Pregnancy and Parenting

A Reddit NLP Project by
Curtis M Hope Hill

Roadmap

1. Background
2. EDA, Data Cleaning, & Preprocessing
3. Classification Models
   3.1. Logistic Regression
   3.2. Random Forests
4. Findings and Conclusions
5. Next Steps

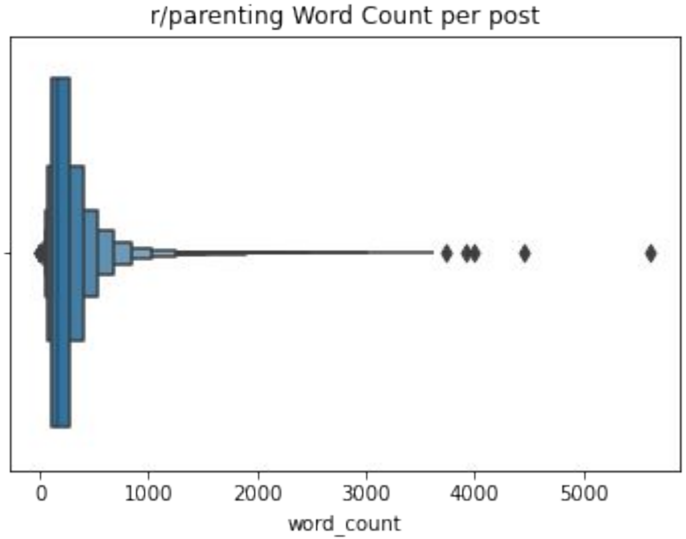
Background

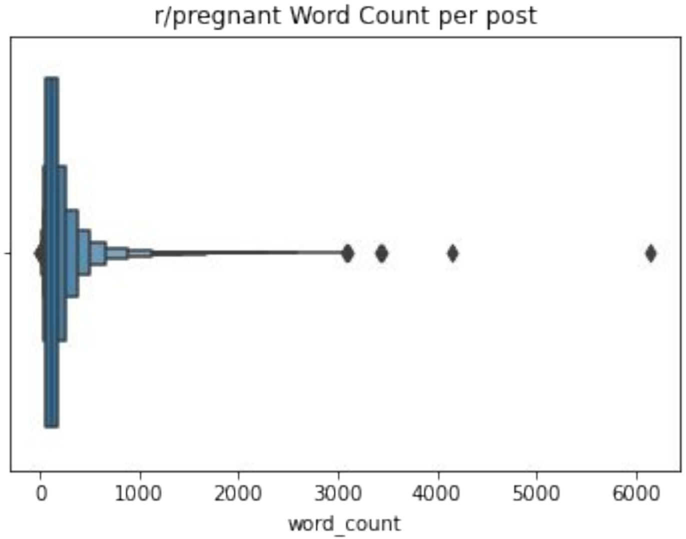
- Pregnancy and parenting are very personally relevant.
- First time parents may be curious about both the pregnancy experience and the transition into parenting.
- Reddit and other social media options likely to have been one of the main ways for community or support while pregnant and/or as parents.

