Mentorship Program for ICML 2021

Building NLP Models for Detecting Drought Impacts

May 6, 2021

Michael Hayes, Ph.D.
Professor
Climatologist

Tsegaye Tadesse, Ph.D.
Research Professor
Applied Climatologist
Remote Sensing Expert

Kelly Smith, Ph.D.
Communications Coordinator &
Assistant Director of NDMC
Drought Impact Expert

Beichen Zhang
Ph.D. Student in Natural Resource Sciences
Climate Assessment and Impacts

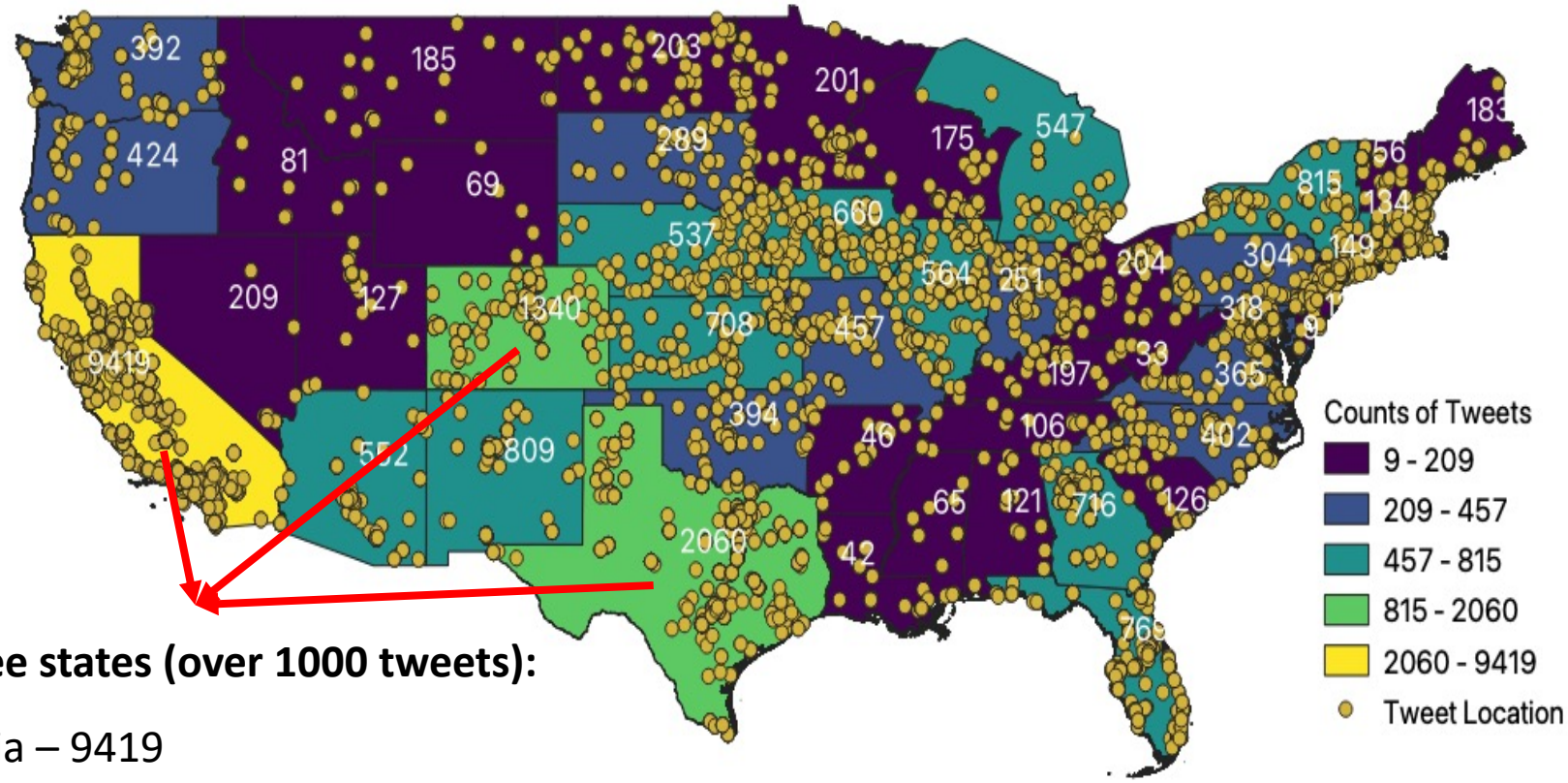
B.S. Major in **Geographic Information Science (GIS, RS, GPS)**
M.S. Major in **Natural Resource Sciences**
with a specialization in **Climate Assessment and Impacts**
Skill sets:

