22:22:39Z	PC	659045293	WindQAQ	Hi @karmel I change the	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-659045293	Conventional revie

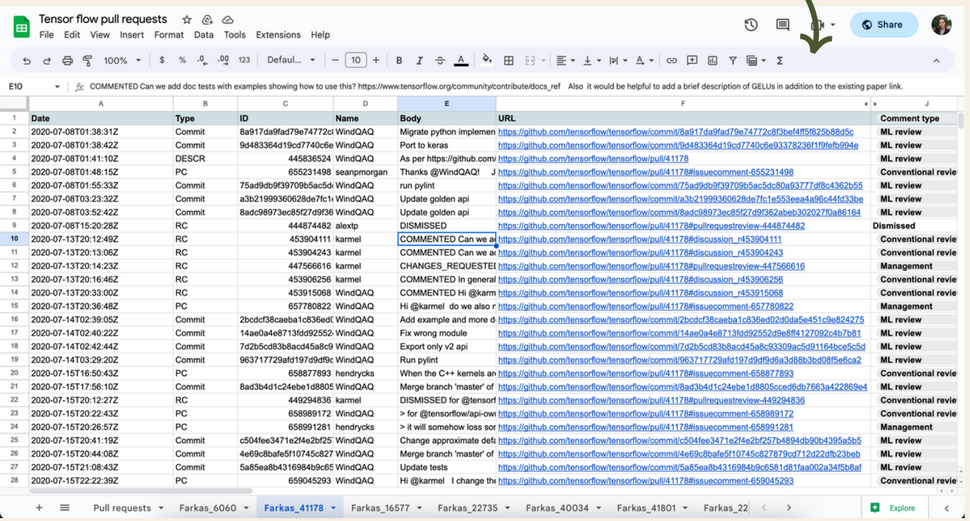
2.Categorization

1

3.Visual Representation

CATEGORIZATION PROCESS

1.Data Collection


Tensor flow pull requests

File Edit View Insert Format Data Tools Extensions Help

Share

E10 COMMENTED Can we add doc tests with examples showing how to use this? https://www.tensorflow.org/community/contribute/docs_ref Also it would be helpful to add a brief description of GELUs in addition to the existing paper link.

	A	B	C	D	E	F	J
1	Date	Type	ID	Name	Body	URL	Comment type
2	2020-07-08T01:38:31Z	Commit	8a917da9fad79e74772c	WindQAQ	Migrate python implemen	https://github.com/tensorflow/tensorflow/commit/8a917da9fad79e74772c8f3bef4ff5f825b88d5c	ML review
3	2020-07-08T01:38:42Z	Commit	9d483364d19cd7740c6e	WindQAQ	Port to keras	https://github.com/tensorflow/tensorflow/commit/9d483364d19cd7740c6e93378236f1f9fefb994e	ML review
4	2020-07-08T01:41:10Z	DESCR	445836524	WindQAQ	As per https://github.com/	https://github.com/tensorflow/tensorflow/pull/41178	ML review
5	2020-07-08T01:48:15Z	PC	655231498	seanpmorgan	Thanks @WindQAQ!	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-655231498	Conventional revie
6	2020-07-08T01:55:33Z	Commit	75ad9db9f39709b5ac5dc	WindQAQ	run pylint	https://github.com/tensorflow/tensorflow/commit/75ad9db9f39709b5ac5dc80a93777df8c4362b55	ML review
7	2020-07-08T03:23:32Z	Commit	a3b21999360628de7fc1	WindQAQ	Update golden api	https://github.com/tensorflow/tensorflow/commit/a3b21999360628de7fc1e553eea4a96c44fd33be	ML review
8	2020-07-08T03:52:42Z	Commit	8adc98973ec85f27d9f36	WindQAQ	Update golden api	https://github.com/tensorflow/tensorflow/commit/8adc98973ec85f27d9f362abeb3020271f0a86164	ML review
9	2020-07-08T15:20:28Z	RC	444874482	alextp	DISMISSED	https://github.com/tensorflow/tensorflow/pull/41178#pullrequestreview-444874482	Dismissed
10	2020-07-13T20:12:49Z	RC	453904111	karmel	COMMENTED Can we a	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453904111	Conventional revie
11	2020-07-13T20:13:06Z	RC	453904243	karmel	COMMENTED Can we a	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453904243	Conventional revie
12	2020-07-13T20:14:23Z	RC	447566616	karmel	CHANGES_REQUESTE	https://github.com/tensorflow/tensorflow/pull/41178#pullrequestreview-447566616	Management
13	2020-07-13T20:16:46Z	RC	453906256	karmel	COMMENTED In general	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453906256	Conventional revie
14	2020-07-13T20:33:00Z	RC	453915068	WindQAQ	COMMENTED Hi @karm	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453915068	Conventional revie
15	2020-07-13T20:36:48Z	PC	657780822	WindQAQ	Hi @karmel do we also r	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-657780822	Management
16	2020-07-14T02:39:05Z	Commit	2bcdcf38caeba1c836ed(	WindQAQ	Add example and more d	https://github.com/tensorflow/tensorflow/commit/2bcdcf38caeba1c836ed02d0da5e451c9e824275	ML review
17	2020-07-14T02:40:22Z	Commit	14ae0a4e8713fdd92552	WindQAQ	Fix wrong module	https://github.com/tensorflow/tensorflow/commit/14ae0a4e8713fdd92552d9e8ff4127092c4b7b81	ML review
18	2020-07-14T02:42:44Z	Commit	7d2b5cd83b8acd45a8c9	WindQAQ	Export only v2 api	https://github.com/tensorflow/tensorflow/commit/7d2b5cd83b8acd45a8c93309ac5d91164bce5c5d	ML review
19	2020-07-14T03:29:20Z	Commit	963717729afd197d9df9c	WindQAQ	Run pylint	https://github.com/tensorflow/tensorflow/commit/963717729afd197d9df9d6a3d88b3bd08f5e6ca2	ML review
20	2020-07-15T16:50:43Z	PC	658877893	hendrycks	When the C++ kernels an	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-658877893	Conventional revie
21	2020-07-15T17:56:10Z	Commit	8ad3b4d1c24ebe1d8805	WindQAQ	Merge branch 'master' of	https://github.com/tensorflow/tensorflow/commit/8ad3b4d1c24ebe1d8805cced6db7663a422869e4	ML review
22	2020-07-15T20:12:27Z	RC	449294836	karmel	DISMISSED for @tensorf	https://github.com/tensorflow/tensorflow/pull/41178#pullrequestreview-449294836	Conventional revie
23	2020-07-15T20:22:43Z	PC	658989172	WindQAQ	> for @tensorflow/api-ow	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-658989172	Conventional revie
24	2020-07-15T20:26:57Z	PC	658991281	hendrycks	> it will somehow loss sor	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-658991281	Management
25	2020-07-15T20:41:19Z	Commit	c504fee3471e2f4e2bf25	WindQAQ	Change approximate def	https://github.com/tensorflow/tensorflow/commit/c504fee3471e2f4e2bf257b4894db90b4395a5b5	ML review
26	2020-07-15T20:44:08Z	Commit	4e69c8bafe5f10745c827	WindQAQ	Merge branch 'master' of	https://github.com/tensorflow/tensorflow/commit/4e69c8bafe5f10745c827879cd712d22dfb23beb	ML review
27	2020-07-15T21:08:43Z	Commit	5a85ea8b4316984b9c65	WindQAQ	Update tests	https://github.com/tensorflow/tensorflow/commit/5a85ea8b4316984b9c6581d81faa002a34f5b8af	ML review
28	2020-07-15T22:22:39Z	PC	659045293	WindQAQ	Hi @karmel I change the	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-659045293	Conventional revie