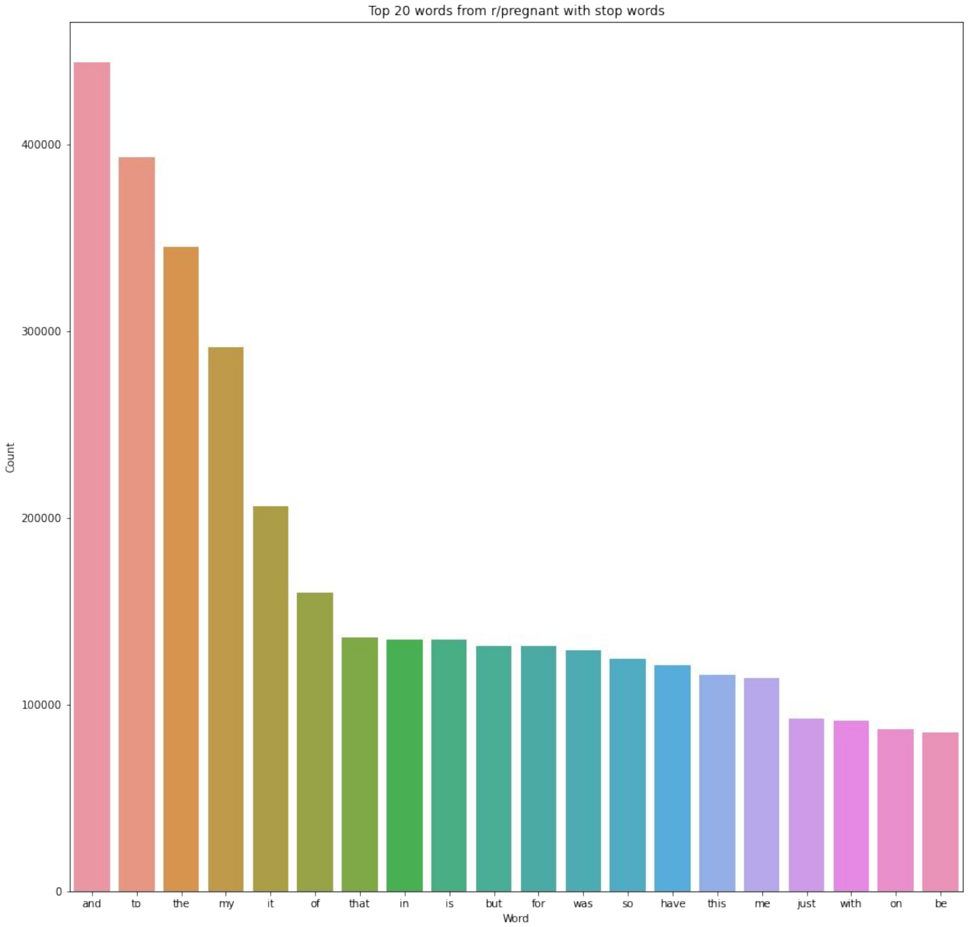
EDA, Cleaning, & Preprocessing

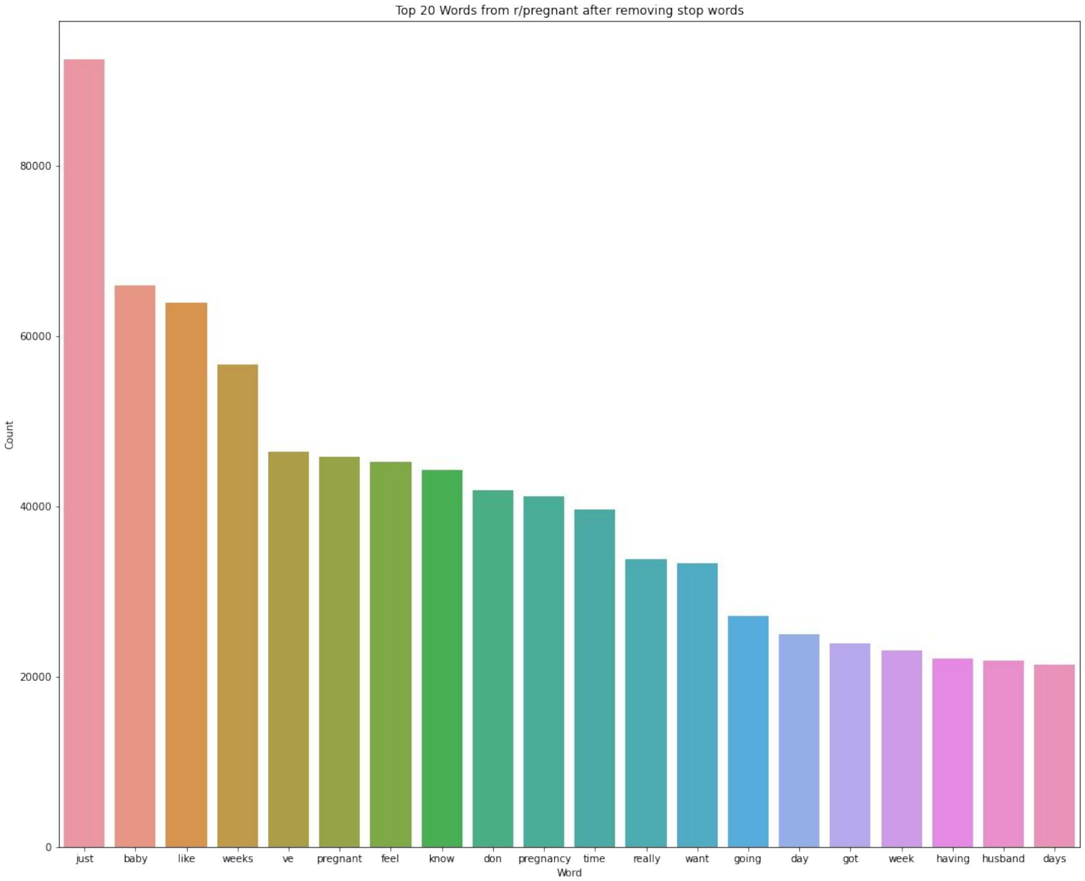
- Dropped nulls & Duplicates
- Dropped posts with links in them
- Lemmatizer to split out each word from posts
- Cleaned up iteratively through process
- Down sampled to 25,000 from each reddit to ease modeling