- Remote sensing (GEE, ERDAS, ENVI)
- GIS (QGIS, ArcGIS, GDAL, etc.)
- Statistical analysis (regression, spatial, time series, etc.)
- Machine Learning (ML) & Deep Learning (DL)

Topic of Ph.D. dissertation: apply **AI** techniques to drought monitoring and **impacts assessment**

Data Description – Tweets with Locations

- **26654** records
- January 9, **2017** – October 8, **2020**
- #drought, #droughtYYYY, #droughtYY, #droughtState (ca,tx)
- Excluded outliers such as bots and spammers, and professional users such as @DroughtCenter.



Top three states (over 1000 tweets):

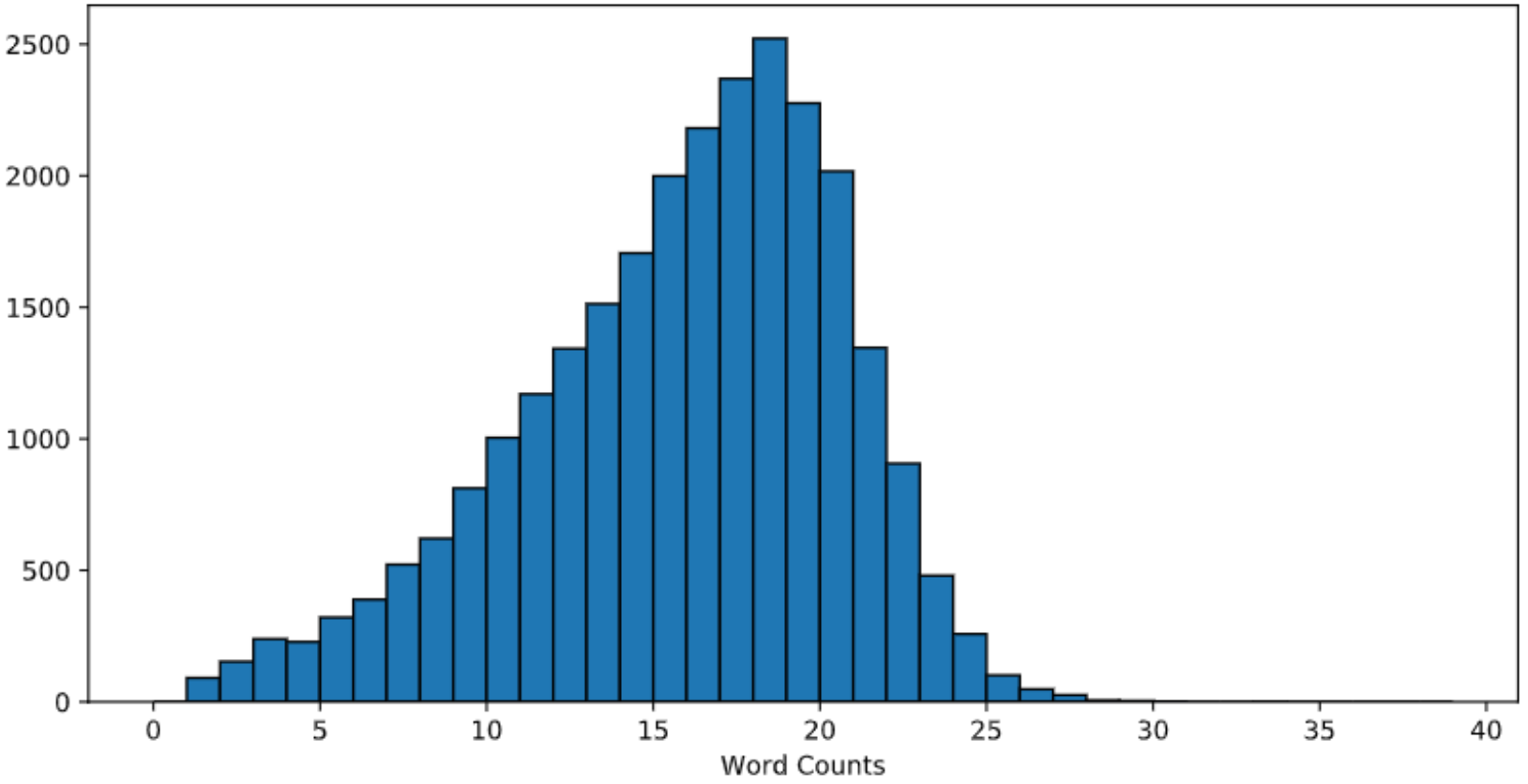
California – 9419

Texas – 2060

Colorado - 1340

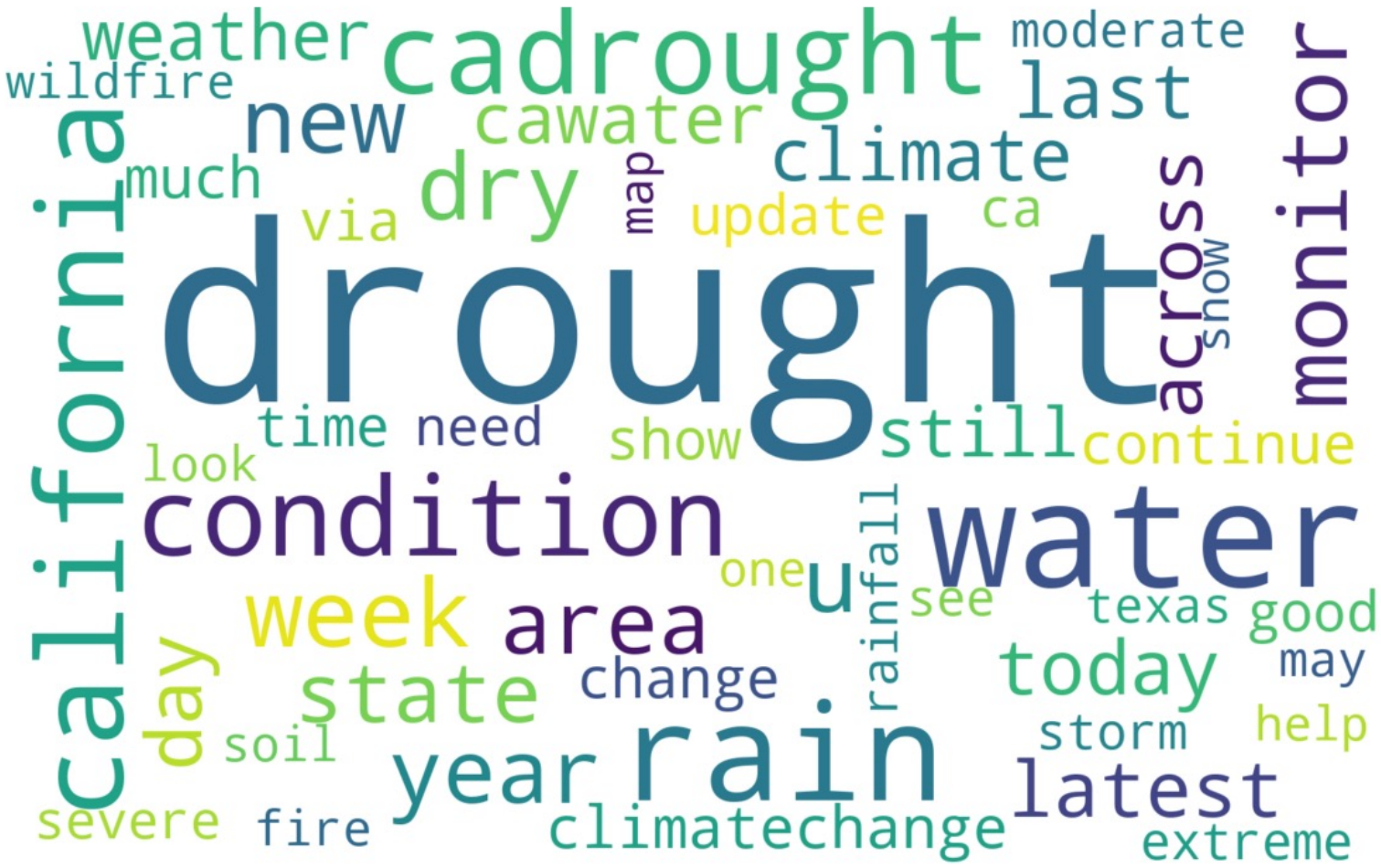
Data Description – Tweets with Locations

