Hospital Playlist Subtitles EP01-03

This dataset contains data about subtitles from the "Hospital Playlist" Netflix TV Series and a complete sentiment analysis using VADER (Valence Aware Dictionary for Sentiment Reasoning) and CardiffNLP's roBERTa from HuggingFace's transformers.
The data has been originally obtained from opensubititles.org and processed for different purposes.

DATASET LINK	DATA CARD AUTHOR(S)
The dataset isn't yet published on any website.	**Stefano de Saraca:** (Owner, Manager)

Authorship

Publishers

PUBLISHING ORGANIZATION(S)	INDUSTRY TYPE(S)	CONTACT DETAIL(S)
LUMSA University - Rome	Academic - Tech	**Contact:** info@lumsa.it, lumsa@pec.it **Website:** www.lumsa.it

Dataset Owners

TEAM(S)	CONTACT DETAIL(S)	AUTHOR(S)
None	**Dataset Owner(s):** Stefano de Saraca **Affiliation:** LUMSA University – Rome **Contact:** s.desaraca@lumsastud.it	Stefano de Saraca, University Student, LUMSA University, 2024 www.opensubtitles.org – Liya Choi

Dataset Overview

DATA SUBJECT(S)

Main characters of the TV series:

1. Ahn Jeong-won
2. Kim Jun-Wan
3. Lee Ik-joon
4. Yang Seok-hyung
5. Chae Song-hwa

Other characters:

0. Secondary

DATASET SNAPSHOT

Size of Dataset	141 (EP1) KB 171 (EP2) KB 169 (EP3) KB
Number of Instances	1218 (EP1) 1471 (EP2) 1464 (EP3)
Number of Fields	13
Labeled Classes	6 (“Character” attribute)
Average Labels Per Instance	1
Number of Files	3

Visualization of the main characteristics of the dataset.

CONTENT DESCRIPTION

Since the dataset is divided in three files each of them has its own characteristics as described in the central column of this table.

The dataset has 13 fields:

- Start
- End
- Text
- Character
- SentenceIndex
- vaderNeg
- vaderNeu
- vaderPos
- vaderCompound
- roBERTaNeg
- roBERTaNeu
- roBERTaPos
- roBERTaMajorSentiment
- vaderMajorSentiment

The character field contains the name of the character that act the specific sentence of the “text” column in the episode itself.

DESCRIPTIVE STATISTICS

Episode 1:

Statistic	count	mean	std	min	25%	50%	75%	max	mode
Start	1218	2337.133	1409.569	7.968	1035.369	2362.799	3538.189	4811.628	7.968
End	1218	2339.241	1409.573	9.886	1039.435	2364.781	3540.806	4812.587	9.886
SentenceIndex	1218	608.500	351.750	0	304.250	608.500	912.750	1217	0
vaderNeg	1218	0.086	0.211	0	0	0	0	1	0
vaderNeu	1218	0.726	0.336	0	0.448	1	1	1	1
vaderPos	1218	0.169	0.285	0	0	0	0.336	1	0
vaderCompound	1218	0.065	0.274	-0.855	0	0	0.226	0.836	0
roBERTaNeg	1218	0.213	0.212	0.001	0.064	0.145	0.264	0.918	0.204

roBERTaNeu	1218	0.599	0.212	0.010	0.482	0.656	0.756	0.935	0.689
roBERTaPos	1218	0.186	0.214	0.004	0.053	0.105	0.202	0.988	0.105
roBERTaMajorSentiment	1218	//	//	//	//	//	//	//	Neutral
vaderMajorSentiment	1218	//	//	//	//	//	//	//	Neutral

Episode 2:

Statistic	count	mean	std	min	25%	50%	75%	max	mode
Start	1471	2260.624	1358.173	21.396	1077.993	2225.890	3364.652	4827.113	21.396
End	1471	2262.714	1358.284	22.896	1079.698	2229.100	3365.757	4829.283	22.896
SentenceIndex	1471	735	424.785	0	367.500	735	1102.500	1470	0
vaderNeg	1471	0.085	0.205	0	0	0	0	1	0
vaderNeu	1471	0.706	0.341	0	0.435	1	1	1	1
vaderPos	1471	0.186	0.299	0	0	0	0.385	1	0
vaderCompound	1471	0.067	0.282	-0.875	0	0	0.226	0.877	0
roBERTaNeg	1471	0.200	0.198	0.001	0.059		0.250	0.934	0.204
roBERTaNeu	1471	0.612	0.205	0.015	0.515	0.137	0.767	0.925	0.689
roBERTaPos	1471	0.186	0.212	0.004	0.054	0.664	0.217	0.983	0.105
roBERTaMajorSentiment	1471	//	//	//	//	//	//	//	Neutral
vaderMajorSentiment	1471	//	//	//	//	//	//	//	Neutral

Episode 3:

Statistic	count	mean	std	min	25%	50%	75%	max	mode
Start	1464	2491.539	1437.342	33.158	1198.613	2532.383	3631.429	5174.169	33.158
End	1464	2493.625	1437.305	34.658	1200.473	2534.318	3633.169	5177.459	34.658
SentenceIndex	1464	731.500	422.764	0	365.750	731.500	1097.250	1463	0

vaderNeg	1464	0.072	0.187	0	0	0	0	1	0.0
vaderNeu	1464	0.712	0.343	0	0.435	1	1	1	1.0
vaderPos	1464	0.195	0.304	0	0	0	0.411	1	0.0
vaderCompound	1464	0.080	0.264	-0.735	0	0	0.226	0.827	0.0
roBERTaNeg	1464	0.180	0.185	0.001	0.051	0.121	0.227	0.931	0.119
roBERTaNeu	1464	0.615	0.212	0.031	0.515	0.670	0.772	0.951	0.539
roBERTaPos	1464	0.203	0.229	0.006	0.059	0.110	0.241	0.966	0.340
roBERTaMajorSentiment	1464	//	//	//	//	//	//	//	Neutral
vaderMajorSentiment	1464	//	//	//	//	//	//	//	Neutral

Above: data overview by column.

Additional Notes: there are three tables, one for every episode analyzed.

Sensitivity of Data

SENSITIVITY TYPE(S)

Pseudonymous Data

FIELD(S) WITH SENSITIVE DATA

Intentionally Collected Sensitive Data

Field Name	Description
Character	The name of the character acted by a specific actor

Unintentionally Collected Sensitive Data

None

SECURITY AND PRIVACY HANDLING

No particular security precautions have been applied because every name of the characters and relative actors is publicly available on the Internet.

RISK TYPE(S)

No Known Risks

SUPPLEMENTAL LINK(S)

//

RISK(S) AND MITIGATION(S)

//

Dataset Version and Maintenance

MAINTENANCE STATUS

Limited Maintenance

The data will not be updated, but any technical issues will be addressed.

VERSION DETAILS

Current Version: 1.0

Last Updated: 25/04/2024

Release Date: 25/04/2024

MAINTENANCE PLAN

Any errors in the dataset will be corrected. Since it's only a subtitles dataset which is already been seen by many people it shouldn't have errors. Yet if one or more get discovered maintenance will be applied right away.

Versioning: every correction round will increment the version number.

Updates: No updates are planned in the close future.

Errors: No errors have been discovered yet.

Feedback: every feedback is very much appreciated and will be taken in consideration for improvements for the dataset.