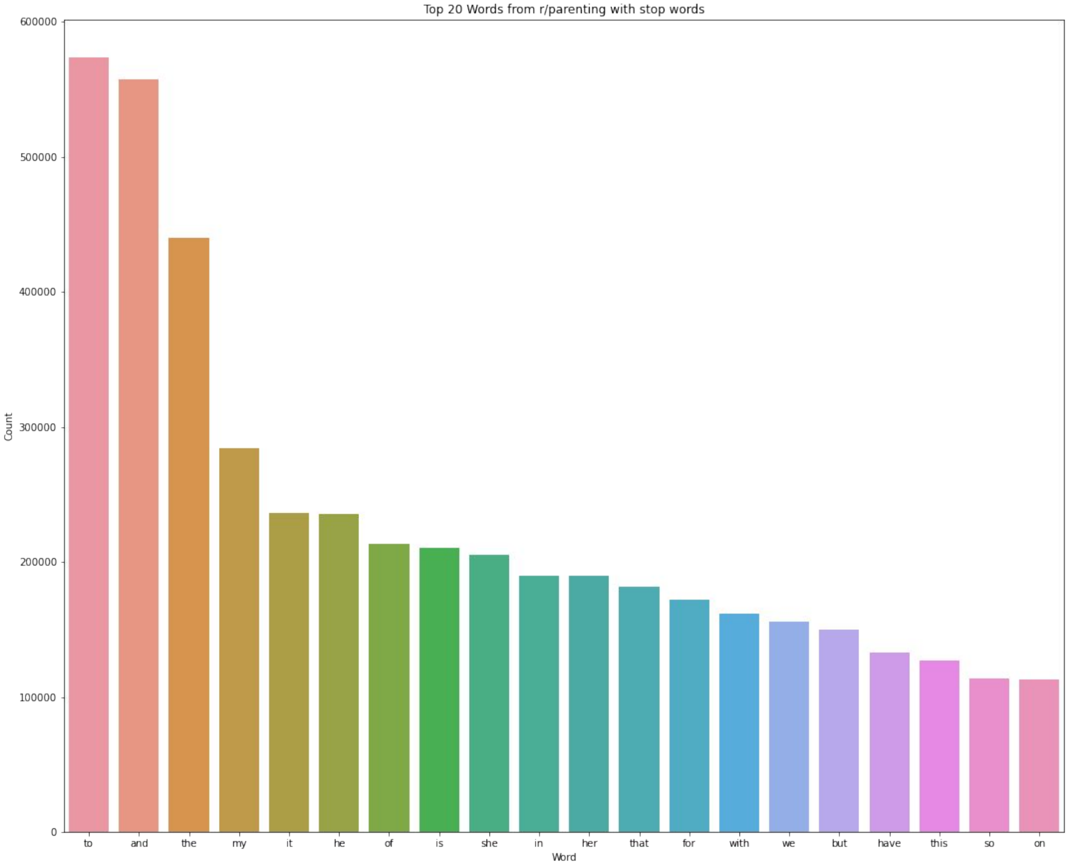
Word Count Distributions

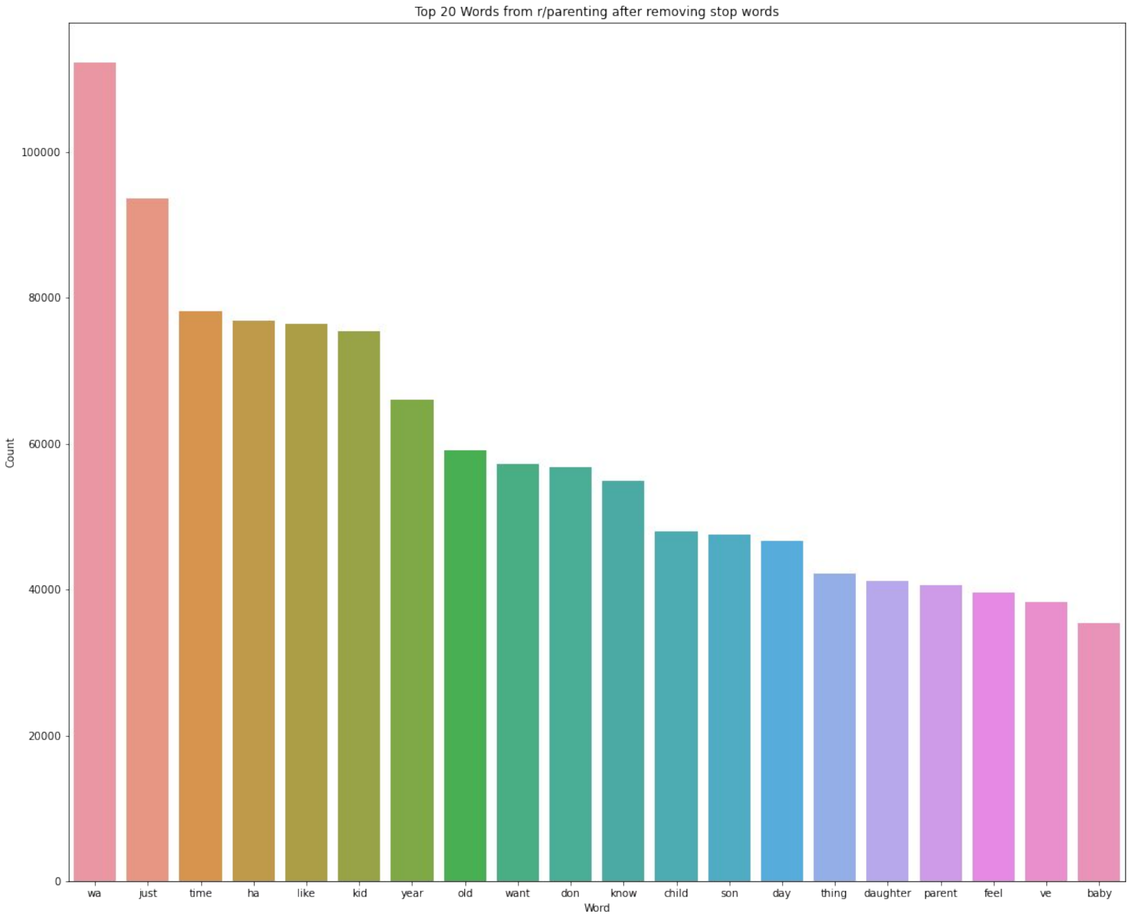





Top 20 words from r/pregnant with stop words
Count
400000
300000
200000
100000
0
Word
and
to
the
my
it
of
that
in
is
but
for
was
so
have
this
me
just
with
on
be


Top 20 Words from r/pregnant after removing stop words
Count
80000
60000
40000
20000
0
just
baby
like
weeks
ve
pregnant
feel
know
don
pregnancy
time
really
want
going
day
got
week
having
husband
days
Word


Top 20 Words from r/parenting with stop words
Count
600000
500000
400000
300000
200000
100000
0
to
and
the
my
it
he
of
is
she
in
her
that
for
with
we
but
have
this
so
on
Word



Top 20 Words from r/parenting after removing stop words

Modeling

- Logistic Regression and Tfidf Vectorizer best model
  - 97% on Train
  - 95% on Test
- Random Forests was heavily overfit
  - 99.9% on Train,
  - 58.9% on Test

Conclusions

- Logistic Regression performed as the best classifier on ‘all_lems’ and appears to be the best option for use when comparing r/pregnant and r/parenting.

Acknowledgements

- Big thanks to all of you who helped me throughout the last week + with parts of this project.
    - Jeffrey Floyd, Mark Harris, Terri John, Matthew Ludwig
    - John Hazard

Questions?