Word counts per Tweet (removed html tags and urls)

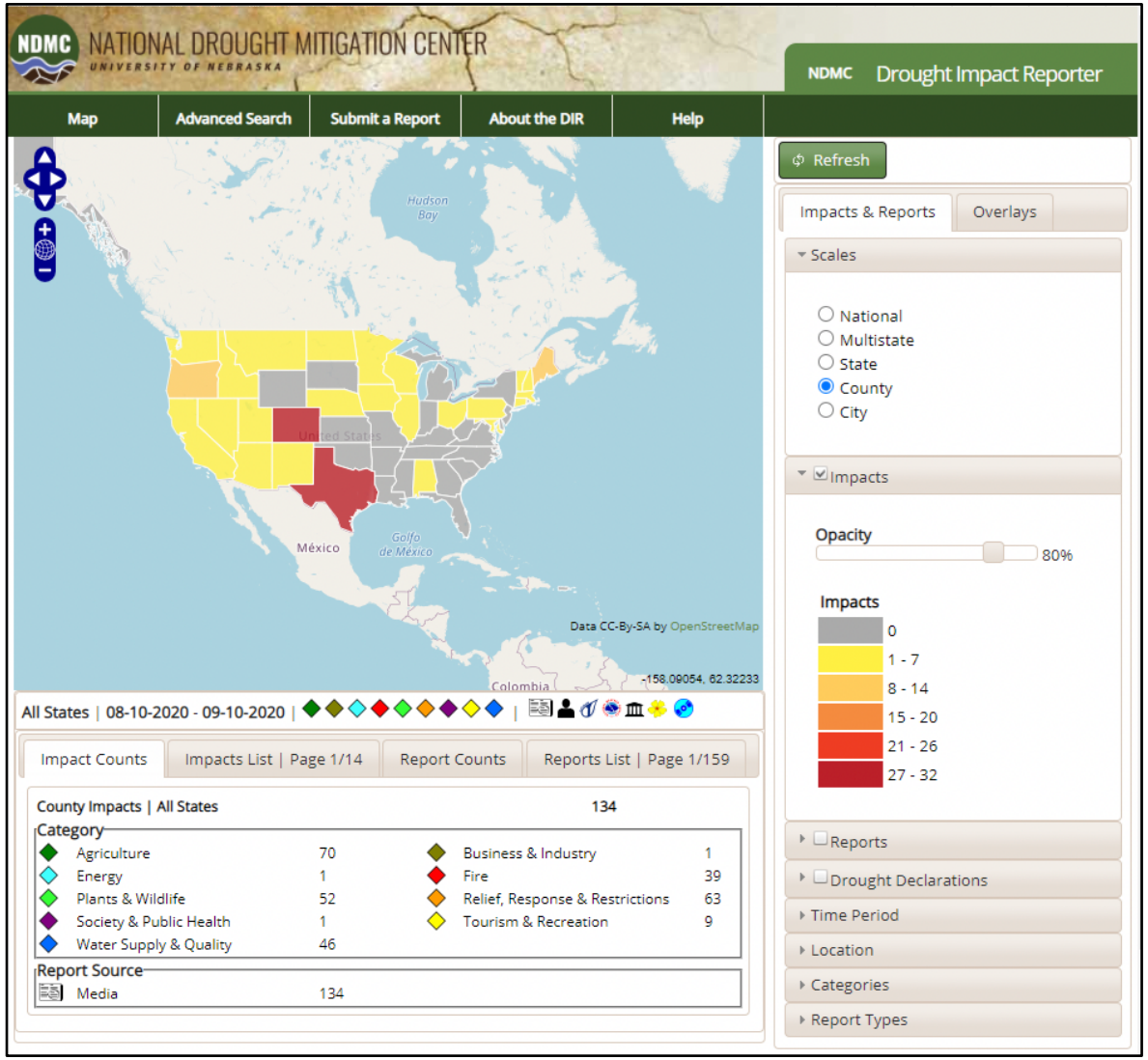


Data Description – Tweets with Locations

Word Cloud
(removed html tags, urls, special characters, stop words)



