**A SENTIMENT ANALYSIS OF COMMENTS ON NEW YORK TIMES ARTICLES**

**Introduction**

Sentiment analysis is a field of application of text mining that analyzes peoples’ opinions, sentiments, evaluations, and emotions from written language such as in social networks, chats, product or service reviews. Given that the emergence of electronic word of mouth statements expressed on the web and especially the evolution of social media has led people to become much more eager than before to share their views and opinions on web regarding various issues (Ravi and Ravi, 2015), there is an ample number of studies involving the sentiment analysis of highly subjective texts like product, service, and movie reviews or Twitter data (Mukwazvure and Supreethi, 2015). However, sentiment analysis on more objective texts like news articles has also gained prominence in research as an approach to extract attitudinal information in newsrooms (Mukwazvure and Supreethi, 2015). Sentiment analysis can serve as an effective way of understanding public opinion on various issues addressed in the news articles based on the associated comments provided by the readers. In this regard, an automatic classification of the comments based on a measure capturing the subjectivity and polarity in a textual input is a more efficient method compared to manually reading and evaluating every comment to discern its semantic orientation (Mukwazvure and Supreethi, 2015). Addressing this issue, current study can be conceived as a preliminary attempt to develop an automatized method in determining the overall sentiment orientation of comments expressed on online news articles associated with various topics.

**Research Question**

The main purpose of this study is to determine how much the topics of New York Times (NYT) online articles are controversial based on the sentiment analysis of reader comments regarding these articles that are associated with different topics. First, the articles were classified into topics using an unsupervised classification method, K-means, and the topic labels were then assigned to the clusters of articles based on the most significant terms found to be associated with them. Then, an aspect based sentiment analysis was conducted to identify the polarity of the comments. That is, after the topic labels of the article clusters were determined, for each topic label the most important target terms (aspects) for the comments were identified and the sentiment scores of the comments associated with a certain cluster of articles were determined based on these significant target terms pertaining to specific topic labels. An overall ternary sentiment orientation (i.e. positive, negative, neutral) evaluation was performed for each article topic. The conclusion on the controversy of a topic was then made on the assumption that the contrast of the overall comment polarity scores of each topic reflects how much the topic associated to a category of news articles is controversial.

**An Overview of the Dataset**

There are two separate but related datasets which were retrieved from Kaggle[[1]](#footnote-1) datasets repository. The first dataset, which was originally called by Kaggle dataset owners as “Articles”, provides various features of NYT articles including article ID number, “newDesk” which is the topic classification label assigned by NYT editors, the keywords pertaining to a specific article, and the “webURL” from which the text of the article can be retrieved[[2]](#footnote-2). The articles were published in NYT in January-May 2017 and January-April 2018. The second dataset, which was originally called by Kaggle dataset owners as “Comments”, provides the comments of individuals on the NYT articles. The original dataset includes 9335 articles. Only the articles (and the related comments) associated with particular topic classification provided by the NYT editors in the “newDesk” column were analyzed in the current study. These topics are “Sports”, “Climate”, “Dining”, “Real Estate”, “Business”, “Politics”, “Culture”, “Arts&Leisure”, “Science”, “Travel”, “Photo”. Consequently, 2394 articles were included in this study. Some peculiar reasons led to this decision. First, the retrieval of article texts from the NYT website is performed in a quite long duration and the retrieval brakes abruptly several times during the process. Second, an unsupervised classification method was used to classify articles. For this purpose, in order to have a guiding reference list in labeling the clusters of articles obtained via K-means clustering, articles associated with these clear-cut topics were selected. Other categories specified in the “newDesk” column, such as “Insider”, “OpEd”, “Special Sections”, “Summary”, “Editorial” were considered as ambiguous categories and the articles associated with these categories were not selected. Lastly, there was a large number of comments on the articles (241785 comments) included in the current study, which was both adequate for performing a text mining tasks and also more reasonable in terms of analysis execution times.

**Methodology and Experimental Results**

First, topic analysis was conducted for articles. Although NYT editors have already had categorized the articles based on their own criteria and the articles were annotated with keywords, these labels and keywords were used only as a guideline for possible topic labels. Before conducting the clustering analysis, a tokenizer was created using Spacy - NLP (Natural Language Processor) to tokenize the article texts. This tokenizer collects tokens which are noun, proper noun, adjective, and verb according to their part of speech in the sentence, turns them into lowercase tokens and lemmatizes them. This tokenizer was assigned as the tokenizer in the TF-IDF vectorizer, which was used to transform the corpus and to map the documents over the vector space. The TF-IDF vectorizer combines natural term frequency with the inverse document frequency. The resulting TF-IDF matrix had a shape of (2396, 53153).

Next, K-means clustering was performed to cluster articles into news categories. The Elbow method was adopted to determine the optimal number of clusters. When the Elbow method was applied, K-means clustering algorithm was run by varying k from 1 to 30 clusters. For each k, the total within-cluster sum of square was calculated. Then, the total within-cluster sum of squares was plotted as a function of the number of clusters k. The optimal number of clusters were determined based on the location of the elbow of the curve, which is the point where the curve starts to become significantly flat and there is relatively little gain from further increasing number of clusters (Thorndike, 1953).

The results revealed that the optimal number of clusters is 21. As a consequence, each article was assigned to one of the 21 new categories. Obviously, these categories were only numerical labels which do not provide any semantic information regarding the topics discussed in the articles. In order to determine a meaningful label for each cluster, which would reflect the main topic discussed in the articles grouped into a particular cluster, the most important terms that can be candidate aspects reflecting the topic were identified.

In order to accomplish this task, first a category map was created using a dictionary where, the keys were the ID numbers of the articles and the values were the categories of the articles, found as a result of the K-means clustering analysis. Next, a simple tokenize function was created which takes one sentence at a time. From each sentence, it takes the lemma of a token and its part of speech. Another function was defined which checks whether the part of speech of a token (collected via the tokenize function) is noun and if it is a noun, takes the context of that noun. That is, it takes three tokens before and three tokens after that noun, if it is the case that these tokens are either noun, proper noun, adjective, or verb. Next, an index was created which is a dictionary, the keys of which are the categories of the articles and the values are another dictionary where the keys are the tokens collected via the second function mentioned in the previous step, and the values are the frequencies of the tokens. For each category, this index was filled in.

The most relevant terms for each category were determined based on their Kullback-Leibler divergence (KL-divergence) scores (Kullback and Leibler, 1951). KL-divergence measurement is a method to determine how two probability distributions are different from each other. KL-divergence scores were computed by multiplying the logarithm of the ratio between the probability of a term’s appearance in a given category and its general probability of appearance in all the categories with the probability of a term’s appearance in that given category. It basically provides the terms that are more specifically related to a given category than to the general corpus. Accordingly, another function to calculate the KL-divergence scores for the tokens for each category of articles was set up, which uses the aforementioned index. For each category of articles, the tokens were ranked in decreasing order according their KL-divergence scores and first 20 of them were collected in a list. As a result of this procedure, for each numeric category, possible aspects associated with the topic of the articles clustered to that category, were discovered. This list can be seen in Table 1 in Appendix. These lists of aspects were used to find a common subsumer title to be used as the topic label for a particular cluster of articles. When identifying the labels based on these aspects, the category entries in the “newDesk” column and the keywords provided by the NYT were also checked and used as guidance. It was observed that the possible titles manifested by the collection of the important aspects, for most of the time corresponded to the news categories already assigned by the NYT editors. The topic labels that were identified are presented in Table 2, below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2: Topic Labels Assigned to Article Clusters[[3]](#footnote-3)** | | | |
| 0 | Sports | 10 | Late Night Shows |
| 1 | Social Networking Services | 11 | TV Programs |
| 2 | Housing | 12 | Real Estate |
| 3 | Wines | 13 | Business |
| 4 | Environmental Protection | 14 | Health Science & Research |
| 5 | Science & Research | 15 | Economy |
| 6 | Dining | 16 | Travel, Vacations, Arts & Leisure |
| 7 | International Trade | 17 | Politics |
| 8 | Theater & Movies | 18 | Sexual Harassment |
| 9 | Baseball | 19 | Technology |

Once the topic label of each cluster of articles has been determined, the comments on the articles on the same topics were collected by merging the two datasets on the article ID numbers. 11 of the comments were discarded from the resulting dataset since they were only numeric entries such as a year or a phone number. As in the first part of the analysis, a category map was created. That is, given that a comment is associated with an article, that comment was assigned to the topic label of that article. So, if in the category map a specific comment ID is searched, then the map provides the corresponding article’s topic label.