Pull requests | Farkas_6060 | Farkas_41178 | Farkas_16577 | Farkas_22735 | Farkas_40034 | Farkas_41801 | Farkas_22 | Explore

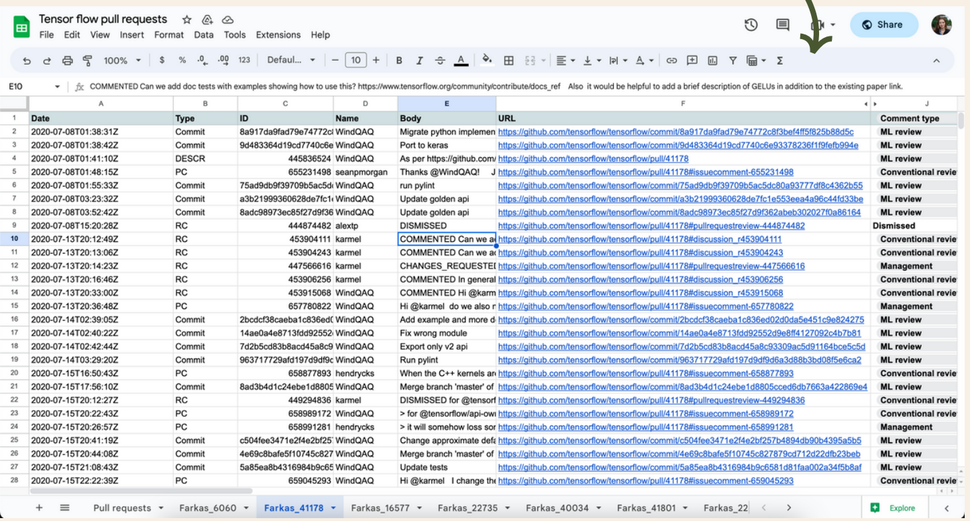
2.Categorization

J	K	L	M	N
Comment type	Nature of comments	Contributed artifacts	Why	Relevent SDLC or ML phase
ML review	Clarification	Code	Code changes	SWEBOK-Software engineering process
ML review	Clarification	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	other	Assisting the changes	SWEBOK-Software engineering process
Conventional review	Comment	Review	Question	SWEBOK-Software engineering management
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
Dismissed	Dismissed	Dismissed	Dismissed	Dismissed
Conventional review	Comment	Link	Asked question	SWEBOK-Software engineering management
Conventional review	Comment	Code	Asked question	SWEBOK-Software engineering management
Management	Pinging	other	Process managment	SWEBOK-Software engineering management
Conventional review	Clarification	Code	Asked question	SWEBOK-Software engineering process
Conventional review	Comment	Code	Question	SWEBOK-Software engineering management
Management	Clarification	Code	Assisting the changes	SWEBOK-Software engineering management
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
Conventional review	Feedback	Code	Explaining changes	SWEBOK-Software engineering management
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
Conventional review	Comment	other	Explaining plan	SWEBOK-Software engineering management
Conventional review	Comment	other	Code review	SWEBOK-Software engineering management
Management	Code review	other	Assisting the changes	SWEBOK-Software engineering management
ML review	Code review	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code review	Code	Approved changes	SWEBOK-Software engineering process
ML review	Code review	Code	Approved changes	SWEBOK-Software engineering management
Conventional review	Request a review	Review	Code review	SWEBOK-Software engineering management

