Methodology

Twitter analysis

This report contains two different analyses of Twitter hashtags: an analysis of the volume of tweets over time mentioning certain hashtags and a content analysis of the major topics mentioned in tweets using a specific subset of hashtags. Each is discussed in greater detail below.

Hashtag volume analysis

To examine the frequency with which the #MeToo hashtag is used on Twitter, researchers used Crimson Hexagon, a Twitter analysis service, to count the total number of tweets per day mentioning #MeToo for the time period starting Oct. 15, 2017, and ending Sept. 30, 2018.

Content and language analysis of tweets referencing #MeToo

In addition to analyzing the frequency with which #MeToo is used on Twitter and the languages used in those tweets, Pew Research Center also conducted a content analysis of tweets referencing the #MeToo hashtag. Researchers selected tweets from five different time periods close to major news events in order to better understand the nature of the conversation occurring around #MeToo during high-volume periods. The five periods chosen were as follows:

Date ranges evaluated in Twitter content analysis of #MeToo hashtags

Date range	Corresponding events	Total number of tweets
Oct. 16-21, 2017	Harvey Weinstein resigns from the board of his entertainment company (Oct. 17)	485,212
Dec. 6-13, 2017	Time Magazine names #MeToo activists as persons of the year (Dec. 6)	151,487
Jan. 8-13, 2018	Numerous presenters and award recipients discuss sexual misconduct at 75th annual Golden Globes Awards (Jan. 7)	167,318
March 9-14, 2018	International Women's Day (March 8)	71,655
April 7-12, 2018	Three members of the Swedish Academy resign their positions, citing allegations against a high-profile figure close to their group (April 6)	92,724

Source: Pew Research Center analysis of all publicly available tweets (obtained using software from Gnip) containing the #MeToo hashtag from each time period. Categories are not mutually exclusive; individual tweets could be assigned to one or more categories.

PEW RESEARCH CENTER

Researchers collected all publicly available tweets (with duplicates removed) during the time periods listed above that contained the #MeToo hashtag. This initial selection process resulted in a total of 968,396 tweets collected. The tweets were collected using Twitter's Gnip API (application

program interface). The language of each tweet was then determined using a Python package called "langdetect" that references the Google language detection library.

After calculating the share of tweets that mentioned different languages, non-English tweets were removed using the same Python package. Tweets were considered non-English if the algorithm determined there was 0% chance the tweet was in English. This selection process resulted in 692,149 English-language tweets mentioning this hashtag during the time periods listed. These tweets were used in the content analysis described below.

Human coding of a subset of tweets

From the above list of English-language tweets, researchers selected a random representative sample of 250 tweets using the simple random sample function in Python. Each of these 250 tweets was hand-coded by Pew Research Center staff into the categories outlined in the table below based on the content of the tweet.

Categories and rules for classification for initial training sample

Category label	Brief description	Other notes
Personal narratives	Tweet contains discussion of harassment or assault that is explicitly about the user	Does not include general comments about feminism
Mentions of celebrities and entertainment	Tweet mentions prominent Hollywood figures, such as Harvey Weinstein or Alyssa Milano	Also includes mentions of Golden Globes, Time Magazine Person of the Year and general discussions of Hollywood
Mentions of politics or political figures	Tweet mentions political or politics-adjacent figures such as Al Franken, Donald Trump or Clarence Thomas.	Also includes references to political parties, the White House or George Soros

Note: Categories and descriptions developed by Pew Research Center coders.

PEW RESEARCH CENTER

Once this initial sample of 250 tweets was grouped into categories, researchers identified the keywords that best differentiated these categories from each other. An automated search for tweets containing this list of keywords for each category was tested against the reliability of the coders. The rate of agreement between the keyword search and the coders was consistently around 75%-100%.

Topic modeling analysis using category keywords

For the final step in this process, researchers calculated the prevalence of each topic across the entirety of the sample of tweets, using an automated process to search for the keywords developed

during the coding process and classify them into the appropriate categories. An individual tweet could mention one or more of these topics and the tweets that mentioned multiple topics were counted in each relevant category.

Survey methods

The American Trends Panel (ATP), created by Pew Research Center, is a nationally representative panel of randomly selected U.S. adults recruited from landline and cellphone random-digit-dial surveys. Panelists participate via monthly self-administered web surveys. Panelists who do not have internet access are provided with a tablet and wireless internet connection. The panel is being managed by GfK.