In this second part, in order to find the most important aspects (target terms) for each category of comments (topics), exactly the same procedure was followed as in the first part. Once the most important 100 aspects for each category of comments were found, they were appended to a list which was then turned into a dictionary where the keys were the categories and the values were the most important 100 aspects. The aspects were considered to be the most significant sub-topics or target terms within a particular category of articles. Table 3[[4]](#footnote-4) in the appendices presents the list of aspects per category of articles are presented in.

In the last part of the analysis, five separate but related functions were defined to conduct the sentiment analysis. The first function simply finds the part of speech of a token in the comment sentences. This function was then embedded in the second function. The second function computes, using the SentiWordNet lexical resource (Esuli and Sebastiani, 2006) interfaced via Python’s Natural Language Toolkit (NLTK)[[5]](#footnote-5), the “positive”, “negative”, and “objective” polarity scores of a token based on its part of speech. SentiWordNet is an extension of WordNet (Fellbaum, 1998), in which three numerical polarity scores of positivity, negativity, and objectivity are assigned to the synsets of words (Baccianella, Esuli, & Sebastiani, 2010). The relationship between the word and its synset is determined by the part of speech of that word. For instance, the word “good” will be annotated with a different ternary of polarity scores depending on its part of speech (e.g. “good” as an “adjective” and “good” as a “noun”) since the associated synsets will be different (the word “good” will convey a more positive sentiment if its POS is adjective than it is noun). So, the main assumption underlying this approach is that different senses of the same term may have different opinion-related properties (Esuli and Sebastiani, 2006).

The third function collects tokens in the sentences of a comment (via the same tokenize function defined initially) and checks whether the token is in the dictionary of the most important 100 aspects of that category and if it is so, it takes the context of that target token. That is, it collects three tokens before and three tokens after that target token, if it is the case that these surrounding tokens are either noun, adjective, adverb, or verb.

The fourth function takes a comment belonging to a particular category, collects the tokens using the third function and then using the second function computes the polarity scores for each token selected. The last function includes all the functions defined and performs all the tasks just by specifying the category of the comments. It gives the overall, column-wise mean of the polarity scores pertaining to a category of comments. A final function was defined just for the simple task of turning the lists of polarity scores of comments into a structured data frame with column names “pos”, “neg”, and “obj” (or “neu”). The polarity scores were then screened in order to determine which topics are controversial. Each of the polarity scores ranges from 0.0 to 1.0, and their sum is 1.0. So, each term has, in the sense indicated by the synset, the three opinion-related properties only to a certain degree (Esuli and Sebastiani, 2006).

Due to the lack of a ground truth that might provide the true sentiment scores determined for instance by human annotators as sentiment ratings of the comments, against which to evaluate the sentiment scores obtained in this analysis, the mean polarity scores were compared to the sentiment scores of comments computed using another sentiment analysis tool called VADER (Valence Aware Dictionary and sEntiment Reasoner). VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media (Hutto and Gilbert, 2014). Since it is a widely used and valid instrument for sentiment analysis, the intensity measures obtained via VADER were treated as a benchmark standard in order to evaluate the performance of the analysis.

Table 4 in Appendices reports, for each category of articles, the descriptive statistics of the positive, negative, and objective (=neutral for VADER lexicon) scores for the comments expressed on different categories of articles. In particular, it presents the aggregated mean scores over all the comments associated with a particular category of articles and also the standard deviation values both of the three different sentiment scores obtained via the method adopted in the current study and of the three sentiment intensity scores obtained via the VADER lexicon. The results indicated that, for both methods, the sentiment scores were comparable. Specifically, in terms of positive-negative polarity, for both methods, the positive and negative mean sentiment scores of the comments were very close to zero for all the article categories. Similarly, in terms of objective-subjective polarity, for both methods, the objectivity or neutrality scores were above 0.6 for all the article categories.

In order to have further evidence on the performance of the aspect based sentiment analysis with SentiWordNet lexicon, the comparability of sentiment scores obtained using this method to the scores obtained using the VADER sentiment analysis tool was also scrutinized by computing the Spearman Rank Order Correlation for each of the three types of polarity scores. More precisely, for these two different methods, the correspondence between the rankings of article categories in terms of the comments’ aggregate polarity scores (separately for “positive”, “negative”, and “objective”) was examined.

The Spearman rank correlation coefficient can be interpreted as follows (Zar, 2014): A correlation coefficient of 0 indicates that the magnitudes of the ranks of one variable are independent of the magnitudes of the ranks of the second variable. A positive correlation indicates that the ranks of one variable tend to increase as the ranks of the other variable increase and a negative correlation indicates that the ranks of one variable tend to decrease as the ranks of the other variable increase. If the sequence of ranks is identical for the two variables, it can be concluded that there was a perfect positive correlation, and r = 1.0. A perfect negative correlation (where r= −1.0) would be one in which the magnitudes of the ranks for one variable vary inversely with the sizes of the ranks of the second.

For three types of polarity scores Spearman’s Rho coefficients were *ρ*pos= 0.809, *ρ*neg= 0.722, and *ρ*obj= 0.427 and the coefficients were statistically significant at p<.01, p<.01, and p=0.06, respectively. For negative polarity scores, the sequence of ranks was almost identical for both sentiment analysis methods since the correlation coefficient was very close to 1. Also, for the positive polarity scores, the sequence of ranks was quite similar since the correlation coefficient was above 0.7. For the objectivity polarity scores, although the magnitude of the relationship was not very strong, the *ρ*- coefficient was statistically significant, indicating that the rank orderings of the objectivity scores are partially similar. These results provided further support for the comparability of the scores obtained using the method developed in the current study to the scores obtained using the standard VADER sentiment analysis tool. Overall, the similarity of the scores for both methods can be interpreted as a qualified performance indicator for the aspect based sentiment analysis with SentiWordNet lexicon method.

**Concluding Remarks**

The results of the sentiment analysis indicated that, for all the news article categories, in terms of positive-negative polarity, the mean scores of the comments were very close to zero, whereas in terms of objective-subjective polarity, the mean scores were all above 0.6. All in all, it can be concluded that none of the news article topics were controversial. In other words, for the article topics identified in the study, individuals provided comments that contain mostly neutral words.

There are some limitations to this study which need to be taken into consideration:

An unsupervised classification method, namely K-means, was used to cluster the news articles into topic categories. Given that the K-means algorithm stops when a local optimum is found, the results may depend on the initial random assignment of each observation to one cluster. Moreover, before running the analysis, one needs to specify the number of clusters to be created. Determining the optimal number of clusters based on the elbow method, is an intuitive but also a subjective way of making a decision. These decisions can have a strong impact on the clustering results obtained. In order to determine the optimal number of clusters, different methods can be adopted. Furthermore, in order to validate the clusters obtained, that is, to ascertain whether the clusters that have been found represent the true subgroups in the data, more elaborative techniques, such as assigning a p-value to a cluster, could be adopted (James et al., 2015).