3.Visual Representation

CATEGORIZATION PROCESS

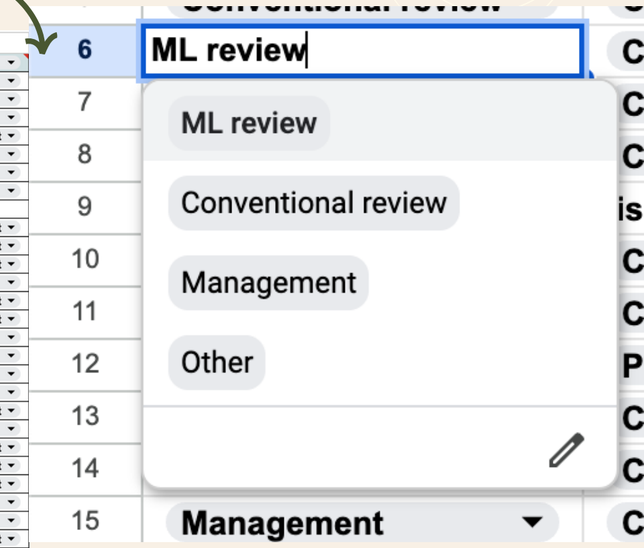
1.Data Collection



	A	B	C	D	E	F	J
1	Date	Type	ID	Name	Body	URL	Comment type
2	2020-07-08T01:38:31Z	Commit	8a917da9fad79e74772c	WindQAQ	Migrate python implemen	https://github.com/tensorflow/tensorflow/commit/8a917da9fad79e74772c8f3bef4ff5f825b88d5c	ML review
3	2020-07-08T01:38:42Z	Commit	9d483364d19cd7740c6e	WindQAQ	Port to keras	https://github.com/tensorflow/tensorflow/commit/9d483364d19cd7740c6e93378236f1f9fefb994e	ML review
4	2020-07-08T01:41:10Z	DESCR	445836524	WindQAQ	As per https://github.com/	https://github.com/tensorflow/tensorflow/pull/41178	ML review
5	2020-07-08T01:48:15Z	PC	655231498	seanspmorgan	Thanks @WindQAQ!	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-655231498	Conventional revie
6	2020-07-08T01:55:33Z	Commit	75ad9db9f39709b5ac5dc	WindQAQ	run pylint	https://github.com/tensorflow/tensorflow/commit/75ad9db9f39709b5ac5dc80a93777df8c4362b55	ML review
7	2020-07-08T03:23:32Z	Commit	a3b21999360628de7fc1e	WindQAQ	Update golden api	https://github.com/tensorflow/tensorflow/commit/a3b21999360628de7fc1e553eea4a96c44fd33be	ML review
8	2020-07-08T03:52:42Z	Commit	8adc98973ec85f27d9f36	WindQAQ	Update golden api	https://github.com/tensorflow/tensorflow/commit/8adc98973ec85f27d9f362abeb3020271f0a86164	ML review
9	2020-07-08T15:20:28Z	RC	444874482	alextp	DISMISSED	https://github.com/tensorflow/tensorflow/pull/41178#pullrequestreview-444874482	Dismissed
10	2020-07-13T20:12:49Z	RC	453904111	karmel	COMMENTED Can we a	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453904111	Conventional revie
11	2020-07-13T20:13:06Z	RC	453904243	karmel	COMMENTED Can we a	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453904243	Conventional revie
12	2020-07-13T20:14:23Z	RC	447566616	karmel	CHANGES_REQUESTE	https://github.com/tensorflow/tensorflow/pull/41178#pullrequestreview-447566616	Management
13	2020-07-13T20:16:46Z	RC	453906256	karmel	COMMENTED In general	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453906256	Conventional revie
14	2020-07-13T20:33:00Z	RC	453915068	karmel	COMMENTED Hi @karm	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453915068	Conventional revie
15	2020-07-13T20:36:48Z	PC	657780822	WindQAQ	Hi @karmel do we also r	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-657780822	Management
16	2020-07-14T02:39:05Z	Commit	2bcdcf38caeba1c836ed(	WindQAQ	Add example and more d	https://github.com/tensorflow/tensorflow/commit/2bcdcf38caeba1c836ed02d0da5e451c9e824275	ML review
17	2020-07-14T02:40:22Z	Commit	14ae0a4e8713fdd92552	WindQAQ	Fix wrong module	https://github.com/tensorflow/tensorflow/commit/14ae0a4e8713fdd92552d9e8ff4127092c4b7b81	ML review
18	2020-07-14T02:42:44Z	Commit	7d2b5cd83b8acd45a8c9	WindQAQ	Export only v2 api	https://github.com/tensorflow/tensorflow/commit/7d2b5cd83b8acd45a8c93309ac5d91164bce5c5d	ML review
19	2020-07-14T03:29:20Z	Commit	963717729afd197d9df9c	WindQAQ	Run pylint	https://github.com/tensorflow/tensorflow/commit/963717729afd197d9df9d6a3d88b3bd08f5e6ca2	ML review
20	2020-07-15T16:50:43Z	PC	658877893	hendrycks	When the C++ kernels an	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-658877893	Conventional revie
21	2020-07-15T17:56:10Z	Commit	8ad3b4d1c24ebe1d8805	WindQAQ	Merge branch 'master' of	https://github.com/tensorflow/tensorflow/commit/8ad3b4d1c24ebe1d8805cced6db7663a422869e4	ML review
22	2020-07-15T20:12:27Z	RC	449294836	karmel	DISMISSED for @tensorf	https://github.com/tensorflow/tensorflow/pull/41178#pullrequestreview-449294836	Conventional revie
23	2020-07-15T20:22:43Z	PC	658989172	WindQAQ	> for @tensorflow/api-ow	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-658989172	Conventional revie
24	2020-07-15T20:26:57Z	PC	658991281	hendrycks	> it will somehow loss scr	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-658991281	Management
25	2020-07-15T20:41:19Z	Commit	c504fee3471e2f4e2bf25	WindQAQ	Change approximate def	https://github.com/tensorflow/tensorflow/commit/c504fee3471e2f4e2bf257b4894db90b4395a5b5	ML review
26	2020-07-15T20:44:08Z	Commit	4e69c8bafe5f10745c827	WindQAQ	Merge branch 'master' of	https://github.com/tensorflow/tensorflow/commit/4e69c8bafe5f10745c827879cd712d22dfb23beb	ML review
27	2020-07-15T21:08:43Z	Commit	5a85ea8b4316984b9c65	WindQAQ	Update tests	https://github.com/tensorflow/tensorflow/commit/5a85ea8b4316984b9c6581d81faa002a34f5b8af	ML review
28	2020-07-15T22:22:39Z	PC	659045293	WindQAQ	Hi @karmel I change the	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-659045293	Conventional revie

2.Categorization

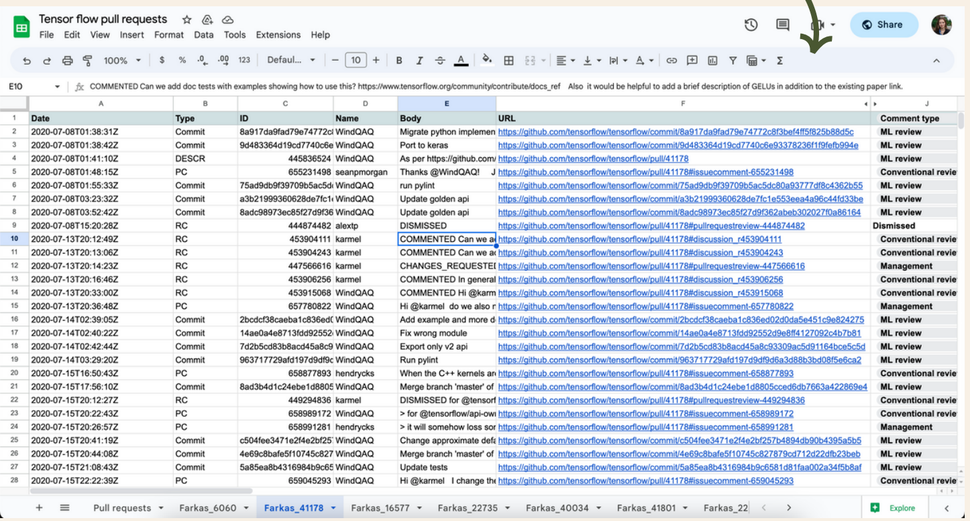
J	K	L	M	N
Comment type	Nature of comments	Contributed artifacts	Why	Relevent SDLC or ML phase
ML review	Clarification	Code	Code changes	SWEBOK-Software engineering process
ML review	Clarification	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	other	Assisting the changes	SWEBOK-Software engineering process
Conventional review	Comment	Review	Question	SWEBOK-Software engineering management
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
Dismissed	Dismissed	Dismissed	Dismissed	Dismissed
Conventional review	Comment	Link	Asked question	SWEBOK-Software engineering management
Conventional review	Comment	Code	Asked question	SWEBOK-Software engineering management
Management	Pinging	other	Process managment	SWEBOK-Software engineering management
Conventional review	Clarification	Code	Asked question	SWEBOK-Software engineering process
Conventional review	Comment	Code	Question	SWEBOK-Software engineering management
Management	Clarification	Code	Assisting the changes	SWEBOK-Software engineering management
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
Conventional review	Feedback	Code	Explaining changes	SWEBOK-Software engineering management
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
Conventional review	Comment	other	Explaining plan	SWEBOK-Software engineering management
Conventional review	Comment	other	Code review	SWEBOK-Software engineering management
Management	Code review	other	Assisting the changes	SWEBOK-Software engineering management
ML review	Code review	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code review	Code	Approved changes	SWEBOK-Software engineering process
ML review	Code review	Code	Approved changes	SWEBOK-Software engineering management
Conventional review	Request a review	Review	Code review	SWEBOK-Software engineering management