NEXT PLANNED UPDATE(S)

EXPECTED CHANGE(S)

No planned updates.	//	//

Example of Data Points

<table>
<tr><th>PRIMARY DATA MODALITY</th><th>SAMPLING OF DATA POINTS</th><th>DATA FIELDS</th></tr>
<tr><td>Text Data
Float Data</td><td>No sampling (full data)</td><td>
<table>
<tr><th>Field Name</th><th>Field Value</th><th>Description</th></tr>
<tr><td>Character</td><td>“Chae Song-hwa”</td><td>Name of character</td></tr>
<tr><td>roBERTaPos</td><td>0.966</td><td>Sentiment score obtained from the analysis</td></tr>
<tr><td>roBERTaMajorSentiment</td><td>“Neutral”</td><td>Major sentiment of the sentence based on the scores obtained from the roBERTa model.</td></tr>
</table>
</td></tr>
</table>

Provenance

Collection

METHOD(S) USED	METHODOLOGY DETAIL(S)	SOURCE DESCRIPTION(S)
Retrieved from open-source website	**Source:** https://www.opensubtitles.com/en **Platform:** Open Subtitles **Is this source considered sensitive or high-risk?** No **Dates of Collection:** MAR 2024 – APR 2024 **Primary modality of collected data:** Text Data **Update Frequency for collected data:** Not defined	https://www.opensubtitles.org/en/subtitles/8233736/hospital-playlist-episode-1-1-en https://www.opensubtitles.org/en/subtitles/8233737/hospital-playlist-episode-1-2-en https://www.opensubtitles.org/en/subtitles/8233738/hospital-playlist-episode-1-3-en https://gotranscript.com/subtitle-converter

Motivations & Intentions

Motivations

PURPOSE(S)	DOMAIN(S) OF APPLICATION	MOTIVATING FACTOR(S)
Text Mining Descriptive Analysis Sentiment Analysis	Sentiment analysis Descriptive analysis	- *Analyzing the sentiment of each character by episode* - *Describing the change of sentiment of each character through episodes*

Intended Use

DATASET USE(S)	SUITABLE USE CASE(S)	UNSUITABLE USE CASE(S)
University exam project	**University Exam**: showing the skills needed to implement tools and theory learned during the lessons on a real-world project. **Personal Portfolio**: showing the ability to perform sentiment analysis on unstructured text using advanced tools and applying different methods to analyze sentiment. Additional Notes: every tool used is publicly available	**TV Series Marketing:** this dataset is not supposed to be used for any kind of marketing strategy or analysis.
	RESEARCH AND PROBLEM SPACE(S)	CITATION GUIDELINES

	Sentiment analysis on tv series	**Guidelines & Steps:** citation of the owner, version and changes applied for third parties is mandatory for any usage of this dataset.
		BiBTeX:
		`
		```
		{ `
		`  author={Stefano de Saraca},`
		`  platform={GitHub},`
		`  number={1},`
		`  rows={1218,1471,1464},`
		`  year={2024},`
		`  publisher={LUMSA University}`
		` }
		```
		 `

Access, Retention, & Wipeout

Access

ACCESS TYPE	DOCUMENTATION LINK(S)	PREREQUISITE(S)
External - Open Access	https://github.com/stefanodesaraca/Hospital-Playlist-Subtitles-Sentiment-Analysis/tree/main/HPDatasets	No prerequisites required
	POLICY LINK(S)	ACCESS CONTROL LIST(S)
	No access policies applied. The dataset is open-source.	No access control lists applied.

Retention

	DURATION	POLICY SUMMARY
	None	None
	PROCESS GUIDE	EXCEPTION(S) AND EXEMPTION(S)
	None	None

Wipeout and Deletion

	DURATION	DELETION EVENT SUMMARY
	None	Bad or illegal usages of the dataset.
	ACCEPTABLE MEANS OF DELETION	POST-DELETION OBLIGATIONS
	Deletion from any owned repository	No known obligations
	OPERATIONAL REQUIREMENT(S)	EXCEPTIONS AND EXEMPTIONS
	None	None

Provenance

Collection

METHOD(S) USED

Open-source website

METHODOLOGY DETAIL(S)

Source: https://www.opensubtitles.org/en/ssearch/sublanguageid-all/idmovie-920439

Platform: www.opensubtitles.org

Is this source considered sensitive or high-risk? No

Dates of Collection: 03/2024-04/2024

Primary modality of collected data:
Text Data

Update Frequency for collected data:
Static

SOURCE DESCRIPTION(S)

[www.opensubtitles.org]: an open-source website where everyone can download tv series or movies' subtitles for free.