In this study, lists of most significant aspects were used to find a common representative title to be used as the topic label for a particular cluster of articles. The collection of the relevant aspects for each cluster of articles were related concepts which could be relatively easily aggregated under a single topic label. Moreover, the topics that revealed themselves by these aspects were in concordance with the news categories already assigned by the NYT human annotators. Nevertheless, a more systematic approach can be adopted for completing this task, such as, trying to find representative words based on the words synsets’ hypernym relations using WordNet, where the overlapping in the hierarchy of synsets could be investigated to determine common subsumer terms for the groups of aspects.[[6]](#footnote-6) Another alternative approach could be representing each of the aspects as vectors reflecting their context by the relevance of their co-occurrences with other words and finding clusters of aspects by conducting, for instance, K-means clustering based on the similarities of these vectors, under the assumption that aspects that share the same context should also share similar meaning.

In the current study, when determining the overall polarity scores assigned to comments, the tokens whose part of speech are noun, adjective, adverb, or verb that surround the target token as a context were taken into account. However, in their prominent study on the development of SentiWordNet, Esuli and Sebastiani (2006) provided evidence that “adverb” and “adjective” synsets are evaluated as subjective much more frequently than “verb” or “name” synsets, indicating that in natural language, opinionated content is most often carried by parts of speech used as modifiers (i.e. adverbs, adjectives) rather than parts of speech used as heads (i.e. verbs, nouns). Similarly, Shelke et al.(2012) argued that the results of previous research provided supporting evidence for a strong correlation between adjectives and subjectivity and that people most commonly used adjectives to depict most of the sentiments. The researchers further stated that most of the adverbs have no prior polarity but when they co-occur with sentiment bearing adjectives, they can have a major role in determining the sentiment of a sentence. Taking into cognizance these facts, future research can concentrate on adjectives only or adjectives in combination with adverbs as parts of speech in determining the sentiment scores.

**References**

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0. In Proc. of LREC-10.

Esuli, A., Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In Proceedings of Language Resources and Evaluation (LREC-2006).

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Hutto, C., Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2015). An Introduction to Statistical Learning with Applications in R. Springer.

Kullback, S., Leibler, R. A. (1951). On Information and Sufficiency. The Annals of Mathematical Statistics, 22(1), 79-86.

Mukwazvure, A., Supreethi, K. P. (2015). A Hybrid Approach to Sentiment Analysis of News Comments. 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 1-6.

Ravi, K., Ravi, V. (2015). A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications. Knowledge-Based Systems, Volume 89, 14-46.

Shelke, N. M., Deshpande, S., Thakre, V. (2012). Survey of Techniques for Opinion Mining. International Journal of Computer Applications, Volume 57, No.13.

Thorndike, R.L. (1953). Who Belongs in the Family? Psychometrika, 18, 267-276.

Zar, J.H. (2014). Spearman Rank Correlation: Overview. In: Wiley StatsRef: Statistics Reference Online, Wiley, Chichester.