3.Visual Representation

CATEGORIZATION PROCESS

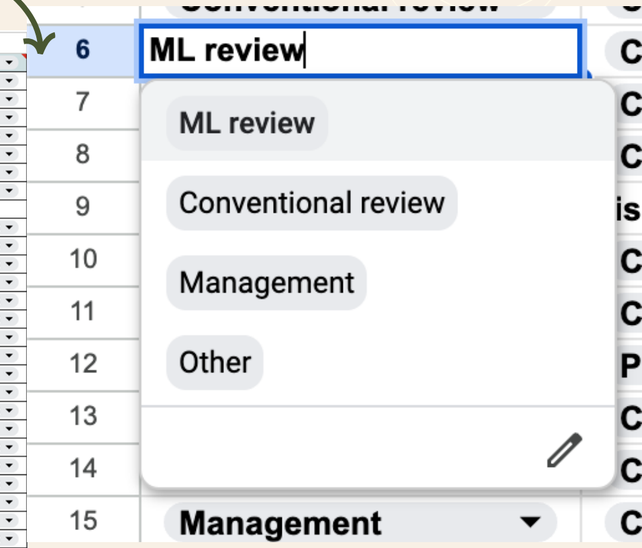
1.Data Collection



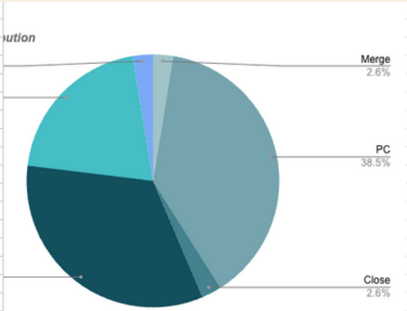
Date	Type	ID	Name	Body	URL	Comment type
2020-07-08T01:38:31Z	Commit	8a917da9fad79e74772c	WindQAQ	Migrate python implemen	https://github.com/tensorflow/tensorflow/commit/8a917da9fad79e74772c8f3bef4ff5f825b88d5c	ML review
2020-07-08T01:38:42Z	Commit	9d483364d19cd7740c6e	WindQAQ	Port to keras	https://github.com/tensorflow/tensorflow/commit/9d483364d19cd7740c6e93378236f1f9fefb994e	ML review
2020-07-08T01:41:10Z	DESCR	445836524	WindQAQ	As per https://github.com/	https://github.com/tensorflow/tensorflow/pull/41178	ML review
2020-07-08T01:48:15Z	PC	655231498	seanpmorgan	Thanks @WindQAQ!	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-655231498	Conventional revie
2020-07-08T01:55:33Z	Commit	75ad9db9f39709b5ac5dc	WindQAQ	run pylint	https://github.com/tensorflow/tensorflow/commit/75ad9db9f39709b5ac5dc80a93777df8c4362b55	ML review
2020-07-08T03:23:32Z	Commit	a3b21999360628de7fc1	WindQAQ	Update golden api	https://github.com/tensorflow/tensorflow/commit/a3b21999360628de7fc1e553eea4a96c44fd33be	ML review
2020-07-08T03:52:42Z	Commit	8adc98973ec85f27d9f36	WindQAQ	Update golden api	https://github.com/tensorflow/tensorflow/commit/8adc98973ec85f27d9f362abeb3020271f0a86164	ML review
2020-07-08T15:20:28Z	RC	444874482	alextp	DISMISSED	https://github.com/tensorflow/tensorflow/pull/41178#pullrequestreview-444874482	Dismissed
2020-07-13T20:12:49Z	RC	453904111	karmel	COMMENTED Can we a	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453904111	Conventional revie
2020-07-13T20:13:06Z	RC	453904243	karmel	COMMENTED Can we a	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453904243	Conventional revie
2020-07-13T20:14:23Z	RC	447566616	karmel	CHANGES_REQUESTE	https://github.com/tensorflow/tensorflow/pull/41178#pullrequestreview-447566616	Management
2020-07-13T20:16:46Z	RC	453906256	karmel	COMMENTED In general	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453906256	Conventional revie
2020-07-13T20:33:00Z	RC	453915068	karmel	COMMENTED Hi @karm	https://github.com/tensorflow/tensorflow/pull/41178#discussion_r453915068	Conventional revie
2020-07-13T20:36:48Z	PC	657780822	WindQAQ	Hi @karmel do we also r	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-657780822	Management
2020-07-14T02:39:05Z	Commit	2bcdcf38caeba1c836ed0	WindQAQ	Add example and more d	https://github.com/tensorflow/tensorflow/commit/2bcdcf38caeba1c836ed02d0da5e451c9e824275	ML review
2020-07-14T02:40:22Z	Commit	14ae0a4e8713fdd92552	WindQAQ	Fix wrong module	https://github.com/tensorflow/tensorflow/commit/14ae0a4e8713fdd92552d9e8ff4127092c4b7b81	ML review
2020-07-14T02:42:44Z	Commit	7d2b5cd83b8acd45a8c9	WindQAQ	Export only v2 api	https://github.com/tensorflow/tensorflow/commit/7d2b5cd83b8acd45a8c93309ac5d91164bce5c5d	ML review
2020-07-14T03:29:20Z	Commit	963717729afd197d9df9c	WindQAQ	Run pylint	https://github.com/tensorflow/tensorflow/commit/963717729afd197d9df9d6a3d88b3bd08f5e6ca2	ML review
2020-07-15T16:50:43Z	PC	658877893	hendrycks	When the C++ kernels an	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-658877893	Conventional revie
2020-07-15T17:56:10Z	Commit	8ad3b4d1c24ebe1d8805	WindQAQ	Merge branch 'master' of	https://github.com/tensorflow/tensorflow/commit/8ad3b4d1c24ebe1d8805cced6db7663a422869e4	ML review
2020-07-15T20:12:27Z	RC	449294836	karmel	DISMISSED for @tensorf	https://github.com/tensorflow/tensorflow/pull/41178#pullrequestreview-449294836	Conventional revie
2020-07-15T20:22:43Z	PC	658989172	WindQAQ	> for @tensorflow/api-ow	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-658989172	Conventional revie
2020-07-15T20:26:57Z	PC	658991281	hendrycks	> it will somehow loss scr	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-658991281	Management
2020-07-15T20:41:19Z	Commit	c504fee3471e2f4e2bf25	WindQAQ	Change approximate def	https://github.com/tensorflow/tensorflow/commit/c504fee3471e2f4e2bf257b4894db90b4395a5b5	ML review
2020-07-15T20:44:08Z	Commit	4e69c8bafe5f10745c827	WindQAQ	Merge branch 'master' of	https://github.com/tensorflow/tensorflow/commit/4e69c8bafe5f10745c827879cd712d22dfb23beb	ML review
2020-07-15T21:08:43Z	Commit	5a85ea8b4316984b9c65	WindQAQ	Update tests	https://github.com/tensorflow/tensorflow/commit/5a85ea8b4316984b9c6581d81faa002a34f5b8af	ML review
2020-07-15T22:22:39Z	PC	659045293	WindQAQ	Hi @karmel I change the	https://github.com/tensorflow/tensorflow/pull/41178#issuecomment-659045293	Conventional revie

