Domain	Source	Type	License
Sub-Domain			
CAPTIONS			
Images	(ElJundi et al., 2020)	Public Dataset	Public Domain
	Flikr30K (Plummer et al., 2015)	Public Dataset	CC0
	WikiCaps (Schamoni et al., 2018)	Public Dataset	CC BY 4.0
	(Rathi, 2020)	Public Dataset	Public Domain
Videos	Vatex (Wang et al., 2019)	Public Dataset	CC BY 4.0
	MultiCapCLIP (Yang et al., 2023)	Public Dataset	BSD-3-Clause license
	(Singh et al., 2022)	Public Dataset	Public Domain
Movies	OpenSubtitles2016 (Lison and Tiedemann, 2016)	Public Dataset	Public Domain
YouTube	youtube.com	Captions	CC
MEDICAL TEXT			
Clinical Reports	i2b2/VA (Uzuner et al., 2011)	On Request Dataset	—
DICTIONARIES			
	almaany.com	Web Article	CC
	dictionary.com	Web Article	CC
ENTERTAINMENT			
	(Al-Khalifa et al., 2022)	Public Dataset	Public Domain
	(Weller and Seppi, 2019)	Public Dataset	MIT License
Jokes	(Jokes)	Public Dataset	Public Domain
	123hindijokes.com	Web List	Public Domain
SPEECH			
Ted Talks	ted.com/talks	Video Transcription	CC BY-NC-ND 4.0
	state.gov/translations/arabic	Web Article	Public Domain
Public Speech	ruscorpora.ru	Document	Personal/Non-Commercial
	whitehouse.gov	Web Article	CC BY 3.0 US
STATEMENTS			
Rumours	Stanceosaurus (Zheng et al., 2022)	Public Dataset	Public Domain
	arabic-quotes.com	Web List	Public Domain
	goodreads.com/quotes	Web List	Public Domain
Quotes	evene.lefigaro.fr	Web List	Personal/Non-Commercial
	storyshala.in	Web List	Public Domain
	infoselection.ru	Web List	Personal/Non-Commercial
	aldiwan.net	Web List	Public Domain
	poetryfoundation.org	Web List	Public Domain
POETRY	poesie-francaise.fr	Web List	Public Domain
	hindionlinejankari.com	Web List	Public Domain
	ruscorpora.ru	Document	Personal/Non-Commercial
LETTERS	oflosttime.com	Web Article	Public Domain
	gutenberg.org	Document	Public Domain
	runivers.ru	Document	Personal/Non-Commercial

Table 20: License or term of use per source (3/3)