**Appendices**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1. First 20 Important Tokens for Each Category of Articles Ranked According to the KL-divergence Scores** | | | | | | | | | | | |
| **0** |  | **1** |  | **2** |  | **3** |  | **4** |  | **5** |  |
| 'team' | 0.05897 | ‘user’ | 0.06823 | ‘price’ | 0.28854 | ‘wine’ | 0.26202 | ‘car’ | 0.02381 | ‘patient’ | 0.01365 |
| 'season' | 0.04194 | ‘Facebook’ | 0.04772 | ‘list’ | 0.27811 | ‘grape’ | 0.04727 | ‘gas’ | 0.02241 | ‘doctor’ | 0.00949 |
| 'Knicks' | 0.03634 | ‘datum’ | 0.04708 | ‘%’ | 0.25160 | ‘bottle’ | 0.03803 | ‘company’ | 0.02211 | ‘drug’ | 0.00944 |
| 'draft' | 0.03195 | ‘company’ | 0.04178 | ‘size’ | 0.16795 | ‘flavor’ | 0.02795 | ‘oil’ | 0.02083 | ‘study’ | 0.00888 |
| 'franchise' | 0.02825 | ‘privacy’ | 0.02721 | ‘bedroom’ | 0.12388 | ‘producer’ | 0.02219 | ‘vehicle’ | 0.01846 | ‘researcher’ | 0.00824 |
| 'Anthony' | 0.02710 | ‘ad’ | 0.02446 | ‘taxis’ | 0.09334 | ‘region’ | 0.01841 | ‘emission’ | 0.01549 | ‘scientist’ | 0.00657 |
| 'pick' | 0.02660 | ‘information’ | 0.02096 | ‘room’ | 0.07973 | ‘vineyard’ | 0.01620 | ‘energy’ | 0.01532 | ‘specie’ | 0.00644 |
| 'guard' | 0.02531 | ‘app’ | 0.01856 | ‘kitchen’ | 0.06893 | ‘fruit’ | 0.01541 | ‘rule’ | 0.01231 | ‘Dr.’ | 0.00561 |
| 'Oakley' | 0.02493 | ‘people’ | 0.01568 | ‘bath’ | 0.06060 | ‘vintage’ | 0.01250 | ‘climate’ | 0.01100 | ‘care’ | 0.00531 |
| 'player' | 0.02379 | ‘platform’ | 0.01494 | ‘living’ | 0.05940 | ‘winemaker’ | 0.01234 | ‘state’ | 0.01075 | ‘brain’ | 0.00520 |
| 'Jackson' | 0.02229 | ‘content’ | 0.01488 | ‘fireplace’ | 0.04622 | ‘red’ | 0.01107 | ‘agency’ | 0.01062 | ‘treatment’ | 0.00506 |
| 'night' | 0.01839 | ‘network’ | 0.01486 | ‘LISTING’ | 0.04008 | ‘aroma’ | 0.01087 | ‘fuel’ | 0.01057 | ‘health’ | 0.00481 |
| 'Dolan' | 0.01773 | ‘medium’ | 0.01409 | ‘house’ | 0.03890 | ‘pinot’ | 0.01082 | ‘driver’ | 0.01040 | ‘frog’ | 0.00421 |
| 'coach' | 0.01717 | ‘tech’ | 0.01347 | ‘market’ | 0.03644 | ‘★’ | 0.01032 | ‘Uber’ | 0.00960 | ‘abuse’ | 0.00412 |
| 'game' | 0.01643 | ‘account’ | 0.01231 | ‘ABOVE’ | 0.03611 | ‘glass’ | 0.01023 | ‘regulation’ | 0.00955 | ‘nursing’ | 0.00405 |
| 'playoff' | 0.01641 | ‘social’ | 0.01171 | ‘floor’ | 0.03398 | ‘vine’ | 0.00925 | ‘coal’ | 0.00937 | ‘research’ | 0.00384 |
| 'president' | 0.01467 | ‘post’ | 0.01138 | ‘week’ | 0.03392 | ‘winery’ | 0.00860 | ‘carbon’ | 0.00898 | ‘animal’ | 0.00372 |
| 'clause' | 0.01440 | ‘executive’ | 0.01036 | ‘counter’ | 0.03240 | ‘tannin’ | 0.00820 | ‘drilling’ | 0.00897 | ‘dog’ | 0.00360 |
| 'river' | 0.01250 | ‘internet’ | 0.00939 | ‘hardwood’ | 0.02943 | ‘acidity’ | 0.00679 | ‘automaker’ | 0.00884 | ‘prescription’ | 0.00349 |
| 'contract' | 0.01201 | ‘election’ | 0.00897 | ‘ceiling’ | 0.02875 | ‘oak’ | 0.00663 | ‘industry’ | 0.00843 | ‘dna’ | 0.00316 |
|  | | | | | | | | | | | |
| **6** |  | **7** |  | **8** |  | **9** |  | **10** |  | **11** |  |
| 'restaurant’ | 0.02250 | ‘trade’ | 0.12370 | ‘play’ | 0.01450 | ‘game’ | 0.05626 | ‘night’ | 0.03473 | ‘episode’ | 0.03782 |
| 'chef’ | 0.01821 | ‘tariff’ | 0.11254 | ‘show’ | 0.01126 | ‘baseball’ | 0.05487 | ‘Trump’ | 0.02742 | ‘character’ | 0.01006 |
| 'recipe’ | 0.01450 | ‘steel’ | 0.06940 | ‘theater’ | 0.01093 | ‘pitcher’ | 0.03261 | ‘rundown’ | 0.02447 | ‘scene’ | 0.00991 |
| 'cook’ | 0.01260 | ‘aluminum’ | 0.03702 | ‘character’ | 0.00986 | ‘inning’ | 0.03100 | ‘comedy’ | 0.02396 | ‘show’ | 0.00898 |
| 'dish’ | 0.01223 | ‘import’ | 0.03364 | ‘movie’ | 0.00944 | ‘season’ | 0.02976 | ‘highlight’ | 0.02218 | ‘week’ | 0.00887 |
| 'food’ | 0.01208 | ‘administration’ | 0.02716 | ‘film’ | 0.00886 | ‘player’ | 0.02944 | ‘newsletter’ | 0.01993 | ‘season’ | 0.00822 |
| 'sauce’ | 0.00981 | ‘country’ | 0.02401 | ‘production’ | 0.00831 | ‘team’ | 0.02826 | ‘tv’ | 0.01690 | ‘man’ | 0.00598 |
| 'tip’ | 0.00827 | ‘China’ | 0.02282 | ‘actor’ | 0.00783 | ‘run’ | 0.01618 | ‘porn’ | 0.01605 | ‘story’ | 0.00515 |
| 'rice’ | 0.00772 | ‘product’ | 0.02024 | ‘song’ | 0.00673 | ‘pitch’ | 0.01542 | ‘hearing’ | 0.01602 | ‘moment’ | 0.00458 |
| 'menu’ | 0.00743 | ‘american’ | 0.01895 | ‘audience’ | 0.00609 | ‘pitching’ | 0.01420 | ‘host’ | 0.01492 | ‘killer’ | 0.00437 |
| 'flavor’ | 0.00705 | ‘war’ | 0.01626 | ‘music’ | 0.00572 | ‘base’ | 0.01238 | ‘president’ | 0.01455 | ‘’’ | 0.00388 |
| 'chicken’ | 0.00667 | ‘chinese’ | 0.01602 | ‘performance’ | 0.00488 | ‘hitter’ | 0.01188 | ‘meyer’ | 0.01248 | ‘death’ | 0.00354 |
| 'cooking’ | 0.00637 | ‘measure’ | 0.01583 | ‘story’ | 0.00482 | ‘ball’ | 0.01155 | ‘interview’ | 0.01132 | ‘series’ | 0.00352 |
| 'suggestion’ | 0.00636 | ‘adviser’ | 0.01380 | ‘stage’ | 0.00400 | ‘league’ | 0.01144 | ‘previous’ | 0.01127 | ‘life’ | 0.00317 |
| 'meat’ | 0.00610 | ‘market’ | 0.01161 | ‘cast’ | 0.00391 | ‘fan’ | 0.01136 | ‘recommendation’ | 0.01125 | ‘plot’ | 0.00294 |
| 'butter’ | 0.00604 | ‘trading’ | 0.01159 | ‘woman’ | 0.00390 | ‘starter’ | 0.01057 | ‘secretary’ | 0.01014 | ‘line’ | 0.00290 |
| 'beef’ | 0.00592 | ‘company’ | 0.01154 | ‘drama’ | 0.00376 | ‘batter’ | 0.00944 | ‘watch’ | 0.01001 | ‘Chuck’ | 0.00288 |
| 'shopping’ | 0.00580 | ‘Trump’ | 0.01054 | ‘actress’ | 0.00368 | ‘reliever’ | 0.00911 | ‘o’brien’ | 0.00970 | ‘father’ | 0.00283 |
| 'pork’ | 0.00536 | ‘president’ | 0.01052 | ‘man’ | 0.00344 | ‘baseman’ | 0.00890 | ‘star’ | 0.00923 | ‘finale’ | 0.00274 |
| 'bean’ | 0.00525 | ‘export’ | 0.01023 | ‘role’ | 0.00343 | ‘manager’ | 0.00887 | ‘Fallon’ | 0.00905 | ‘zombie’ | 0.00272 |
| **Table 1. (continued)** | | | | | | | | | | | |