COLLECTION CADENCE

Static

Data was collected once from a single source.

DATA INTEGRATION

www.opensubtitles.org
Included Fields
(Data fields that were collected and are included in the dataset.)

Field Name	Description
Start	The second which marks the start of the sentence.
End	The second which marks the end of the sentence.
Text	The text of the sentence.

Excluded Fields
No excluded fields collected.

DATA PROCESSING

Description: no processing executed during the collection of the dataset. All of the transformation will be described in the specific paragraph.

Collection Criteria

DATA SELECTION

DATA INCLUSION

DATA EXCLUSION

www.opensubtitles.org: No selection methods were applied.	www.opensubtitles.org: Only what was available on the source website was collected.	www.opensubtitles.org: No data was excluded during the collection process.

Relationship to Source

USE & UTILITY(IES)	BENEFIT AND VALUE(S)	LIMITATION(S) AND TRADE-OFF(S)
Open-source: every data used is freely available on their website and doesn't require any signing up or similar.	**Open-source:** the benefit of collecting the data from this source is that since it's possible for everyone to access useful data to execute their own analyses.	No tradeoffs found, other than not already having the name of the speaker for every sentence.

Version and Maintenance

	FIRST VERSION	NOTE(S) AND CAVEAT(S)
	Release date: 04/2024 **Link to dataset:** Hospital Playlist EP01-03 Subtitles, https://github.com/stefanodesaraca/Hospital-Playlist-Subtitles-Sentiment-Analysis/tree/main/HPDatasets **Status:** Static **Size of Dataset:** 141 (EP1) KB, 171 (EP2) KB, 169 (EP3) KB **Number of Instances:** 1218 (EP1), 1471 (EP2), 1464 (EP3)	Since the subtitles of the episodes don't change overtime, there's no need for maintenance, if not for technical problems.
CADENCE	**LAST AND NEXT UPDATE(S)**	**CHANGES ON UPDATE(S)**
Static	**Date of last update:** 10/04/2024 **Total Data points affected:** 4153 (Total from all episodes combined) **Data points updated:** 4153 **Data points added:** 0 **Data points removed:** 0 **Date of next update:** None	Unknown

Human and Other Sensitive Attributes

<table>
<tr><th>SENSITIVE HUMAN ATTRIBUTE(S)</th><th>INTENTIONALITY</th><th>RATIONALE</th></tr>
<tr><td>Full Name</td><td><b>Intentionally Collected Attributes</b>
<table>
<tr><th>Field Name</th><th>Description</th></tr>
<tr><td>Character</td><td>Contains name and surname of the character (not the actor’s name)</td></tr>
</table></td><td>Being the dataset mainly made of subtitles from a tv series (which of course includes people as the main subjects) it’s very difficult to hide the full name of the character, but since they’re referred to the character itself and not to the actor data sensibility shouldn’t a big issue.</td></tr>
<tr><td></td><td>SOURCE(S)</td><td>METHODOLOGY DETAIL(S)</td></tr>
<tr><td></td><td>www.opensubtitles.org: every data has been collected from this open-source website.</td><td>No particular methods, practices or other additional processes have been applied.</td></tr>
</table>

DISTRIBUTION(S)

Episode 1

Character (Full Name)

	Character
Count (Sentences)	Ahn Jeong-won: 204 Kim Jun-wan: 117 Lee Ik-joon: 23 Yang Seok-hyung: 51 Chae Song-hwa: 169 Secondary: 654
Mode	Secondary
Min (Minimum Number of Sentences Acted)	Lee Ik-joon
Max (Maximum) Number of Sentences Acted)	Secondary (Secondary Characters) Ahn Jeong-won (5 Main Characters)
Range (Possible Values)	Ahn Jeong-won Kim Jun-wan Lee Ik-joon Yang Seok-hyung Chae Song-hwa Secondary