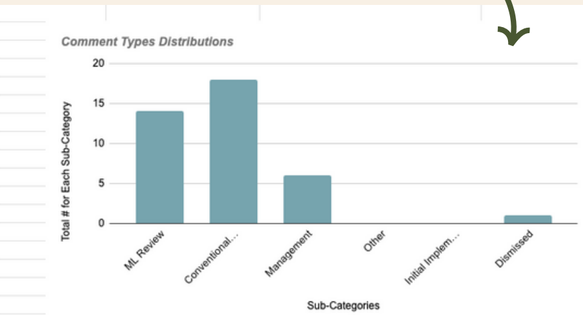
2.Categorization

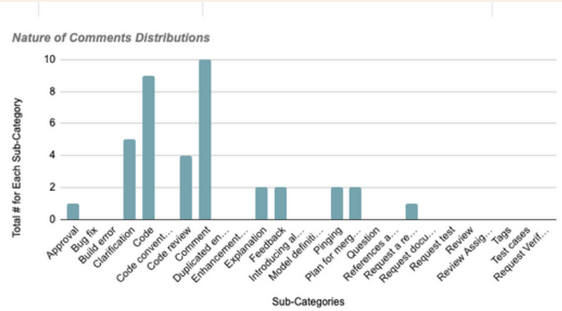
Comment type	Nature of comments	Contributed artifacts	Why	Relevent SDLC or ML phase
ML review	Clarification	Code	Code changes	SWEBOK-Software engineering process
ML review	Clarification	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	other	Assisting the changes	SWEBOK-Software engineering process
Conventional review	Comment	Review	Question	SWEBOK-Software engineering management
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
Dismissed	Dismissed	Dismissed	Dismissed	Dismissed
Conventional review	Comment	Link	Asked question	SWEBOK-Software engineering management
Conventional review	Comment	Code	Asked question	SWEBOK-Software engineering management
Management	Pinging	other	Process managment	SWEBOK-Software engineering management
Conventional review	Clarification	Code	Asked question	SWEBOK-Software engineering process
Conventional review	Comment	Code	Question	SWEBOK-Software engineering management
Management	Clarification	Code	Assisting the changes	SWEBOK-Software engineering management
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
Conventional review	Feedback	Code	Explaining changes	SWEBOK-Software engineering management
ML review	Code	Code	Assisting the changes	SWEBOK-Software engineering process
Conventional review	Comment	other	Explaining plan	SWEBOK-Software engineering management
Conventional review	Comment	other	Code review	SWEBOK-Software engineering management
Management	Code review	other	Assisting the changes	SWEBOK-Software engineering process
ML review	Code review	Code	Assisting the changes	SWEBOK-Software engineering process
ML review	Code review	Code	Approved changes	SWEBOK-Software engineering process
ML review	Code review	Code	Approved changes	SWEBOK-Software engineering management
Conventional review	Request a review	Review	Code review	SWEBOK-Software engineering management

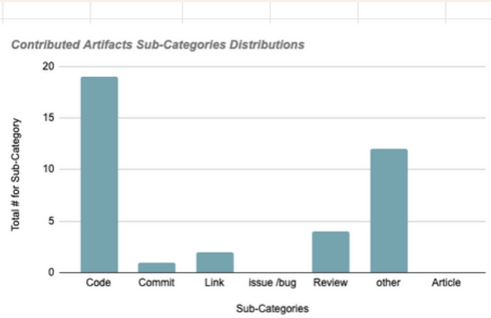


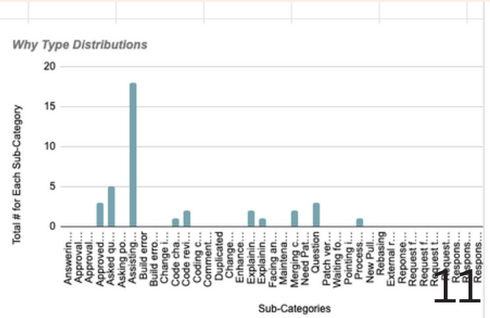
3.Visual Representation











Manual Classification

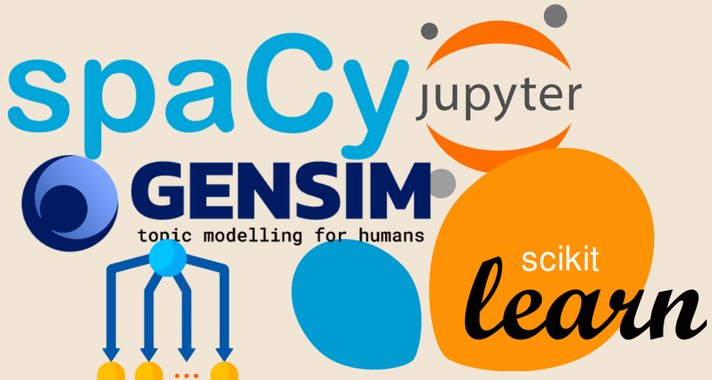
PROS	CONS
• Allowed Data Familiarization	• Time-Consuming & Costly
• Quality Control ○ Able to double check and compare with other results	• Oversight/ Inconsistencies
• Precision and Accuracy ○ Human judgement combined with collaboration	• Human Error: Typos, Bias

HOW CAN WE IMPROVE?

INTRODUCING

Machine learning

TO THE PROJECT...



GETTING READY TO TRAIN!

- Code Breakdown:
  a. **Import a .csv file containing a combined version of all the sheets**
  b. Give a Category Value for each sub-category in "Comment type" column
  c. Remove stop words and lemmatize the text from the "Body" column, store them in a vector list
  d. Give each vector a vectorized value (turning text into a numerical representation)
  e. Create training and test values

```
#Read in and print the csv file
df = pd.read_csv("combined_tf_pr.csv")
print(df.shape)
#print the 39 rows
df.head(607)
```

GETTING READY TO TRAIN!

- Code Breakdown:
  a. ~~Import a .csv file containing a combined version of all the sheets~~
  b. **Give a Category Value for each sub-category in "Comment type" column**
  c. Remove stop words and lemmatize the text from the "Body" column, store them in a vector list
  d. Give each vector a vectorized value (turning text into a numerical representation)
  e. Create training and test values

Why is this important?

Ensures efficiency in data processing!

```
#New Column: Gives a unique number to each of these categories
df['category_val'] = df['Comment type'].map({'Conventional review' : 0, 'ML review': 1, 'Dismissed' : 2, 'Management' : 3, 'Other' : 4,  'initial implementation' : 5, 'empty' : 6})
df.head(607)
```

GETTING READY TO TRAIN!

- Code Breakdown:
  a. ~~Import a .csv file containing a combined version of all the sheets~~
  b. ~~Give a Category Value for each sub-category in "Comment type" column~~
  c. **Remove stop words and lemmatize the text from the "Body" column, store them in a vector list**
  d. Give each vector a vectorized value (turning text into a numerical representation)
  e. Create training and test values

Why use nltk and spcay?