| **12** |  | **13** |  | **14** |  | **15** |  | **16** |  | **17** |  |
| 'apartment’ | 0.03152 | ‘company’ | 0.01972 | ‘study’ | 0.04213 | ‘rate’ | 0.05680 | ‘art’ | 0.00379 | ‘athlete’ | 0.02393 |
| 'building’ | 0.02987 | ‘loan’ | 0.00673 | ‘disease’ | 0.03880 | ‘market’ | 0.05516 | ‘artist’ | 0.00276 | ‘medal’ | 0.01941 |
| 'rent’ | 0.01513 | ‘worker’ | 0.00562 | ‘risk’ | 0.03360 | ‘stock’ | 0.04138 | ‘city’ | 0.00274 | ‘sport’ | 0.01668 |
| 'op’ | 0.01229 | ‘customer’ | 0.00525 | ‘cancer’ | 0.03244 | ‘economy’ | 0.03995 | ‘painting’ | 0.00185 | ‘tennis’ | 0.01402 |
| 'home’ | 0.01200 | ‘industry’ | 0.00497 | ‘exercise’ | 0.02483 | ‘growth’ | 0.03825 | ‘museum’ | 0.00182 | ‘gold’ | 0.01376 |
| 'bedroom’ | 0.01156 | ‘employee’ | 0.00468 | ‘blood’ | 0.02472 | ‘percent’ | 0.02831 | ‘life’ | 0.00180 | ‘tournament’ | 0.01305 |
| 'neighborhood' | 0.01136 | ‘student’ | 0.00445 | ‘weight’ | 0.02161 | ‘inflation’ | 0.02498 | ‘day’ | 0.00164 | ‘woman’ | 0.01137 |
| 'co’ | 0.01046 | ‘service’ | 0.00433 | ‘heart’ | 0.02076 | ‘interest’ | 0.02348 | ‘bar’ | 0.00159 | ‘event’ | 0.01049 |
| 'house’ | 0.00989 | ‘consumer’ | 0.00419 | ‘researcher’ | 0.01749 | ‘wage’ | 0.02249 | ‘image’ | 0.00148 | ‘skater’ | 0.01033 |
| 'estate’ | 0.00984 | ‘fund’ | 0.00416 | ‘fat’ | 0.01732 | ‘investor’ | 0.02072 | ‘coffee’ | 0.00148 | ‘team’ | 0.00965 |
| 'housing’ | 0.00939 | ‘government’ | 0.00403 | ‘people’ | 0.01682 | ‘increase’ | 0.01913 | ‘wall’ | 0.00141 | ‘golf’ | 0.00762 |
| 'unit’ | 0.00899 | ‘store’ | 0.00377 | ‘patient’ | 0.01642 | ‘job’ | 0.01770 | ‘place’ | 0.00136 | ‘olympic’ | 0.00746 |
| '-’ | 0.00894 | ‘card’ | 0.00371 | ‘diet’ | 0.01526 | ‘economist’ | 0.01714 | ‘work’ | 0.00135 | ‘competition’ | 0.00652 |
| 'tenant’ | 0.00892 | ‘school’ | 0.00361 | ‘cell’ | 0.01476 | ‘bank’ | 0.01512 | ‘video’ | 0.00134 | ‘champion’ | 0.00631 |
| 'board’ | 0.00869 | ‘contractor’ | 0.00345 | ‘health’ | 0.01426 | ‘tax’ | 0.01499 | ‘street’ | 0.00131 | ‘flag’ | 0.00586 |
| 'neighbor’ | 0.00802 | ‘tax’ | 0.00345 | ‘diabetes’ | 0.01171 | ‘unemployment’ | 0.01334 | ‘beach’ | 0.00125 | ‘skier’ | 0.00560 |
| 'broker’ | 0.00781 | ‘bank’ | 0.00321 | ‘brain’ | 0.01084 | ‘spending’ | 0.01307 | ‘child’ | 0.00119 | ‘match’ | 0.00553 |
| 'resident’ | 0.00779 | ‘payment’ | 0.00316 | ‘age’ | 0.01059 | ‘Fed’ | 0.01206 | ‘visitor’ | 0.00118 | ‘single’ | 0.00552 |
| 'landlord’ | 0.00737 | ‘technology’ | 0.00313 | ‘level’ | 0.01051 | ‘investment’ | 0.01198 | ‘book’ | 0.00117 | ‘ski’ | 0.00532 |
| 'city’ | 0.00704 | ‘system’ | 0.00313 | ‘percent’ | 0.00998 | ‘bond’ | 0.01162 | ‘whiskey’ | 0.00115 | ‘skating’ | 0.00496 |
|  | | | | | | | | | | | |
| **18** |  | **19** |  | **20** |  |  |  |  |  |  |  |
| 'room’ | 0.06279 | ‘Mr.’ | 0.01386 | ‘team’ | 0.03882 |  |  |  |  |  |  |
| 'bedroom’ | 0.05746 | ‘president’ | 0.00778 | ‘player’ | 0.03881 |  |  |  |  |  |  |
| 'bathroom’ | 0.05293 | ‘news’ | 0.00667 | ‘game’ | 0.03172 |  |  |  |  |  |  |
| 'floor’ | 0.04022 | ‘campaign’ | 0.00655 | ‘football’ | 0.01531 |  |  |  |  |  |  |
| 'house’ | 0.03576 | ‘harassment’ | 0.00600 | ‘league’ | 0.01438 |  |  |  |  |  |  |
| 'property’ | 0.03345 | ‘election’ | 0.00585 | ‘coach’ | 0.01356 |  |  |  |  |  |  |
| 'fireplace’ | 0.03144 | ‘woman’ | 0.00554 | ‘season’ | 0.01304 |  |  |  |  |  |  |
| 'buyer’ | 0.02913 | ‘Trump’ | 0.00536 | ‘basketball’ | 0.01277 |  |  |  |  |  |  |
| 'foot’ | 0.02657 | ‘executive’ | 0.00489 | ‘fan’ | 0.00986 |  |  |  |  |  |  |
| 'area’ | 0.02511 | ‘voter’ | 0.00488 | ‘club’ | 0.00985 |  |  |  |  |  |  |
| 'dining’ | 0.02354 | ‘network’ | 0.00484 | ‘sport’ | 0.00868 |  |  |  |  |  |  |
| 'price’ | 0.02300 | ‘allegation’ | 0.00469 | ‘college’ | 0.00751 |  |  |  |  |  |  |
| 'kitchen’ | 0.02198 | ‘candidate’ | 0.00456 | ‘no’ | 0.00727 |  |  |  |  |  |  |
| 'door’ | 0.02043 | ‘medium’ | 0.00414 | ‘tournament’ | 0.00702 |  |  |  |  |  |  |
| 'ceiling’ | 0.01987 | ‘conservative’ | 0.00394 | ‘.’ | 0.00695 |  |  |  |  |  |  |
| 'home’ | 0.01928 | ‘firm’ | 0.00386 | ‘championship’ | 0.00594 |  |  |  |  |  |  |
| 'bath’ | 0.01919 | ‘party’ | 0.00370 | ‘playoff’ | 0.00577 |  |  |  |  |  |  |
| 'living’ | 0.01904 | ‘lawyer’ | 0.00342 | ‘cheerleader’ | 0.00568 |  |  |  |  |  |  |
| 'wood’ | 0.01770 | ‘gun’ | 0.00333 | ‘soccer’ | 0.00551 |  |  |  |  |  |  |
| 'stone’ | 0.01707 | ‘interview’ | 0.00328 | ‘quarterback’ | 0.00507 |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3. Most Important Aspects (Target Terms) Ranked According to the KL-divergence Scores** | | | | | | | | | |
| **0** |  | **1** |  | **2** |  | **3** |  | **4** |  |
| 'team’ | 0.03644 | ‘datum’ | 0.04475 | ‘house’ | 0.07263 | ‘wine’ | 0.40932 | ‘car’ | 0.03342 |
| 'player’ | 0.03300 | ‘user’ | 0.03496 | ‘home’ | 0.05568 | ‘bottle’ | 0.05045 | ‘fuel’ | 0.02245 |
| 'sport’ | 0.02917 | ‘privacy’ | 0.02408 | ‘taxis’ | 0.03205 | ‘glass’ | 0.04501 | ‘coal’ | 0.02073 |
| 'game’ | 0.02379 | ‘Facebook’ | 0.02390 | ‘property’ | 0.02912 | ‘grape’ | 0.02474 | ‘auto’ | 0.01730 |
| 'football’ | 0.02144 | ‘information’ | 0.02359 | ‘area’ | 0.01446 | ‘taste’ | 0.01956 | ‘gas’ | 0.01727 |
| 'athlete’ | 0.01585 | ‘medium’ | 0.02127 | ‘bedroom’ | 0.01424 | ‘winery’ | 0.01039 | ‘emission’ | 0.01599 |
| 'fan’ | 0.01443 | ‘account’ | 0.01273 | ‘estate’ | 0.01412 | ‘vineyard’ | 0.00990 | ‘oil’ | 0.01439 |
| 'basketball’ | 0.01271 | ‘ad’ | 0.01242 | ‘price’ | 0.01313 | ‘red’ | 0.00986 | ‘standard’ | 0.01424 |
| 'coach’ | 0.01103 | ‘company’ | 0.01237 | ‘tax’ | 0.01188 | ‘tasting’ | 0.00888 | ‘air’ | 0.01402 |