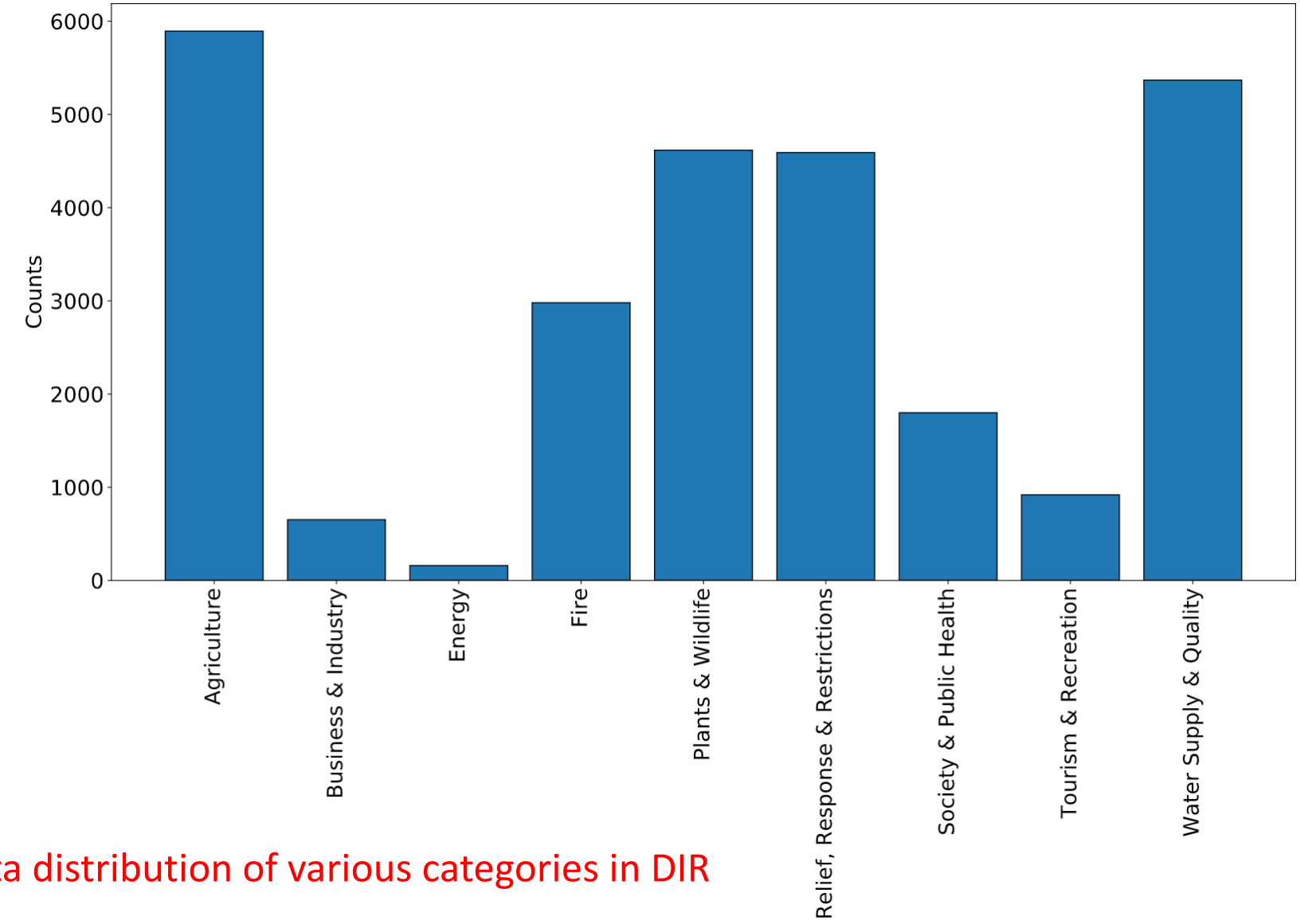
Data Description – Drought Impact Reporter (DIR)