Episode 2

Character (Full Name)

	Character
Count (Sentences)	Ahn Jeong-won: 90 Kim Jun-wan: 90 Lee Ik-joon: 133 Yang Seok-hyung: 34 Chae Song-hwa: 228 Secondary: 896
Mode	Secondary
Min (Minimum Number of Sentences Acted)	Yang Seok-hyung

Max (Maximum) Number of Sentences Acted)	Secondary (Secondary Characters) Chae Song-hwa (5 Main Characters)
Range (Possible Values)	Ahn Jeong-won Kim Jun-wan Lee Ik-joon Yang Seok-hyung Chae Song-hwa Secondary

Episode 3

Character (Full Name)

	Character
Count (Sentences)	Ahn Jeong-won: 65 Kim Jun-wan: 240 Lee Ik-joon: 239 Yang Seok-hyung: 15 Chae Song-hwa: 95 Secondary: 810
Mode	Secondary
Min (Minimum Number of Sentences Acted)	Yang Seok-hyung
Max (Maximum) Number of Sentences Acted)	Secondary (Secondary Characters) Kim Jun-wan (5 Main Characters)
Range (Possible Values)	Ahn Jeong-won Kim Jun-wan Lee Ik-joon Yang Seok-hyung Chae Song-hwa Secondary

Above:
Basic descriptive analysis for each episode of the “Character” human attribute (since it’s the only human attribute present in the dataset).

KNOWN CORRELATIONS

RISK(S) AND MITIGATION(S)

	No known correlations.	//

Extended Use

Use with Other Data

SAFETY LEVEL	KNOWN SAFE DATASET(S) OR DATA TYPE(S)	BEST PRACTICES
Safe to use with other data	No known datasets or data types which this one could be joined or aggregated with.	Always mention that this dataset is made of sentences which are told by invented characters and not real people.
	KNOWN UNSAFE DATASET(S) OR DATA TYPE(S)	LIMITATION(S) AND RECOMMENDATION(S)
	//	*Same as mentioned in “Best practices”.*

Forking & Sampling

SAFETY LEVEL	ACCEPTABLE SAMPLING METHOD(S)	BEST PRACTICE(S)
Safe to fork and/or sample	Random Sampling Weighted Sampling Unknown	**Forking**: if forked always mention the original dataset with a link to the original repository where it’s located. **Sampling**: if sampled always mention the original data shape and the sampled one including the sampling method.
	RISK(S) AND MITIGATION(S)	LIMITATION(S) AND RECOMMENDATION(S)
	None	**Limitations:** Small size of the dataset. **Recommendations:** Always double check if the original version has been forked and not an already forked one.

Use in ML or AI Systems

DATASET USE(S)	NOTABLE FEATURE(S)	USAGE GUIDELINE(S)
Sentiment Analysis Using Neural Networks	**Text:** the text analyzed by the neural network (roBERTa).	**Usage Guidelines:** be sure that the text hasn’t been edited if the dataset has been modified by third parties after a fork operation, otherwise the results could be different.
	DISTRIBUTION(S)	KNOWN CORRELATION(S)

	//	Unknown

SPLIT STATISTICS

None

Transformations

Synopsis

TRANSFORMATION(S) APPLIED	FIELD(S) TRANSFORMED	LIBRARY(IES) AND METHOD(S) USED

Cleaning Missing Values

Converting Data Types

Joining Input Sources

Converting Classes Names

Removing Stopwords

Converting Characters to ASCII

Replacing Useless Characters

Lowering Every Character

Joining Input Sources, Converting Classes Names

Field Name	Source & Target
Character	Character -> Character

Removing Stopwords, Converting “-“ to “ “, Cleaning Emojis, Converting to ASCII

Field Name	Source & Target
Text	Text -> Text

Method: multiple transformations have been carried on, including:

- **Joining Input Sources**: adding the “Character” column through an Excel spreadsheet.
- **Converting Missing Values**: the only missing values that were present in the dataset were located in the “Characters” column where they have been replaced by the “Secondary” class. This means that the sentences on those rows were told by secondary characters.
- **Converting Data Types**: converting the character numbers to their real names.
- **Removing Stopwords:** using the “re” (RegEx) library and tokenizing every sentence it has been possible to remove every stopwords.
- **Converting Classes Names**: since the characters were originally identified by a number during the data cleaning process every number has been replaced by the corresponding name of the character.
- **Converting Characters to ASCII:** by using the clean-text library it has been possible to convert every character to its closest ASCII one.
- **Replacing Useless Characters**: the useless dashes that were present in some sentences were replaced by simple spaces.
- **Lowering Every Character**: every character of every sentence has been lowered through the lower() Python base function.

Platforms, tools, or libraries:
Excel Spreadsheets
Pandas
Clean-Text
re (Regular Expressions)

Transformation Results: <Provide results, outcomes, and actions taken because of the transformations. Include visualizations where available.>

Additional Notes: all libraries used have been applied in Python scripts only.

Breakdown of Transformations

<table>
<tr><th>CLEANING MISSING VALUE(S)</th><th>METHOD(S) USED</th><th>COMPARATIVE SUMMARY</th></tr>
<tr><td><b>Characters:</b> every character had a number and if the sentence was told by a secondary one the column field was empty.
During the transformation process every NaN value (empty field) was replaced by a “Secondary” string.</td><td><b>Pandas:</b> by using the Pandas fillna() method every NaN has been replaced by a 0 and then by the “Secondary” string.</td><td>&lt;Summarize here. Include links, tables, visualizations where available>
<table><tr><th>Field Name</th><th>Diff</th></tr><tr><td>Character</td><td>NaN -> 0</td></tr><tr><td>Character</td><td>0 -> “Secondary”</td></tr></table></td></tr>
<tr><th>RESIDUAL & OTHER RISK(S)</th><th>HUMAN OVERSIGHT MEASURE(S)</th><th>ADDITIONAL CONSIDERATIONS</th></tr>
<tr><td>None</td><td>None</td><td>None</td></tr>
<tr><th>CONVERTING DATA TYPE(S)</th><th>METHOD(S) USED</th><th>COMPARATIVE SUMMARY</th></tr>
<tr><td>All fields of the “Characters” column have been affected by the data types conversion.</td><td>First of all, every value of the “Character” column has been converted into an integer value represented by a 16bit word. After that they have been converted to the string type.</td><td><table><tr><th>Field Name</th><th>Diff</th></tr><tr><td>Character</td><td>1 -> “1”</td></tr></table></td></tr>
<tr><th>RESIDUAL & OTHER RISK(S)</th><th>HUMAN OVERSIGHT MEASURE(S)</th><th>ADDITIONAL CONSIDERATIONS</th></tr>
<tr><td>None</td><td>None</td><td>None</td></tr>
<tr><th>JOINING INPUT SOURCES</th><th>METHOD(S) USED</th><th>COMPARATIVE SUMMARY</th></tr>
<tr><td><ul><li>Original dataset obtained from www.opensubtitles.org</li><li>Additional column containing the character name for each row.</li><li>Sentiment analysis scores</li></ul></td><td>Although it can’t really be identified as a join the original dataset has been “joined” with the character number which represent the speaker of that specific sentence.