Ensures better coverage!

```
print(nltk_stopwords)
print("\n")
print(spacy_stopwords)
print("\n")
print(combined_stopwords)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yo

{'side', 'due', 'might', 'unless', 'hereby', 'us', 'a', 'towards', 'forty', 'thereafter', 'have', 'beforehand', 'this', 'his', 'no', 'seemed

{'side', 'hereupon', 'y', 'out', 'aren', 'does', 'due', 'before', "shan't", 'might', 'done', 'namely', 'unless', 'hereby', 'us', 'how', 'ele
```

GETTING READY TO TRAIN!

- Code Breakdown:
  a. ~~Import a .csv file containing a combined version of all the sheets~~
  b. ~~Give a Category Value for each sub-category in "Comment type" column~~
  c. ~~Remove stop words and lemmatize the text from the "Body" column, store them in a vector list~~
  d. **Give each vector a vectorized value (turning text into a numerical representation)**
  e. Create training and test values

Why is this important?

helps bridge the gap between human and machine.
eg.
woman -> queen, man -> king

```
#New Column: Gives. vectorized value of each input value
df['vector'] = df['Body'].apply(lambda text: preprocess_and_vectorize(text))
df.head(607)
```

```
[array([-2.4752  ,  0.52168 ,  0.16867 ,  3.2994  ,  2.1103  ,  0.25    ,
         0.046571,  2.7635  , -1.3151  ,  1.1516  ,  3.4943  ,  1.0532  ,
        -4.1382  ,  3.3488  ,  2.7279  ,  0.023003,  1.3536  ,  1.4741  ,
        -2.6494  , -2.9689  , -2.6738  , -2.7591  , -1.2577  ,  1.3407  ,
         0.58053 ,  0.32727 , -3.3395  , -0.35698 , -1.2736  ,  1.3963  ,
        -1.476   , -1.4323  , -1.1268  , -0.25453 , -1.9975  , -2.7572  ,
         1.3462  , -1.0832  ,  0.91696 , -3.0349  , -0.76496 , -1.9992  ,
         0.063289,  3.5607  , -2.7648  ,  0.76119 ,  2.2463  , -1.5844  ,
        -2.0389  , -1.5829  ,  1.0624  ,  2.8675  , -2.8235  , -4.5756  ,
        -2.7501  , -0.085979, -5.9145  ,  1.1754  ,  1.1318  , -0.38072 ,
        -3.2722  , -2.1684  , -0.4984  ,  0.88358 ,  5.0679  ,  2.0915  ,
        -2.7009  ,  0.22031 ,  2.2098  , -0.36808 ,  2.1906  , -3.6019  ,
         4.1686  ,  1.5225  , -0.43594 ,  0.42891 , -1.6773  , -0.28204 ,
        -0.59196 ,  1.0297  , -2.7454  , -0.89824 , -2.0003  ,  0.72807 ,
         1.0572  , -0.78476 ,  1.3047  , -4.2398  , -0.47198 ,  0.21958 ,
         2.6356  ,  0.53105 , -2.7675  ,  0.1523  , -0.71488 ,  2.7661  ,
         0.22556 , -0.48985 , -0.24394 ,  0.26594 ,  2.0607  ,  0.61695 ,
```

GETTING READY TO TRAIN!

- Code Breakdown:
  a. ~~Import a .csv file containing a combined version of all the sheets~~
  b. ~~Give a Category Value for each sub-category in "Comment type" column~~
  c. ~~Remove stop words and lemmatize the text from the "Body" column, store them in a vector list~~
  d. ~~Give each vector a vectorized value (turning text into a numerical representation)~~
  e. **Create training and test values**

Why is this important?

Allows the model to learn patterns from the training data and then its accuracy can be assessed on unseen data,

```
#Splitting the vectors of df into sets of X_train and testing with X_test. then splitting category_val into Y_train and testing with Y_test
X_train, X_test, Y_train, Y_test = train_test_split(
                                                    df['vector'],   # The data needing to split
                                                    df['category_val'],  # the category_values associated to each vector
                                                    test_size=0.2,  #testing 20% and training 80% of the data set
                                                    random_state =2023  #ensuring consistency in splitting
                                                    )
```

training (80%), testing (20%)

SUPERVISED LEARNING

CLASSIFICATION REPORT

Precision: The proportion of positive identifications that were actually correct. A model that produces no false positives has a precision of 1.0.

Recall: The proportion of actual positives that were identified correctly.

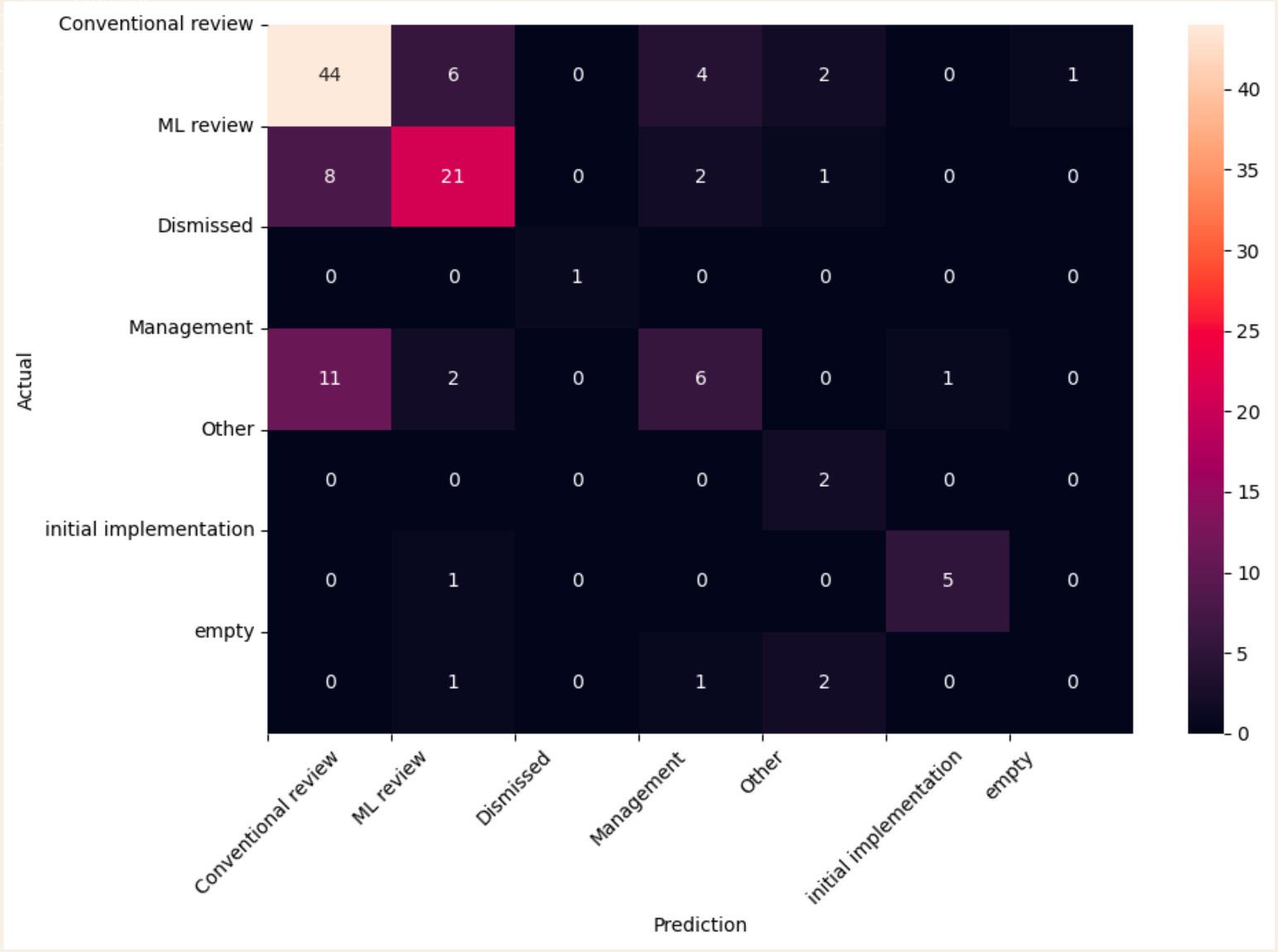
f1-score: Avg of precision and recall value. Best value at 1 and worst at 0.