https://droughtreporter.unl.edu/map/

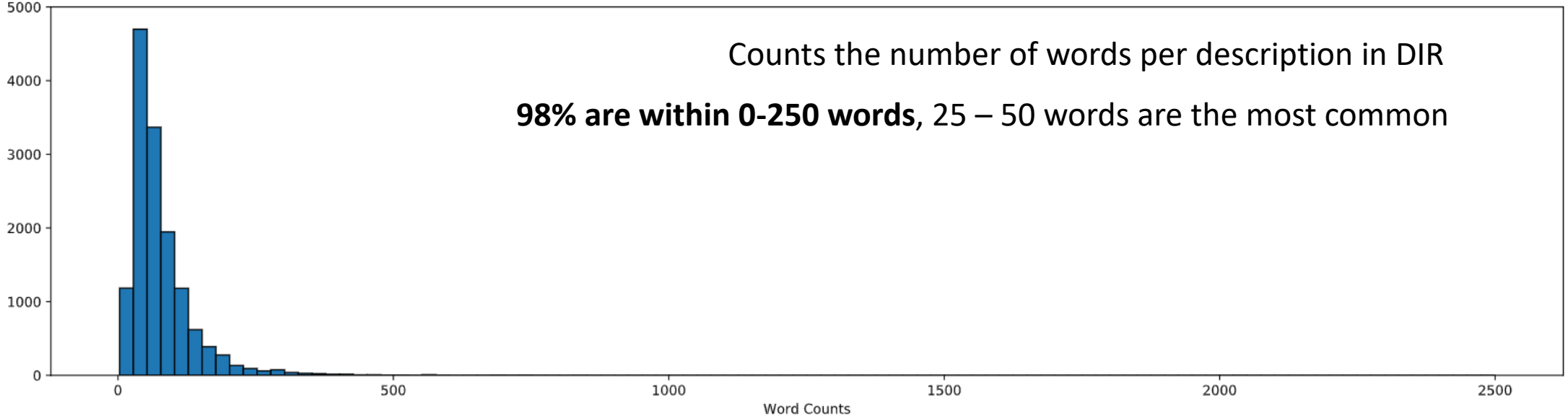
- The first national comprehensive drought database.
- Multiple reporting resources.
- Nine categories of drought impacts.
- Flexible spatial resolution and available historical records (2005 – current, updated in 2011).
- Date, title, description, locations, impacts (labels). **Multi-class text classification.**
- **2011 – 2020, state, county, and city-level impacts, 14178 records.**

Data Description – Drought Impact Reporter (DIR)



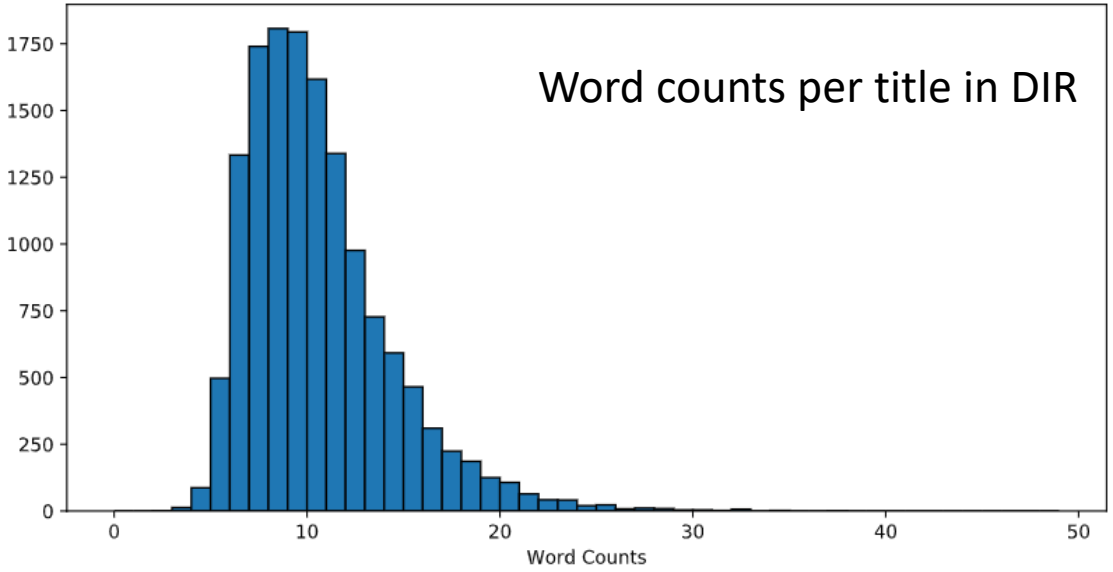
Imbalanced data distribution of various categories in DIR

