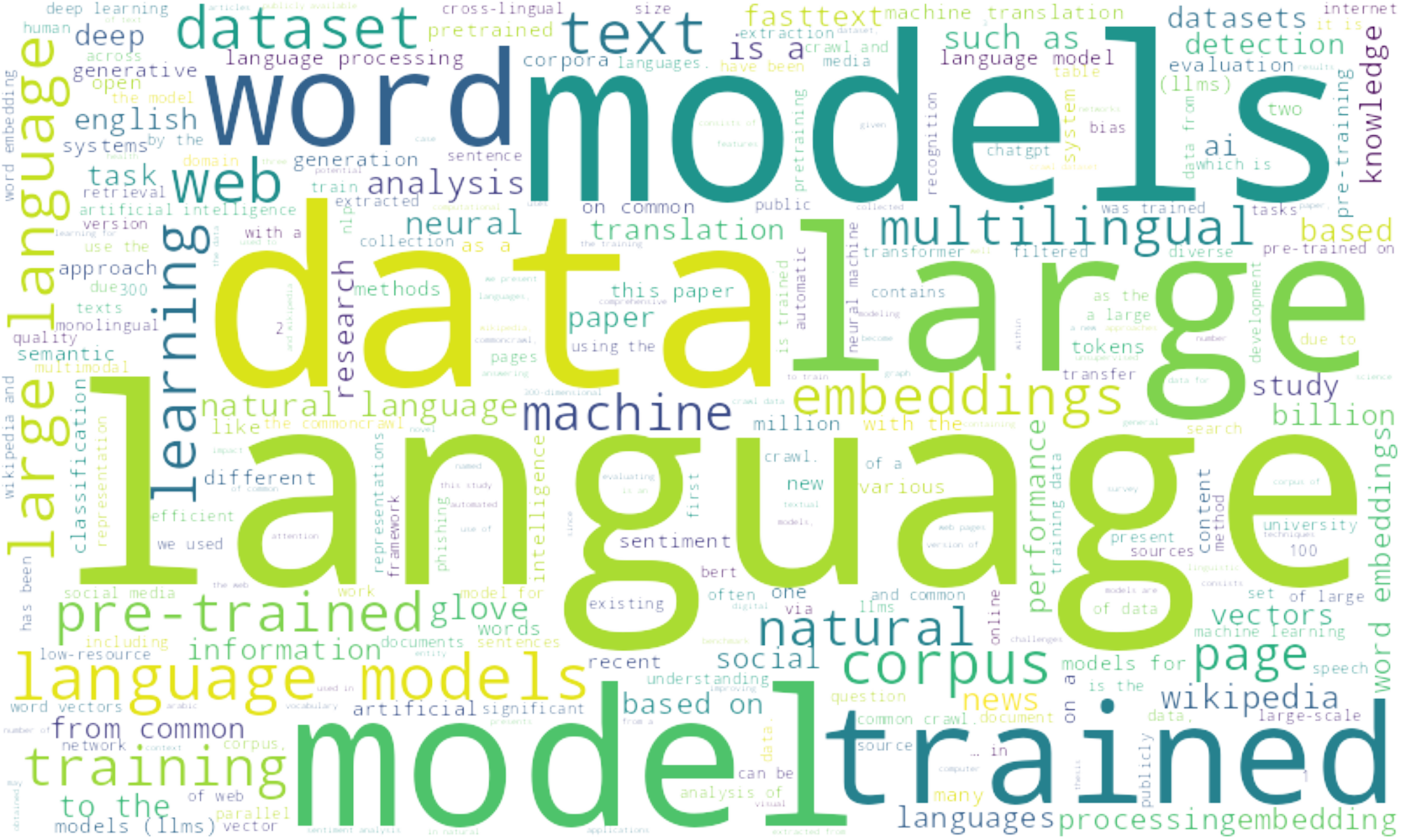
models
word
data
large
language
trained
dataset
text
deep
web
learning
large language
multilingual
embeddings
machine
natural language
pre-trained
language models
training
corpus
natural
social
page
wikipedia
vectors
word embeddings
knowledge
datasets
detection
such as
is a
fasttext
pretrained
cross-lingual
machine translation
language processing
generative
open
english
systems
task
analysis
neural
translation
based
research
methods
this paper
paper
tokens
study
billion
performance
classification
different
efficient
we used
sentiment
bert
glove
information
from common
to the
models (llms)
based on
news
languages
processing
embedding
chatgpt
ai
evaluation
(llms)
two
pre-training
recognition
pretraining
system
bias
transfer
transformer
filtered
diverse
pre-trained on
was trained
tasks
on common
public
collected
extracted
generation
sentence
train
retrieval
artificial intelligence
version
with a
use the
approach
due
300
texts
monolingual
quality
semantic
multimodal
wikipedia and
has been
social media
work
model for
existing
often
one
and common
llms
via
digital
recent
understanding
question
common crawl
document
source
can be
analysis of
many
publicly
data
large-scale
speech
machine learning
set of large
of data
models for
on a
is the
word vectors
arabic
vocabulary
artificial
significant
presents
from a
network
corpus
parallel
vector
sentiment analysis
in natural
applications
extracted from
computer
thesis
obtained
buy
of web
low-resource
including
documents
sentences
words
representations
framework
phishing
automated
use of
attention
intelligence
automatic
is trained
neural machine
representation
the commoncrawl
like
million
with the
crawl.
of a
various
new
first
evaluating
is an
textual
models,
web pages
version of
training data
survey
corpus of
linguistic
consists
techniques
100
university
present
sources
content
method
search
general
containing
crawl data
graph
to train
300-dimensional
answering
wikipedia,
commoncrawl,
comprehensive
languages,
we present
pages
using the
contains
modeling
within
number
unsupervised
a large
as the
a new approaches
development
due to
science
data for
crawl dataset
paper,
tasks
pre-trained on
diverse
filtered
transformer
well
collected
public
computational
utah
nlp
extracted
collection
as a
the training
learning for
used to
the data
version
artificial intelligence
retrieval
health
domain
three
potential
case
features
consists of
given
networks
results
table
recognition
system
bias
chatgpt
data from
which is
pre-training
knowledge
internet
it is
media
have been
languages.
crawl and
corpora
extraction
dataset.
size
deep learning
articles
publicly available
of text
human
across
the model
by the
word embedding
generative
language model
evaluation
(llms)
two
ai
knowledge