Accuracy: Proportion of all predictions that were correct.

	labels	precision	recall	f1-score
Conventional review	0	0.70	0.77	0.73
ML review	1	0.68	0.66	0.67
Dismissed	2	1.00	1.00	1.00
Management	3	0.46	0.30	0.36
Other	4	0.29	1.00	0.44
initial implementation	5	0.83	0.83	0.83
empty	6	0.00	0.00	0.00
accuracy				0.65

The overall accuracy is 0.65 -> the model correctly predicted the class 65% of the time on the test set!

CONFUSION MATRIX



What does this tell us? Identifies the overall accuracy of the model AND also which specific classes the model is struggling with.

How to read?

- The diagonal from the top-left to the bottom-right represent the true positive counts for each class
- The rest are false predictions
- A higher TP values = higher accuracy.
- Actual:
- Prediction:

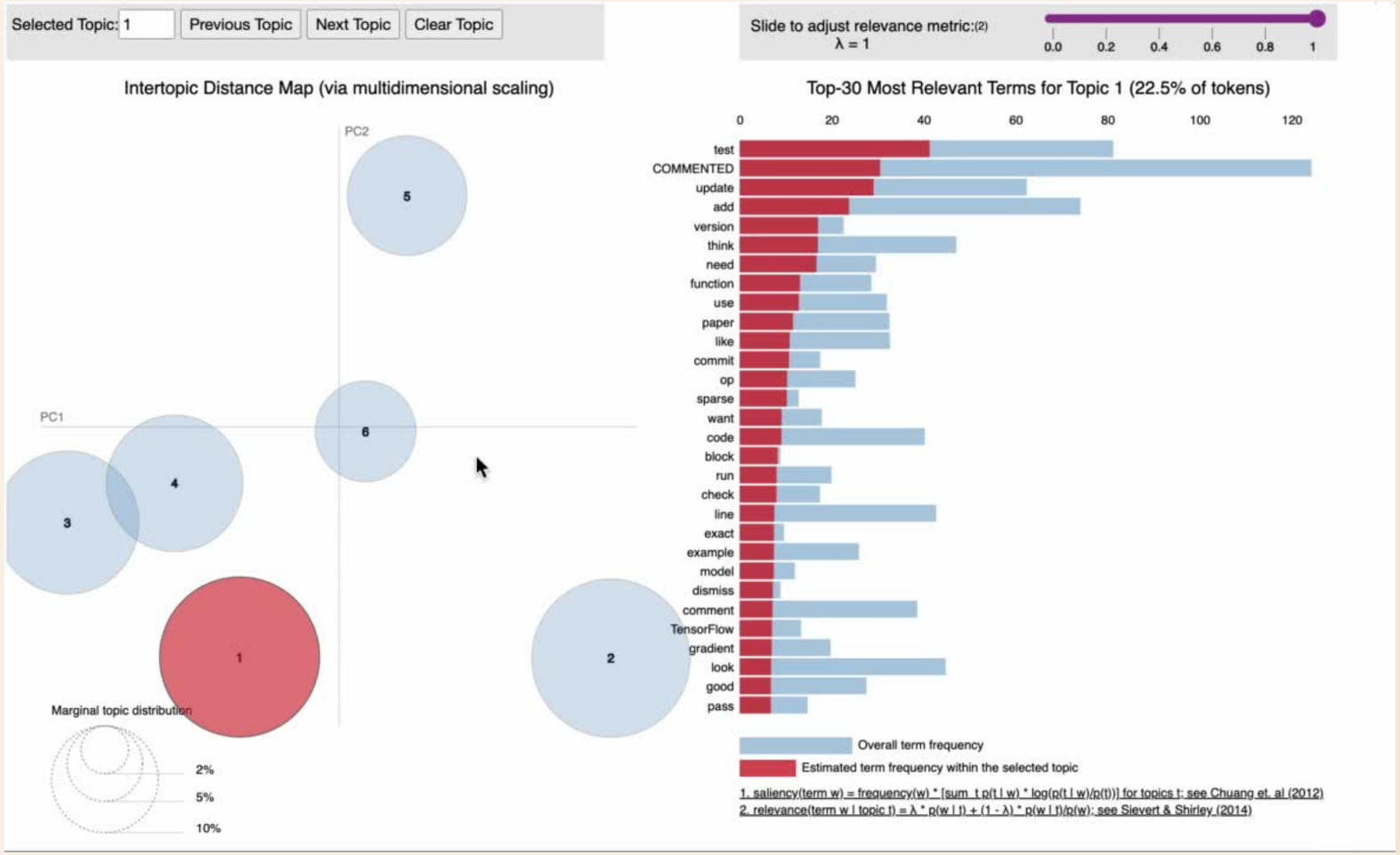
Supervised Learning

PROS	CONS
• High Accuracy ○ Assumption: The Given Data is Also Accurate	• Need to Provide Labeled Data: Time-Consuming & Costly Process
• Strong Predictive Power	• Oversight/ Inconsistencies: From Bias results That The Machine Can Pick Up On
• Data and Time Efficiency	• Data Processing: ○ Lots of Time is Needed to Review the Results and Ensure Accuracy

HOW CAN WE IMPROVE?

UNSUPERVISED LEARNING

PYLDAVIS GENSIM



What does this tell us?

- How many topics there are and their relevance.
- How topics relate to each other.
- The most important terms in the entire corpus and within each topic.

How to read?

- On the Left:
  - The larger the bubble the bigger the topic relevance
  - Distance between bubbles indicates the similarity between topics - closer together are more similar

- On the right:
  - This list shows the overall most descriptive/relevant terms in the corpus.

Unsupervised Learning

PROS	CONS
• No Need for Labeled Data	• No clear objective with the data
• Great for Data Exploration	• Sensitivity for Preprocessing
• Data and Time Efficiency	• Data Processing: ◦ Lots of Time is Needed to Review the Results and Ensure Accuracy

HOW CAN WE IMPROVE?