Data Description – Drought Impact Reporter (DIR)

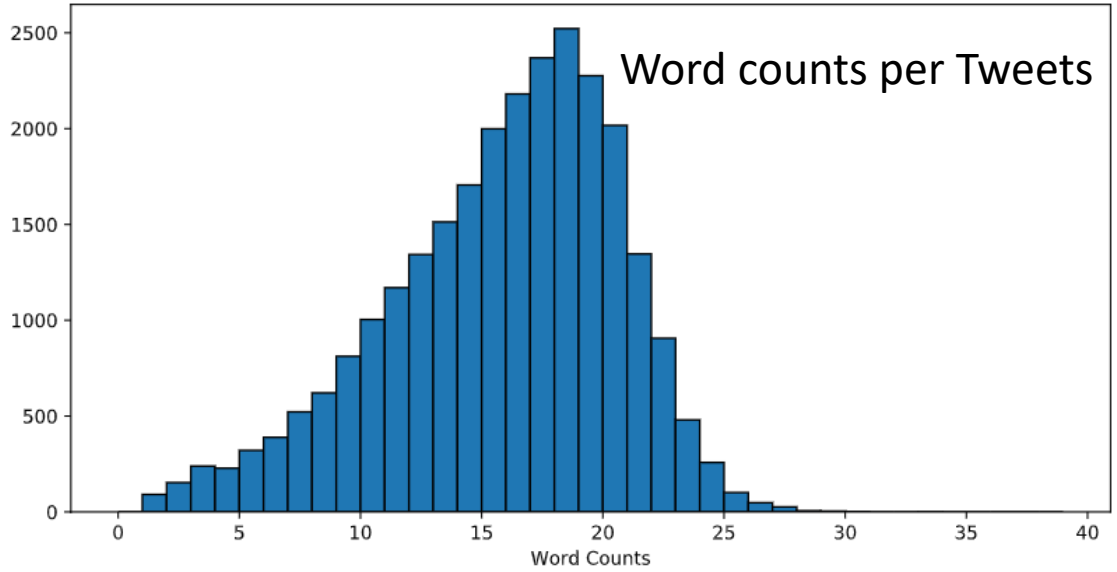


Counts the number of words per description in DIR

98% are within 0-250 words, 25 – 50 words are the most common



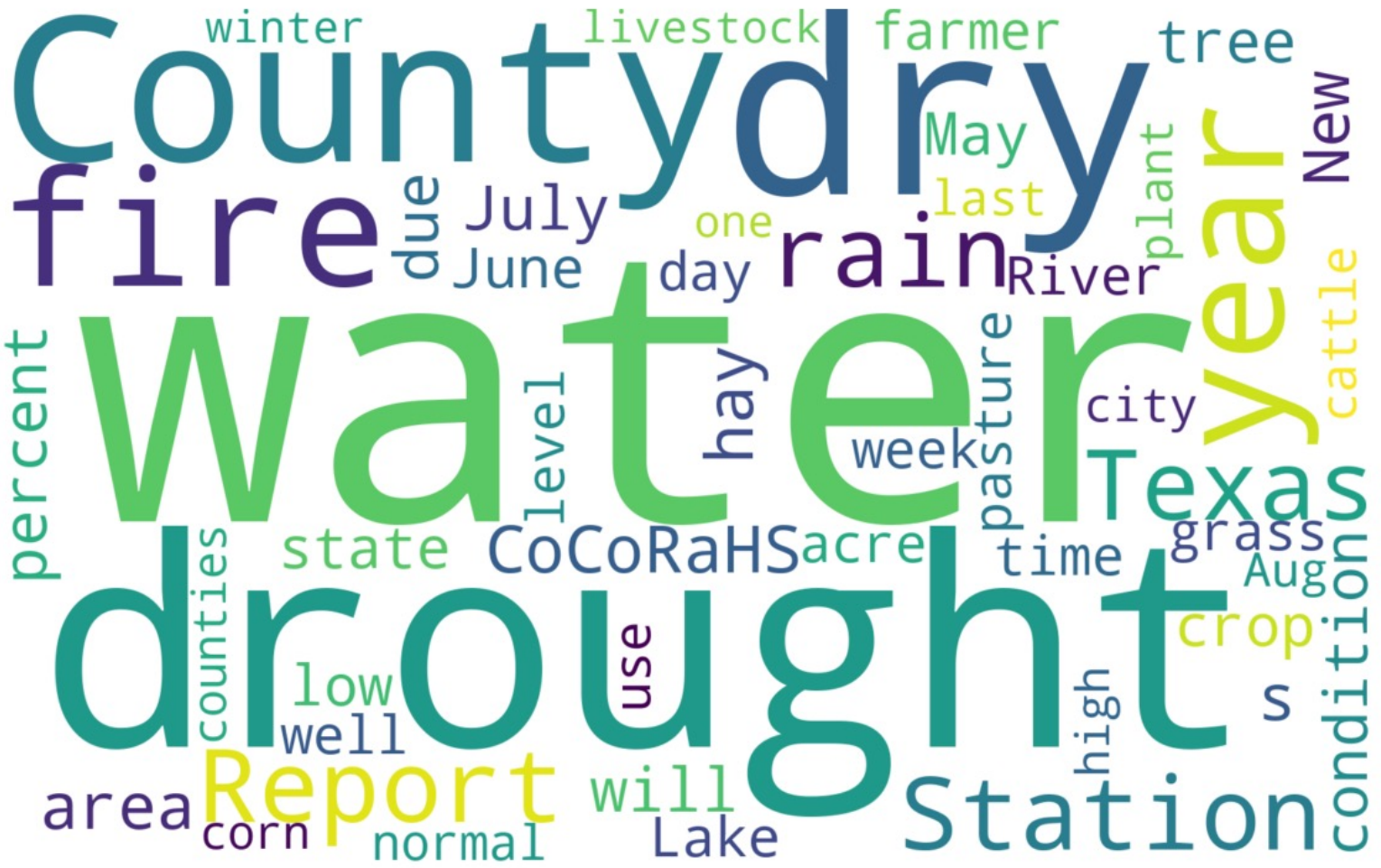
Word counts per title in DIR



Word counts per Tweets

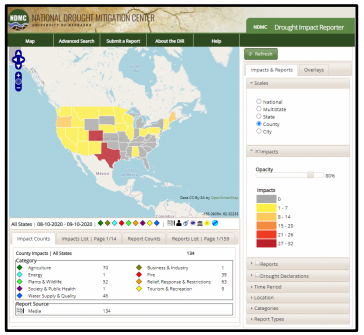
Data Description – Drought Impact Reporter (DIR)

Word Cloud



Research Ideas

DIR – Training Dataset (14178)



Pre-trained word-embedding models + RNN
or
Pretrained BERT-based models

Identify if Tweets indicate types of droughts impacts

Tweets – Test Dataset (26654)

Questions:

1. 9 multi-classes: too many for models to learn based on the size of dataset?
2. Study regions - the U.S. vs top states. Tweeting habits are different in different states.
3. Calibration – filter some Tweets to calibrate the results.
4. Innovation in DL/NLP models.

Gantt Plot

TASK TITLE	PCT OF TASK COMPLETE	May							
		Week1		Week2		Week3		Week4	
Identify research topics	0%	*	*						
Pre-process training and test sets	20%		*	*	*				
Build and test models	60%			*	*	*	*		
Compare and analyze results	80%						*	*	
Write and revise the conference paper	100%							*	*

Thank you!
Welcome to comments and questions!