| 'college’ | 0.00830 | ‘social’ | 0.00988 | ‘hampton’ | 0.01084 | ‘food’ | 0.00851 | ‘industry’ | 0.01398 |
| 'woman’ | 0.00822 | ‘platform’ | 0.00980 | ‘condo’ | 0.01060 | ‘fruit’ | 0.00832 | ‘pollution’ | 0.01213 |
| 'cheerleader’ | 0.00732 | ‘internet’ | 0.00845 | ‘floor’ | 0.01053 | ‘producer’ | 0.00806 | ‘vehicle’ | 0.01213 |
| 'play’ | 0.00631 | ‘friend’ | 0.00775 | ‘bath’ | 0.01043 | ‘flavor’ | 0.00797 | ‘energy’ | 0.01105 |
| 'league’ | 0.00624 | ‘data’ | 0.00752 | ‘apartment’ | 0.00936 | ‘acidity’ | 0.00742 | ‘administration’ | 0.00844 |
| 'school’ | 0.00579 | ‘fb’ | 0.00654 | ‘wealth’ | 0.00919 | ‘riesling’ | 0.00717 | ‘manufacturer’ | 0.00824 |
| 'ball’ | 0.00450 | ‘FB’ | 0.00628 | ‘mansion’ | 0.00890 | ‘quality’ | 0.00707 | ‘climate’ | 0.00722 |
| 'NFL’ | 0.00448 | ‘people’ | 0.00608 | ‘room’ | 0.00745 | ‘restaurant’ | 0.00671 | ‘environment’ | 0.00676 |
| 'soccer’ | 0.00437 | ‘election’ | 0.00586 | ‘city’ | 0.00732 | ‘price’ | 0.00669 | ‘regulation’ | 0.00651 |
| 'win’ | 0.00414 | ‘business’ | 0.00581 | ‘foot’ | 0.00668 | ‘store’ | 0.00597 | ‘automaker’ | 0.00586 |
| 'man’ | 0.00394 | ‘advertising’ | 0.00527 | ‘bookend’ | 0.00640 | ‘sweetness’ | 0.00591 | ‘water’ | 0.00575 |
|  | | | | | | | | | |
| **5** |  | **6** |  | **7** |  | **8** |  | **9** |  |
| 'patient’ | 0.01676 | ‘food’ | 0.03685 | ‘trade’ | 0.05961 | ‘movie’ | 0.04064 | ‘game’ | 0.16697 |
| 'doctor’ | 0.01275 | ‘restaurant’ | 0.02837 | ‘tariff’ | 0.04098 | ‘film’ | 0.02769 | ‘baseball’ | 0.09442 |
| 'pain’ | 0.00974 | ‘recipe’ | 0.01894 | ‘steel’ | 0.03101 | ‘actor’ | 0.01534 | ‘pitcher’ | 0.05696 |
| 'drug’ | 0.00959 | ‘salt’ | 0.01150 | ‘war’ | 0.02043 | ‘show’ | 0.01386 | ‘inning’ | 0.04453 |
| 'care’ | 0.00759 | ‘meat’ | 0.01135 | ‘country’ | 0.01364 | ‘play’ | 0.01223 | ‘pitch’ | 0.03969 |
| 'life’ | 0.00608 | ‘dish’ | 0.01100 | ‘China’ | 0.00935 | ‘music’ | 0.01213 | ‘batter’ | 0.03841 |
| 'medication’ | 0.00608 | ‘chef’ | 0.01099 | ‘aluminum’ | 0.00890 | ‘theater’ | 0.01049 | ‘ball’ | 0.03394 |
| 'treatment’ | 0.00572 | ‘cook’ | 0.01083 | ‘world’ | 0.00838 | ‘award’ | 0.01008 | ‘fan’ | 0.02731 |
| 'health’ | 0.00477 | ‘egg’ | 0.01019 | ‘product’ | 0.00834 | ‘woman’ | 0.00849 | ‘player’ | 0.02386 |
| 'hospital’ | 0.00458 | ‘cooking’ | 0.00979 | ‘US’ | 0.00788 | ‘character’ | 0.00793 | ‘team’ | 0.02303 |
| 'physician’ | 0.00412 | ‘rice’ | 0.00933 | ‘import’ | 0.00783 | ‘production’ | 0.00776 | ‘strike’ | 0.01946 |
| 'depression’ | 0.00365 | ‘pan’ | 0.00920 | ‘manufacturing’ | 0.00658 | ‘performance’ | 0.00744 | ‘time’ | 0.01884 |
| 'surgery’ | 0.00325 | ‘cream’ | 0.00811 | ‘economy’ | 0.00650 | ‘audience’ | 0.00725 | ‘pitching’ | 0.01817 |
| 'year’ | 0.00311 | ‘meal’ | 0.00779 | ‘Trump’ | 0.00624 | ‘song’ | 0.00717 | ‘mound’ | 0.01780 |
| 'effect’ | 0.00304 | ‘bread’ | 0.00760 | ‘industry’ | 0.00619 | ‘musical’ | 0.00665 | ‘play’ | 0.01251 |
| 'therapy’ | 0.00298 | ‘beef’ | 0.00747 | ‘market’ | 0.00542 | ‘opera’ | 0.00525 | ‘commercial’ | 0.01235 |
| 'symptom’ | 0.00294 | ‘chicken’ | 0.00736 | ‘price’ | 0.00514 | ‘cast’ | 0.00487 | ‘bat’ | 0.01136 |
| 'work’ | 0.00289 | ‘butter’ | 0.00688 | ‘policy’ | 0.00507 | ‘actress’ | 0.00468 | ‘clock’ | 0.01108 |
| 'retirement’ | 0.00288 | ‘flavor’ | 0.00679 | ‘export’ | 0.00494 | ‘director’ | 0.00461 | ‘sport’ | 0.01038 |
| 'study’ | 0.00284 | ‘fish’ | 0.00676 | ‘deficit’ | 0.00493 | ‘story’ | 0.00459 | ‘change’ | 0.00983 |
|  |  |  |  |  |  |  |  |  |  |
| **Table 3. (continued)** | | | | | | | | | |
| **10** |  | **11** |  | **12** |  | **13** |  | **14** |  |
| 'show’ | 0.06667 | ‘episode’ | 0.04483 | ‘apartment’ | 0.02753 | ‘airline’ | 0.01095 | ‘cancer’ | 0.02661 |
| 'night’ | 0.05270 | ‘show’ | 0.03178 | ‘rent’ | 0.02156 | ‘passenger’ | 0.00992 | ‘diet’ | 0.02295 |
| 'Fallon’ | 0.03177 | ‘season’ | 0.02964 | ‘building’ | 0.02146 | ‘flight’ | 0.00773 | ‘weight’ | 0.02133 |
| 'humor’ | 0.03093 | ‘character’ | 0.02283 | ‘neighbor’ | 0.02085 | ‘employee’ | 0.00681 | ‘study’ | 0.01835 |
| 'host’ | 0.03069 | ‘scene’ | 0.02159 | ‘house’ | 0.01596 | ‘customer’ | 0.00668 | ‘exercise’ | 0.01724 |
| 'guest’ | 0.02950 | ‘series’ | 0.01424 | ‘housing’ | 0.01366 | ‘loan’ | 0.00522 | ‘fat’ | 0.01234 |
| 'fallon’ | 0.02924 | ‘plot’ | 0.01066 | ‘tenant’ | 0.01343 | ‘seat’ | 0.00489 | ‘breast’ | 0.01166 |
| 'interview’ | 0.02329 | ‘spy’ | 0.00802 | ‘city’ | 0.01277 | ‘service’ | 0.00372 | ‘disease’ | 0.01139 |
| 'late’ | 0.02258 | ‘Chuck’ | 0.00690 | ‘property’ | 0.01263 | ‘student’ | 0.00371 | ‘food’ | 0.01044 |
| 'comedy’ | 0.02065 | ‘writer’ | 0.00613 | ‘neighborhood’ | 0.01248 | ‘plane’ | 0.00349 | ‘sugar’ | 0.01015 |
| 'comedian’ | 0.01762 | ‘recap’ | 0.00554 | ‘landlord’ | 0.01226 | ‘company’ | 0.00339 | ‘blood’ | 0.00786 |
| 'joke’ | 0.01574 | ‘story’ | 0.00512 | ‘home’ | 0.01139 | ‘store’ | 0.00330 | ‘calorie’ | 0.00718 |
| 'Trump’ | 0.01242 | ‘finale’ | 0.00484 | ‘op’ | 0.00901 | ‘pension’ | 0.00319 | ‘effect’ | 0.00716 |
| 'hair’ | 0.01222 | ‘tape’ | 0.00452 | ‘co’ | 0.00883 | ‘pay’ | 0.00313 | ‘body’ | 0.00709 |
| 'Colbert’ | 0.01174 | ‘zombie’ | 0.00441 | ‘dog’ | 0.00846 | ‘school’ | 0.00299 | ‘health’ | 0.00690 |
| 'tv’ | 0.01081 | ‘agent’ | 0.00426 | ‘rental’ | 0.00768 | ‘worker’ | 0.00277 | ‘heart’ | 0.00662 |
| 'talk’ | 0.01036 | ‘end’ | 0.00420 | ‘owner’ | 0.00761 | ‘business’ | 0.00252 | ‘diabetes’ | 0.00660 |
| 'entertainment’ | 0.00958 | ‘Elizabeth’ | 0.00407 | ‘estate’ | 0.00740 | ‘teacher’ | 0.00242 | ‘risk’ | 0.00650 |