HOW DID WE IDENTIFY AND CLASSIFY MESSAGES CONTAINED IN A PULL REQUEST FOR MACHINE LEARNING INITIATIVES?

RESULTS

Able to identify what types of Pull Request comments are the most useful when it comes to the software development process.

e.g. If comment is classified as a Machine Learning message then a Machine Learning Engineer can comment back

FUTURE STEPS

Expand the code coverage to include the rest of the Categories such as "Nature of Comments", "Contributed Artifacts" etc..

BUMPS/OBSTACLES

Understanding and scheduling the amount of time necessary to complete each part of the project.

TIME MANAGEMENT

Reading and researching the best ways to approach the machine learning aspect of the project based on my current limited knowledge

ML KNOWLEDGE

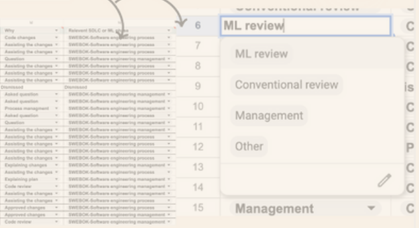
Understanding the various outputs and reflecting on what they mean based on the given data and how to fix the issues that are being represented incorrectly

DEBUGGING

CATEGORIZATION PROCESS

1.Data Collection

2.Categorization



16 Sheets, each varying from 7 to 85 rows

3.Visual Representation

IMPLEMENTATION PROCESS

Critical Components

- Code Breakdown:
  - Import a .csv file containing a combined version of all the sheets
  - Give a Category Value for each sub-category in "Comment type" column
  - Remove stop words and lemmatize the text from the "Body" column, store them in a vector list
  - Give each vector a vectorized value (turning text into a numerical representation)
  - Create training and test values

```
#New Column: Gives a unique number to each of these categories
df['category_val'] = df['Comment type'].map({'Conventional review' : 0, 'ML review': 1, 'Dismissed' : 2, 'Management' : 3, 'Other' : 4,  'initial implementation' : 5, 'empty' : 6})
df.head(607)
```

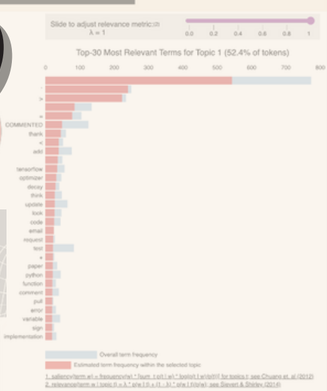
```
#New Column: Gives, vectorized value of each input value
df['vector'] = df['Body'].apply(lambda text: preprocess_and_vectorize(text))
df.head(607)
```

```
#Splitting the vectors of df into sets of X_train and testing with X_test. then splitting category_val into Y_train and testing with Y_test
X_train, X_test, Y_train, Y_test = train_test_split(
                                   df['vector'],   # The data needing to split
                                   df['category_val'],  # The category values associated to each vector
    test_size=0.2,  # 20% for testing, 80% for training 80% of the data set
    random_state =2023  # Ensures consistency in splitting
    )
```

```
#Converting the dataset into a Python list
X_train = X_train.tolist()
X_test = X_test.tolist()
```

VISUALIZATION PROCESS

pyLDAvis Gensim



- 6 different categories = 6 different "bubbles"
- **What does this tell us?**
  - How many topics there are and their relevance.
  - How topics relate to each other.
  - The most important terms in the entire corpus and within each topic.
  - How certain terms are distributed across topics.
- **How to read?**
  - On the Left:
    - Each bubble represents a topic (0 to 6)
    - The larger the bubble the bigger the topic relevance
    - Distance between bubbles indicates the similarity between topics - closer together are more similar.
  - On the right:
    - This list shows the overall most descriptive/relevant terms in the corpus.
    - Can hover over each word in the list and will highlight the most relevant bubbles and vice versa

CLASSIFICATION REPORT

The proportion of positive identifications that were actually correct. *A model that produces no false positives has a precision of 1.0.*

The proportion of actual positives that were identified correctly.

The mean of the precision and the recall value. Best value at 1 (perfect precision and recall) and worst at 0.

	precision	recall	f1-score	support
0	0.69	0.79	0.74	57
1	0.70	0.59	0.64	32
2	1.00	1.00	1.00	1
3	0.50	0.35	0.41	20
4	0.29	1.00	0.44	2
5	0.83	0.83	0.83	6
6	0.00	0.00	0.00	1
accuracy			0.65	122
macro avg	0.57	0.65	0.58	122
weighted avg	0.64	0.65	0.64	122

Proportion of all predictions that were correct.
The overall accuracy is 0.65, = the model correctly predicted the class 65% of the time on the test set.

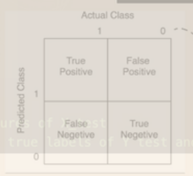
Independently for each class and then takes the average, treating all classes equally.

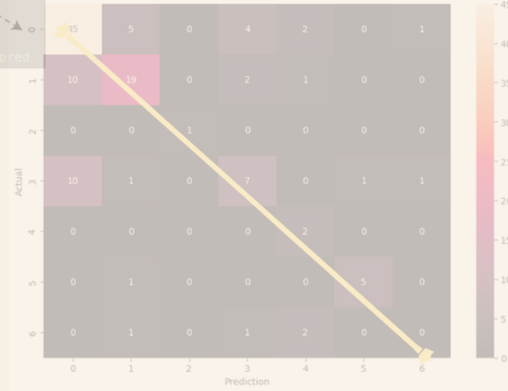
For each class and then takes a weighted average based on the number of true instances for each class

```
#Providing the values of the predicted
y_pred = clf.predict(X_test)  # predicting on the test features
print(classification_report(Y_test, Y_pred))  # comparing the true labels (Y_test) and Y_pred
```

VISUALIZATION PROCESS

Confusion Matrix





- 6 different categories = 6 x 6 matrix representation
- **What does this tell us?** Helps helps identify the overall accuracy of the model AND also which specific classes the model is struggling with.
- **How to read?**
  - The diagonal from the top-left to the bottom-right represents the true positive counts for each class
  - For example, at Row of Class5 and column of Class1 we see a value of 1 -> this means that **1** instance that is actually in class5 was incorrectly predicted as in class1
  - A higher TP values = higher accuracy .

QUESTIONS?

SO WHICH METHOD IS BETTER?

	PROS	CONS
Manual	• Data Familiarization • Quality Control • Precision and Accuracy	• Time-Consuming & Costly • Oversight/ Inconsistencies • Human Error: Typos

Supervised	• High Accuracy
	  ◦ Assumption: The Given Data is Also Accurate
	  ◦ Strong Predictive Power
	  ◦ Data and Time Efficiency
		• Time-Consuming & Costly Process
		• Oversight/ Inconsistencies:
		• Data Processing:

Unsupervised	• No Need for Labeled Data
	• Great for Data Exploration
	• Data and Time Efficiency
		• No clear objective with the data
		• Sensitivity:
		  ◦ Must Filter Through the Words to Avoid
		• Data Processing
		  ◦ Lots of Time is Needed to Review the Results and Ensure Accuracy