At the end of the sentiment analysis multiple columns containing the scores for each method used have been joined with the original dataset through an ”inner join”.</td><td>None</td></tr>
</table>

<table>
<tr><td>RESIDUAL & OTHER RISK(S)</td><td>HUMAN OVERSIGHT MEASURE(S)</td><td>ADDITIONAL CONSIDERATIONS</td></tr>
<tr><td>None</td><td>None</td><td>None</td></tr>
<tr><td>CLASSES NAMES CONVERSION</td><td>METHOD(S) USED</td><td>COMPARATIVE SUMMARY</td></tr>
<tr><td><b>Character:</b> every character has a number assigned and every row has a “Character” field, which will be converted based on a dictionary to the corresponding character.</td><td>First of all, a dictionary that contains numbers and characters names is created.
After that through a lambda function inside the pandas apply() one every character number is converted to the one’s name.</td><td>&lt;Summarize here. Include links, tables, visualizations where available.&gt;<table><tr><th>Field Name</th><th>Diff</th></tr><tr><td>Character</td><td>5 -&gt; “Chae Song-hwa”</td></tr></table></td></tr>
<tr><td>RESIDUAL & OTHER RISK(S)</td><td>HUMAN OVERSIGHT MEASURE(S)</td><td>ADDITIONAL CONSIDERATIONS</td></tr>
<tr><td>None</td><td>None</td><td>None</td></tr>
<tr><td>REMOVING STOPWORDS</td><td>METHOD(S) USED</td><td>COMPARATIVE SUMMARY</td></tr>
<tr><td>This transformation effected only the “Text” column in which sentences are stored.
Each sentence is then tokenized and with a list comprehension and RegEx it has been possible to remove all stopwords.</td><td><b>re, NLTK:</b> using the re (Regular Expressions) and NLTK’s word_tokenize() function it has been possible to remove all stopwords that were previously loaded from a text file with a list of words that are useless to the analysis (stopwords).</td><td><i>Example: assuming we have these stopwords: [my, is]</i><table><tr><th>Field Name</th><th>Diff</th></tr><tr><td>Text</td><td>Hello, my name is Stefano -> Hello, name Stefano</td></tr></table><b>Above:</b> Simple example to understand stopwords removal.</td></tr>
<tr><td>RESIDUAL & OTHER RISK(S)</td><td>HUMAN OVERSIGHT MEASURE(S)</td><td>ADDITIONAL CONSIDERATIONS</td></tr>
<tr><td>None</td><td>None</td><td>Potentially some stopwords could be excluded from the removal because of them missing in the stopwords text file, but after a quick check it’s possible to just add them afterwords.</td></tr>
<tr><td>CONVERTING CHARACTERS TO ASCII, LOWERING EVERY CHARACTER</td><td>METHOD(S) USED</td><td>COMPARATIVE SUMMARY</td></tr>
<tr><td>Every character of every sentence in the “Text” column has been lowered and converted to its closest ASCII one.</td><td><b>Base Python, clean-text:</b> using the base Python function lower() and the clean() one from the clean-text library it was possible to convert every character that wasn’t ASCII to one so to not have problems during the sentiment analysis.</td><td><i>Example:</i><table><tr><th>Field Name</th><th>Diff</th></tr><tr><td>Text</td><td>Å -&gt; a</td></tr></table><b>Above:</b> note that this is just an example</td></tr>
<tr><td>RESIDUAL & OTHER RISK(S)</td><td>HUMAN OVERSIGHT MEASURE(S)</td><td>ADDITIONAL CONSIDERATIONS</td></tr>
<tr><td>Some characters could be wrongly converted because of their non-ASCII nature, so they could just be</td><td>None</td><td>None</td></tr>
</table>

<table>
<tr><td>not recognized by the sentiment analysis algorithm.</td><td></td><td></td></tr>
<tr><td>REPLACING USELESS CHARACTERS</td><td>METHOD(S) USED</td><td>COMPARATIVE SUMMARY</td></tr>
<tr><td>Every dash (-) has been replaced by a space since some sentences contained some of them.
The only column affected by this transformation was the “Text” one.</td><td><b>Base Python</b>: by using the base python function replace() it has been possible to transform dashes (-) into simple spaces.</td><td><i>Example:</i><table><tr><th>Field Name</th><th>Diff</th></tr><tr><td>Text</td><td>This-is-a-dash ->
This is a dash</td></tr></table><b>Above:</b> simple example of the replace() function applied with the dash removal transformation</td></tr>
<tr><td>RESIDUAL & OTHER RISK(S)</td><td>HUMAN OVERSIGHT MEASURE(S)</td><td>ADDITIONAL CONSIDERATIONS</td></tr>
<tr><td>None</td><td>None</td><td>None</td></tr>
</table>