| 'Jimmy’ | 0.00952 | ‘Philip’ | 0.00390 | ‘unit’ | 0.00724 | ‘government’ | 0.00236 | ‘statin’ | 0.00625 |
| 'monologue’ | 0.00827 | ‘Jimmy’ | 0.00379 | ‘noise’ | 0.00719 | ‘cost’ | 0.00236 | ‘sleep’ | 0.00612 |
|  | | | | | | | | | |
| **15** |  | **16** |  | **17** |  | **18** |  | **19** |  |
| 'market’ | 0.04229 | ‘portrait’ | 0.00513 | ‘gun’ | 0.00911 | ‘woman’ | 0.04666 | ‘car’ | 0.05752 |
| 'stock’ | 0.03510 | ‘art’ | 0.00446 | ‘news’ | 0.00721 | ‘man’ | 0.02292 | ‘driver’ | 0.05066 |
| 'economy’ | 0.02393 | ‘artist’ | 0.00287 | ‘voter’ | 0.00689 | ‘harassment’ | 0.02201 | ‘taxi’ | 0.02887 |
| 'tax’ | 0.02335 | ‘word’ | 0.00279 | ‘Trump’ | 0.00627 | ‘behavior’ | 0.01312 | ‘vehicle’ | 0.01731 |
| 'rate’ | 0.02161 | ‘coffee’ | 0.00278 | ‘election’ | 0.00612 | ‘sexual’ | 0.01289 | ‘Uber’ | 0.01533 |
| 'growth’ | 0.01615 | ‘place’ | 0.00263 | ‘party’ | 0.00465 | ("O'Reilly", | 0.00892 | ‘company’ | 0.01516 |
| 'cut’ | 0.01406 | ‘article’ | 0.00255 | ‘medium’ | 0.00465 | ‘victim’ | 0.00797 | ‘ride’ | 0.01180 |
| 'wage’ | 0.01156 | ‘city’ | 0.00246 | ‘vote’ | 0.00393 | ‘news’ | 0.00794 | ‘service’ | 0.01129 |
| '%’ | 0.01138 | ‘book’ | 0.00244 | ‘impeachment’ | 0.00378 | ‘rape’ | 0.00769 | ‘cab’ | 0.01022 |
| 'job’ | 0.01061 | ‘life’ | 0.00224 | ‘campaign’ | 0.00372 | ‘Fox’ | 0.00707 | ‘self’ | 0.00972 |
| 'interest’ | 0.00977 | ‘painting’ | 0.00196 | ‘candidate’ | 0.00361 | ‘abuse’ | 0.00595 | ‘drive’ | 0.00899 |
| 'increase’ | 0.00923 | ‘museum’ | 0.00191 | ‘president’ | 0.00353 | ‘allegation’ | 0.00587 | ‘technology’ | 0.00858 |
| 'inflation’ | 0.00868 | ‘gypsy’ | 0.00191 | ‘hannity’ | 0.00309 | ‘predator’ | 0.00548 | ‘road’ | 0.00752 |
| 'money’ | 0.00783 | ‘woman’ | 0.00177 | ‘trump’ | 0.00295 | ‘advertiser’ | 0.00471 | ‘human’ | 0.00751 |
| 'investment’ | 0.00693 | ‘show’ | 0.00176 | ‘lie’ | 0.00284 | ‘accusation’ | 0.00452 | ‘law’ | 0.00748 |
| 'debt’ | 0.00641 | ‘character’ | 0.00174 | ‘law’ | 0.00279 | ‘settlement’ | 0.00447 | ‘accident’ | 0.00713 |
| 'unemployment’ | 0.00612 | ‘beauty’ | 0.00165 | ‘country’ | 0.00266 | ‘power’ | 0.00390 | ‘pedestrian’ | 0.00670 |
| 'investor’ | 0.00595 | ‘trip’ | 0.00162 | ‘people’ | 0.00253 | ‘assault’ | 0.00388 | ‘city’ | 0.00616 |
| 'credit’ | 0.00572 | ‘cruise’ | 0.00159 | ‘client’ | 0.00240 | ‘sex’ | 0.00381 | ‘transportation’ | 0.00569 |
| 'economic’ | 0.00554 | ‘tourist’ | 0.00153 | ‘truth’ | 0.00233 | ‘justice’ | 0.00378 | ‘traffic’ | 0.00532 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 4.**  **The Descriptive Statistics of the Sentiment Scores for the Comments Expressed on Different Categories of Articles** | | | | | | | | | | | | | |
|  | **Sentiment Scores Obtained via ABSA** | | | | | |  | **Sentiment Scores Obtained via VADER** | | | | | |
|  | polarity | count | mean | std | min | max | polarity | count | mean | std | min | max |
| SPORTS | pos | 11866 | 0.0795 | 0.0729 | 0.0008 | 1.0000 | pos | 16327 | 0.1368 | 0.1224 | 0.0000 | 1.0000 |
| neg | 11385 | 0.0584 | 0.0610 | 0.0004 | 0.8750 | neg | 0.0802 | 0.0885 | 0.0000 | 1.0000 |
| obj | 12767 | 0.8113 | 0.1338 | 0.0417 | 1.0000 | neu | 0.7830 | 0.1330 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| SOCIAL NETWORK SERVICES | pos | 6935 | 0.0625 | 0.0500 | 0.0010 | 0.7917 | pos | 8247 | 0.1012 | 0.0934 | 0.0000 | 1.0000 |
| neg | 6780 | 0.0564 | 0.0559 | 0.0004 | 0.7500 | neg | 0.0826 | 0.0855 | 0.0000 | 1.0000 |
| obj | 7299 | 0.8122 | 0.1253 | 0.1354 | 1.0000 | neu | 0.8160 | 0.1163 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| HOUSING | pos | 558 | 0.0820 | 0.0696 | 0.0021 | 0.4583 | pos | 726 | 0.1259 | 0.1368 | 0.0000 | 1.0000 |
| neg | 528 | 0.0526 | 0.0531 | 0.0003 | 0.4375 | neg | 0.0608 | 0.0948 | 0.0000 | 1.0000 |
| obj | 591 | 0.8050 | 0.1315 | 0.2500 | 1.0000 | neu | 0.8132 | 0.1497 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| WINES | pos | 1458 | 0.0904 | 0.0712 | 0.0016 | 0.7500 | pos | 1780 | 0.1390 | 0.1154 | 0.0000 | 1.0000 |
| neg | 1410 | 0.0540 | 0.0528 | 0.0007 | 0.6875 | neg | 0.0461 | 0.0636 | 0.0000 | 0.6430 |
| obj | 1548 | 0.8038 | 0.1208 | 0.0625 | 1.0000 | neu | 0.8148 | 0.1202 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| ENVIRONMENTAL PROTECTION | pos | 7185 | 0.0640 | 0.0560 | 0.0006 | 0.7500 | pos | 9326 | 0.1041 | 0.0992 | 0.0000 | 1.0000 |
| neg | 7034 | 0.0625 | 0.0637 | 0.0003 | 0.8750 | neg | 0.0968 | 0.1004 | 0.0000 | 1.0000 |
| obj | 7715 | 0.8291 | 0.1210 | 0.1250 | 1.0000 | neu | 0.7991 | 0.1253 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| SCIENCE & RESEARCH | pos | 14359 | 0.0734 | 0.0600 | 0.0008 | 0.8095 | pos | 18210 | 0.1152 | 0.1104 | 0.0000 | 1.0000 |
| neg | 14051 | 0.0666 | 0.0569 | 0.0005 | 0.7500 | neg | 0.0952 | 0.0927 | 0.0000 | 1.0000 |
| obj | 15239 | 0.8000 | 0.1219 | 0.0625 | 1.0000 | neu | 0.7895 | 0.1266 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| DINING | pos | 8646 | 0.0903 | 0.0898 | 0.0010 | 1.0000 | pos | 12248 | 0.1442 | 0.1372 | 0.0000 | 1.0000 |
| neg | 8238 | 0.0556 | 0.0569 | 0.0003 | 0.8125 | neg | 0.0508 | 0.0800 | 0.0000 | 1.0000 |
| obj | 9387 | 0.8042 | 0.1423 | 0.0982 | 1.0000 | neu | 0.8050 | 0.1439 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| INTERNATIONAL TRADE | pos | 11050 | 0.0632 | 0.0542 | 0.0007 | 0.6667 | pos | 13097 | 0.1039 | 0.1003 | 0.0000 | 1.0000 |
| neg | 10734 | 0.0591 | 0.0555 | 0.0002 | 0.8750 | neg | 0.0971 | 0.0914 | 0.0000 | 1.0000 |
| obj | 11601 | 0.8251 | 0.1154 | 0.0833 | 1.0000 | neu | 0.7991 | 0.1229 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| THEATER & MOVIES | pos | 8812 | 0.0972 | 0.0800 | 0.0013 | 1.0000 | pos | 11429 | 0.1577 | 0.1275 | 0.0000 | 1.0000 |
| neg | 8479 | 0.0597 | 0.0618 | 0.0005 | 0.7500 | neg | 0.0705 | 0.0868 | 0.0000 | 1.0000 |
| obj | 9371 | 0.7926 | 0.1346 | 0.0417 | 1.0000 | neu | 0.7718 | 0.1351 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| BASEBALL | pos | 2491 | 0.0644 | 0.0558 | 0.0013 | 0.6875 | pos | 2879 | 0.1138 | 0.1017 | 0.0000 | 0.7180 |
| neg | 2440 | 0.0472 | 0.0382 | 0.0010 | 0.4500 | neg | 0.0656 | 0.0779 | 0.0000 | 1.0000 |
| obj | 2567 | 0.8261 | 0.1087 | 0.2375 | 1.0000 | neu | 0.8206 | 0.1155 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| LATE NIGHT SHOWS | pos | 1162 | 0.0831 | 0.0707 | 0.0032 | 0.8750 | pos | 1365 | 0.1376 | 0.1160 | 0.0000 | 1.0000 |
| neg | 1144 | 0.0715 | 0.0638 | 0.0007 | 0.5000 | neg | 0.0901 | 0.0900 | 0.0000 | 0.7140 |
| obj | 1215 | 0.7735 | 0.1293 | 0.1250 | 1.0000 | neu | 0.7723 | 0.1256 | 0.0000 | 1.0000 |
| **Table 4. (continued)** | | | | | | | | | | | | | |
|  | **Sentiment Scores Obtained via ABSA** | | | | | |  | **Sentiment Scores Obtained via VADER** | | | | | |
| TV PROGRAMS | pos | 5077 | 0.0866 | 0.0779 | 0.0014 | 1.0000 | pos | 6472 | 0.1181 | 0.1241 | 0.0000 | 1.0000 |
| neg | 4929 | 0.0611 | 0.0606 | 0.0004 | 0.8125 | neg | 0.0799 | 0.0871 | 0.0000 | 1.0000 |
| obj | 5402 | 0.7891 | 0.1356 | 0.0417 | 1.0000 | neu | 0.8020 | 0.1372 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| REAL ESTATE | pos | 4471 | 0.0704 | 0.0693 | 0.0011 | 1.0000 | pos | 5531 | 0.1137 | 0.1178 | 0.0000 | 1.0000 |
| neg | 4348 | 0.0551 | 0.0545 | 0.0004 | 0.7500 | neg | 0.0632 | 0.0789 | 0.0000 | 1.0000 |
| obj | 4752 | 0.8092 | 0.1318 | 0.1250 | 1.0000 | neu | 0.8231 | 0.1280 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| BUSINESS | pos | 26786 | 0.0626 | 0.0550 | 0.0005 | 0.7500 | pos | 35749 | 0.1094 | 0.1024 | 0.0000 | 1.0000 |
| neg | 26094 | 0.0589 | 0.0609 | 0.0003 | 0.8750 | neg | 0.0899 | 0.0904 | 0.0000 | 1.0000 |
| obj | 28793 | 0.8233 | 0.1223 | 0.0625 | 1.0000 | neu | 0.8007 | 0.1229 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| HEALTH SCIENCE & RESEARCH | pos | 6363 | 0.0707 | 0.0595 | 0.0007 | 1.0000 | pos | 7985 | 0.1076 | 0.1060 | 0.0000 | 1.0000 |
| neg | 6235 | 0.0641 | 0.0566 | 0.0005 | 0.7500 | neg | 0.0842 | 0.0879 | 0.0000 | 1.0000 |
| obj | 6733 | 0.8124 | 0.1201 | 0.0625 | 1.0000 | neu | 0.8082 | 0.1229 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| ECONOMY | pos | 10380 | 0.0618 | 0.0531 | 0.0004 | 0.7500 | pos | 12564 | 0.1179 | 0.1097 | 0.0000 | 1.0000 |
| neg | 10227 | 0.0563 | 0.0525 | 0.0005 | 0.8125 | neg | 0.0895 | 0.0908 | 0.0000 | 1.0000 |
| obj | 10910 | 0.8280 | 0.1129 | 0.0625 | 1.0000 | neu | 0.7925 | 0.1287 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| TRAVEL, VACATIONS, ARTS & LEISURE | pos | 20849 | 0.0935 | 0.0877 | 0.0006 | 1.0000 | pos | 29297 | 0.1404 | 0.1338 | 0.0000 | 1.0000 |
| neg | 19807 | 0.0640 | 0.0683 | 0.0003 | 0.8750 | neg | 0.0718 | 0.0879 | 0.0000 | 1.0000 |
| obj | 22563 | 0.7962 | 0.1429 | 0.0417 | 1.0000 | neu | 0.7877 | 0.1410 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| POLITICS | pos | 23126 | 0.0707 | 0.0614 | 0.0008 | 0.7500 | pos | 31197 | 0.1125 | 0.1099 | 0.0000 | 1.0000 |
| neg | 22408 | 0.0684 | 0.0704 | 0.0004 | 0.8750 | neg | 0.0947 | 0.0959 | 0.0000 | 1.0000 |
| obj | 24892 | 0.8021 | 0.1378 | 0.0625 | 1.0000 | neu | 0.7928 | 0.1302 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| SEXUAL HARASSMENT | pos | 7861 | 0.0827 | 0.0780 | 0.0009 | 0.7780 | pos | 10772 | 0.1121 | 0.1199 | 0.0000 | 1.0000 |
| neg | 7674 | 0.0785 | 0.0782 | 0.0006 | 0.8750 | neg | 0.1097 | 0.1080 | 0.0000 | 1.0000 |
| obj | 8468 | 0.7901 | 0.1429 | 0.0625 | 1.0000 | neu | 0.7782 | 0.1404 | 0.0000 | 1.0000 |
|  | | | | | | |  | | | | | |
| TECHNOLOGY | pos | 5017 | 0.0636 | 0.0572 | 0.0012 | 0.7500 | pos | 6584 | 0.1117 | 0.1127 | 0.0000 | 1.0000 |
| neg | 4841 | 0.0580 | 0.0568 | 0.0003 | 0.7500 | neg | 0.0852 | 0.0932 | 0.0000 | 1.0000 |
| obj | 5345 | 0.8207 | 0.1247 | 0.1250 | 1.0000 | neu | 0.8030 | 0.1294 | 0.0000 | 1.0000 |

1. The datasets were retrieved from the following website: https://www.kaggle.com/aashita/nyt-comments [↑](#footnote-ref-1)
2. The NYT article texts were retrieved from their URL addresses using a Python code. The duration of this process is quite long. The texts were added to a list which was then saved to a local Excel file, in order not to run the code each time when the dataset is needed. [↑](#footnote-ref-2)
3. Since some of the topics identified overlapped, the articles associated with these topics were grouped under the same label. For instance, articles clustered under the label “Sports” but were related both to “football” and “basketball” and those clustered under the label “Sports” but were mainly related to “basketball” were merged together under the label “Sports”. [↑](#footnote-ref-3)
4. Due to space limitations, only the first 20 of the 100 aspects are presented here. [↑](#footnote-ref-4)
5. NLTK is a collection of Python libraries for Natural Language Processing which provides an interface to WordNet and SentiWordNet. [↑](#footnote-ref-5)
6. In fact, this method was applied but the containment relations of the words provided very general notions such as “entity”, “whole”, “object”, thus, since the results were not very useful, they were not reported in the study. [↑](#footnote-ref-6)