Data in this report are drawn from the panel wave conducted May 29-June 11, 2018, among 4,594 respondents. The margin of sampling error for the full sample of 4,594 respondents is plus or minus 2.4 percentage points.

Members of the American Trends Panel were recruited from several large, national landline and cellphone random-digit-dial (RDD) surveys conducted in English and Spanish. At the end of each survey, respondents were invited to join the panel. The first group of panelists was recruited from the 2014 Political Polarization and Typology Survey, conducted Jan. 23-March 16, 2014. Of the 10,013 adults interviewed, 9,809 were invited to take part in the panel and a total of 5,338 agreed to participate.[1] The second group of panelists was recruited from the 2015 Pew Research Center Survey on Government, conducted Aug. 27-Oct. 4, 2015. Of the 6,004 adults interviewed, all were invited to join the panel, and 2,976 agreed to participate.[2] The third group of panelists was recruited from a survey conducted April 25-June 4, 2017. Of the 5,012 adults interviewed in the survey or pretest, 3,905 were invited to take part in the panel and a total of 1,628 agreed to participate.[3]

The ATP data were weighted in a multistep process that begins with a base weight incorporating the respondents' original survey selection probability and the fact that in 2014 some panelists were subsampled for invitation to the panel. Next, an adjustment was made for the fact that the propensity to join the panel and remain an active panelist varied across different groups in the

[1] When data collection for the 2014 Political Polarization and Typology Survey began, non-internet users were subsampled at a rate of 25%, but a decision was made shortly thereafter to invite all non-internet users to join. In total, 83% of non-internet users were invited to join the panel.
[2] Respondents to the 2014 Political Polarization and Typology Survey who indicated that they were internet users but refused to provide an email address were initially permitted to participate in the American Trends Panel by mail but were no longer permitted to join the panel after Feb. 6, 2014. Internet users from the 2015 Pew Research Center Survey on Government who refused to provide an email address were not permitted to join the panel.
[3] White, non-Hispanic college graduates were subsampled at a rate of 50%.

sample. The final step in the weighting uses an iterative technique that aligns the sample to population benchmarks on a number of dimensions. Gender, age, education, race, Hispanic origin and region parameters come from the U.S. Census Bureau's 2016 American Community Survey. The county-level population density parameter (deciles) comes from the 2010 U.S. decennial census. The telephone service benchmark comes from the July-December 2016 National Health Interview Survey and is projected to 2017. The volunteerism benchmark comes from the 2015 Current Population Survey Volunteer Supplement. The party affiliation benchmark is the average of the three most-recent Pew Research Center general public telephone surveys. The internet access benchmark comes from the 2017 ATP Panel Refresh Survey. Respondents who did not previously have internet access are treated as not having internet access for weighting purposes. Sampling errors and statistical tests of significance take into account the effect of weighting. Interviews are conducted in both English and Spanish, but the Hispanic sample in the American Trends Panel is predominantly native born and English speaking.

The following table shows the unweighted sample sizes and the error attributable to sampling that would be expected at the 95% level of confidence for different groups in the survey:

Group	Unweighted sample size	Plus or minus ...
Social media users	4,316	2.5 percentage points

In addition to sampling error, one should bear in mind that question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of opinion polls.

The May 2018 wave had a response rate of 84% (4,594 responses among 5,486 individuals in the panel). Taking account of the combined, weighted response rate for the recruitment surveys (10.1%) and attrition from panel members who were removed at their request or for inactivity, the cumulative response rate for the wave is 2.4%.[4]

[4] Approximately once per year, panelists who have not participated in multiple consecutive waves are removed from the panel. These cases are counted in the denominator of cumulative response rates.

Topline questionnaire

2018 PEW RESEARCH CENTER'S AMERICAN TRENDS PANEL
WAVE 35 MAY 2018
FINAL TOPLINE

ASK IF SOCIAL MEDIA USER (SNSUSER=1) [N=4,316]:

SM14 Thinking about the content you SEE on social media, approximately how much content would you say is about… **[RANDOMIZE]**

	A great deal	Some	Only a little	None	No Answer
ITEM A NOT SHOWN					
b. Sexual harassment or assault					
May 29-Jun 11, 2018	29	36	24	10	1