Annotations & Labeling

ANNOTATION WORKFORCE TYPE	ANNOTATION CHARACTERISTIC(S)	ANNOTATION DESCRIPTION(S)
Unlabeled	None	None
	ANNOTATION DISTRIBUTION(S)	ANNOTATION TASK(S)
	None	None

Human Annotators

	ANNOTATOR DESCRIPTION(S)	ANNOTATOR TASK(S)
	None	None
LANGUAGE(S)	LOCATION(S)	GENDER(S)
None	None	None

Validation Types

METHOD(S)	BREAKDOWN(S)	DESCRIPTION(S)
Data Type Validation Classes Validation Values Range Validation	**Data Type Validation:** **Number of Data Points Validated:** 4.153 **Fields Validated:** - Character **Classes Validation:** **Fields Validated:** - Character **Values Range Validation:** **Number of Data Points Validated:** 4.153 **Fields Validated:** - Character - vaderPos - vaderNeu - vaderNeg - vaderCompound - roBERTaPos - roBERTaNeu - roBERTaNeg	**Data Type Validation:** All of the characters names are strings and that has been confirmed thanks to the “raise” condition that makes sure that if during the transformation process an error occurred the code will stop running communicating something is wrong. **Classes Validation:** Since every character had its own number identifier and the fillna()function from the Pandas library has been used, as explained before, every row will have a value that’s in the dictionary below: "1": "Ahn Jeong-won", "2": "Kim Jun-Wan", "3": "Lee Ik-joon", "4": "Yang Seok-hyung", "5": "Chae Song-hwa", "0": "Secondary” **Values Range Validation:** The “Character” column values (and so the range too) have been validated as explained in the previous “Classes Validation” section. The other columns had their ranges validated from the describe() function from the Pandas library which summarizes the entire data of the columns and includes the min and max for all of them.

Description of Human Validators

	CHARACTERISTIC(S)	DESCRIPTION(S)
	//	//

LANGUAGE(S)	LOCATION(S)	GENDER(S)
//	//	//

Terms of Art

Concepts and Definitions referenced in this Data Card

Pandas	Dictionary	re (Regula Expressions)
Definition: popular Python library for data science, data wrangling and more. Source: https://pandas.pydata.org/docs/index.html	Definition: basic data structure made of key-value pairs.	Definition: regular expressions (RegEx) are strings of characters which can identify a recurrent pattern in a text. Source: https://docs.python.org/3/library/re.html
ASCII	NaN	Join
Definition: ASCII (American Standard Code for Information Interchange) is a characters codification code which contains 2^8 characters (256 [0, 255]). Source: https://en.wikipedia.org/wiki/ASCII	Definition: abbreviation of “Not a Number”. Source: https://en.wikipedia.org/wiki/NaN	Definition: merging operation between two or more objects.
Sentiment Analysis	VADER (Valence Aware Dictionary and sEntiment Reasoner)	roBERTa
Definition: it’s a sub-field of the bigger “Text Mining” which describes theory, techniques and algorithms behind the sentiment (opinion, polarity, etc.) of a given text.	Definition: VADER is a rule-based sentiment analysis tool which is based on the “Bag of Words” approach. Source: https://medium.com/@rslavanyageetha/vader-a-comprehensive-guide-to-sentiment-analysis-in-python-c4f1868b0d2e	Definition: a neural network model developed by CardiffNLP, based on Google’s BERT model and trained on twitter’s tweets which is capable of analyzing the sentiment of a given text. Source: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment