Automatic Detection of Fake News on Social Media Platforms

Marten Risius
Neu Ulm University of Applied Sciences

Automatic Detection of Fake News on Social Media Platforms

Completed Research Paper

Christian Janze
Goethe University Frankfurt
Theodor-W.-Adorno-Platz 4
60323 Frankfurt, Germany
janze@wiwi.uni-frankfurt.de

Marten Risius
University of Mannheim
L 15, 1-6
68161 Mannheim, Germany
risius@uni-mannheim.de

Abstract

This study investigates how fake news shared on social media platforms can be automatically identified. Drawing on the Elaboration Likelihood Model and previous studies on information quality, we develop and test an explorative research model on Facebook news posts during the U.S. presidential election 2016. The study examines how cognitive, visual, affective and behavioral cues of the news posts as well as of the addressed user community can be used by machine learning classifiers to identify fake news fully automatically. The best performing configurations achieve a stratified 10-fold cross validated predictive accuracy of more than 80%, and a recall rate (share of correctly identified fake news) of nearly 90% on a balanced data sample solely based on data directly available on Facebook. Platform operators and users can draw on the results to identify fake news on social media platforms - either automatically or heuristically.

Keywords: Fake News, Machine Learning, Classification, Detection

Introduction

Fake news are fabricated misinformation from allegedly confidable sources devoid of supportive objective facts designed to mislead recipients. While hoaxes and propaganda are well established concepts in traditional media, fake news recently gained attention by being predominantly and personalized disseminated through social media with allegedly effects on elections (e.g., Philippines, USA) (Mozur and Scott, 2016; Allcott and Gentzkow, 2017). Furthermore, fake news' usage is not limited to elections. Incumbent political parties use fake news as means of "public diplomacy" (Kragh and Åsberg, 2017). Beyond politics, fake news exert a direct and persistent effect on the economy as well. For example, the share price of United Airlines dropped by 76% in a matter of minutes after fake news on its bankruptcy emerged (Carvalho, Klagge and Moench, 2011). Thus, considering the societal and economic impact of fake news, their detection is an important topic. However, in a recent study, approximately 75% of adults in the United States were unable to identify fake news as such (Silverman and Singer-Vine, 2016). Thus, in order to overcome the harmful effects of targeted and widespread misinformation, it seems necessary to help users to identify fake news.

Building on the Elaboration Likelihood Model (ELM) and existing empirical work in the field of user-generated content (UGC), we design an explorative study. In particular, we examine how cognitive, visual, affective and behavioral cues of a Facebook news posting as well as the associated comments allow for the prediction of fake news using machine learning methods. Thus, the research question of our study is *how to fully automatically identify fake news using information immediately apparent on social media platforms*.

In our study, we utilize ground-truth data of human fact-checked fake and non-fake news articles posted during the U.S. presidential election 2016. Specifically, we draw on a balanced sample of 460 Facebook postings of nine left-wing, right-wing and mainstream media outlets as well as the 125,725 associated user-comments. Next to an in-depth analysis of factors related to the information source and social judgment helping to explain fake news, we utilize a multitude of machine learning classifiers to predict fake news. Within a stratified 10-fold cross validation of the model along various performance metrics, we show how our best performing configurations achieve a predictive accuracy of more than 80%, and a recall rate (share of correctly identified fake news) of almost 90% on a balanced data sample. As our approach does not rely on any domain specifics (e.g. term frequencies) and works without taking into consideration any data that is not directly available on Facebook, our results enable platform operators to build generalizable fake content detection systems.

The remaining portion of this paper is structured as follows: Section two outlines the theoretical background from the perspective of the ELM, related work in the realm of UGC and our derived exploratory research model. Section 3 presents details of our research design including the data collection and feature engineering procedures as well as the model evaluation strategy. Section 4 presents the results and evaluation of our study as well as a discussion of the findings and limitations. Section 5 concludes the practical and theoretical implications of the study.

Background and Research Model

Dual process theories are commonly referenced to explain differences in the formation of attitudes and persuasion in online environments. The most prominent example is the ELM of persuasion to explain differences in information processing of UGC (Gilovich, Keltner and Nisbett, 2010). It posits that information is either processed through a central (or "systematic") route that considers the logic and cogency of the message supplemented by individually related experiences, memories or images. Alternatively, an argument is processed through the peripheral (or "heuristic") route affected by superficial aspects such as the alleged expertise of the source, for example, insinuated through its appearance (Petty and Cacioppo, 1986) . The exercising of either mode of elaboration depends on the personal relevance of a message, the individual knowledge about the issue at hand and the feeling of responsibility for an outcome (Gilovich et al., 2010). Accordingly, different types of textual and visual cues like source credibility (Zhang, Zhao, Cheung and Lee, 2014), apparent expertise and trustworthiness (Ayeh, 2015; Zhiwei Liu and Park, 2015; Park and Nicolau, 2015), as well as perceived similarity between recipient and contributor (Shan, 2016) have been related to individually evaluated UGC quality. To detect fake news, we consider various cognitive and visual cues of the information source – most of which have previously been found to affect online information quality. Furthermore, a substantial body of social psychology established the impact of social judgment on the presented information (e.g., through pluralistic ignorance or leveling and sharpening of the information) (Gilovich et al., 2010). Social media environments provide us with the unique opportunity to simultaneously consider the audience's social judgment in addition to the characteristics of the information source. We assess the social judgment through the environmental attitudes in terms of its constituting features: affect, behavior and cognition (Gilovich et al., 2010) .

While fraud has been investigated in other traditional business fields like accounting (e.g., profiling of manipulator's characteristics or earnings management), only recently have researchers begun to analyze the manipulation of online reviews (N. Hu, Bose, Koh and Liu, 2012). Considering that no research has yet attempted to detect fake news to the best of our knowledge, we exploratively transfer a comprehensive set of proxies from the related literature on UGC information quality.

Information Source

The most apparent cues of messages in the Facebook feed are the textual and graphic features of a post. Thus, we assume that unethical sites submitting fake news will set the individual's interpretive framework by manipulating the immediately perceived written and visual characteristics. Regarding the textual cues, a growing body of research investigates the role of cognitive cues in terms of various message and title features on the quality of information. Textual length is generally considered to be a proxy for the amount of information provided. Thus, *word count* of the title and message have repeatedly been found to determine information quality (Mudambi and Schuff, 2010; Pan and Zhang, 2011; Korfiatis, García-Bariocanal and Sánchez-Alonso, 2012; Cheng and Ho, 2015; Zhiwei Liu and Park, 2015; Park and Nicolau, 2015; Fang, Ye, Kucukusta and Law, 2016; Qazi et al., 2016; Salehan and Kim, 2016). Emotionality provides insights complementary to the purely factual information. As such, sentiment *polarity* is considered as a major determinant of informational quality (Mudambi and Schuff, 2010; Ghose and Ipeirotis, 2011; Salehan and Kim, 2016; Yin, Mitra and Zhang, 2016). Closely related to the emotional valence, extremeness of opinions – both positive or negative – affect the usefulness of information (Cao, Duan and Gan, 2011; Park and Nicolau, 2015). Thus, we incorporate *loudness* of a message to approximate its explicitness. Furthermore, *readability* denotes the effort necessary to comprehend a text dependent on textual features (e.g., word frequency, sentence length, and lexical density) and reader characteristics (e.g., level of education) (DuBay, 2004) , which is generally associated with information quality (Mudambi and Schuff, 2010; Cao et al., 2011; Ghose and Ipeirotis, 2011; Korfiatis et al., 2012; Fang et al., 2016). People refer to experts or reputable others in order to validate their message or make them more viable. Therefore, our research model also acknowledges whether a post contains a *citation*. Lastly, *questions* affect knowledge exchange quality on social media platforms (Seebach, 2012) and help draw attention when predominantly exposed in a text title (Siering, Zimmermann and Haferkorn, 2014). Thus, we assess whether or not a post contains a question in order to detect fake news.

Regarding visual cues, current technical challenges in obtaining automated graphical information detain referable studies on information quality or fake news. However, research on the Elaboration Likelihood Model has found that *faces* can affect the perception of information by means of signaling sympathy, expertise or attractiveness (Gilovich et al., 2010). Thus, we assess whether a picture contains a face or simply non-human objects. Similarly, the mood conveyed through a picture can be manipulated through its tone with lighter colors commonly indicating rather happy feelings. The research model therefore contains the brightness of the picture as a mean for fake news.

Social Judgment

Social judgment comprehends various biases of information processing attributable to the attitudes prevalent in the social context (Gilovich et al., 2010). Social media platforms provide the unique opportunity to assess the three constituting components of attitudes: affection, behavior and cognition. Thus, our research model contains proxies for the respective attitude cues based on the responses from the social environment that received the message. *Cognitive cues* among the audience refer to the same knowledge related proxies that were elaborated regarding the information source. *Behavioral cues* comprehend the community actions that are influenced by the attitude. While respective research in Facebook is scarce, previous literature has investigated the role of these functionally equivalent features on Twitter. Sharing content generally demonstrates the interest in and connectedness with the retweeted content to one's own friends within one's network (Boyd, Golder and Lotan, 2010). Thus, sharing messages demonstrates a better connection with the source of the information to others. Comments increase the post's share of voice and subsequently the spread of a message, which increases awareness for the present issue (Risius and Beck, 2015). By tagging others in messages users can strike up a conversation with the recipient, intentionally reply to a previous message or – in case of an ongoing conversation – both (Honey and Herring, 2009). In any case it demonstrates the personal relevance of the topic (Krüger, Stieglitz and Potthoff, 2012; Bruns and Stieglitz, 2014). Likes serve as a positive feedback for the sender signaling that a user expresses positive agreement with a message (Kosinski, Stillwell and Graepel, 2013). The *affective cues* refer to the emotions and personal feelings about an attitude object. For this purpose, Facebook introduced the possibility for users to disclose five different emotional responses. Since differentiated emotions have been found to provide incremental information over the general (dis)liking (Risius, Akolk and Beck, 2015), we consider the distinct emotional cues. Overall, considering these deliberations and the insights from related literature we derive the present study's explorative research model (Figure 1).

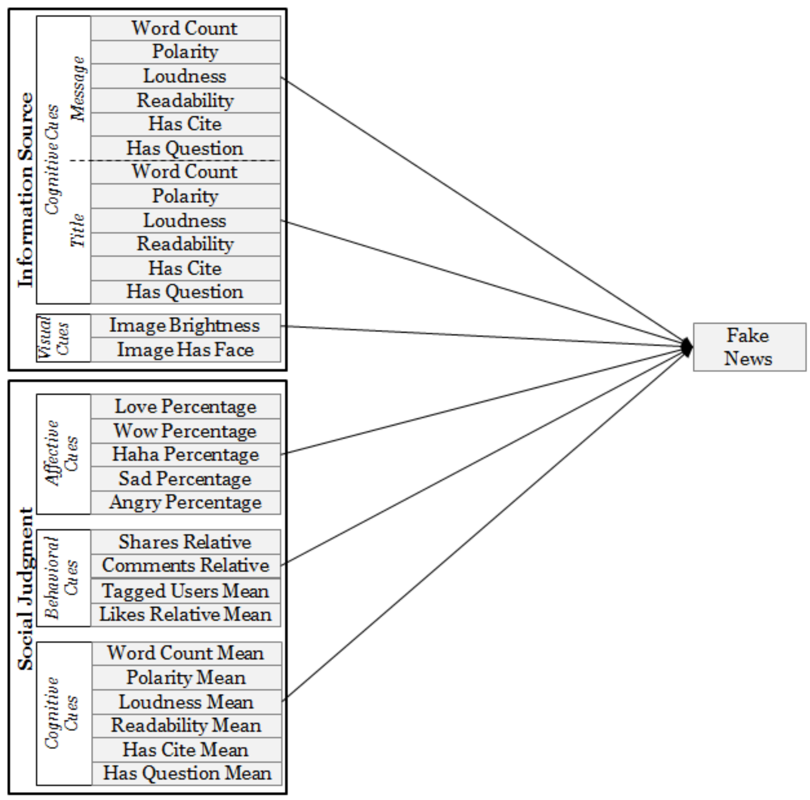

Figure 1. Exploratory Research Model to Identify Fake News

Research Methodology

Data Sample and Feature Engineering

We rely on ground-truth labeling of fake and non-fake news postings on Facebook from BuzzFeed, (Singer-Vine, 2017), which we augment by retrieving additional data via the Facebook developer API. BuzzFeed selected a total of nine self-proclaimed news pages which are active on Facebook and are verified - three left-wing associated pages (*The Other 98%*, 3.24M fans; *Addicting Info*, 1.22M fans; *Occupy Democrats*, 4.14M fans), three right-wing associated pages (Eagle Rising, 0.62M fans; Right Wing News, 3.38M fans; Freedom Daily, 1.36M fans) and three mainstream associated outlets (*Politico*, 1.18M fans; *CNN Politics*, 1.9M fans; *ABC News Polics*, 0.46M fans). (Silverman, Strapagie, Shaban, Hall and Singer-Vine, 2016).

From these nine pages, BuzzFeed fact-checked every post created over a period of seven weekdays (Sept. 19-23. and Sept 26-27, 2016) (Silverman et al., 2016). Each post was randomly assigned to a BuzzFeed rater, which then fact-checked its content and subsequently assigned it to one of four categories ("mostly true", mixture of true and false", "mostly false" and "no factual content"). See Table 1 for additional details on the categories used. In case a human rater was unsure about a specific category, they could also indicate this accordingly. Afterwards, a second rater was assigned to fact-check and rate the same post. In case of a discrepancy of the two ratings, a third reviewer was assigned to resolve the issue. As a sanity check, posts in the final sample that were assigned the label "mostly false" were fact-checked again (Silverman et al., 2016).

Table 1. BuzzFeed Rating Categories (Silverman et al., 2016)	
Category	Description
Mostly True	"The post and any related link or image are based on factual information and portray it accurately. This lets them interpret the event/info in their own way, so long as they do not misrepresent events, numbers, quotes, reactions, etc., or make information up. This rating does not allow for unsupported speculation or claims."
Mixture of True and False	"Some elements of the information are factually accurate, but some elements or claims are not. This rating should be used when speculation or unfounded claims are mixed with real events, numbers, quotes, etc., or when the headline of the link being shared makes a false claim but the text of the story is largely accurate. It should also only be used when the unsupported or false information is roughly equal to the accurate information in the post or link. Finally, use this rating for news articles that are based on unconfirmed information."
Mostly False	"Most or all of the information in the post or in the link being shared is inaccurate. This should also be used when the central claim being made is false."
No Factual Content	"This rating is used for posts that are pure opinion, comics, satire, or any other posts that do not make a factual claim. This is also the category to use for posts that are of the 'Like this if you think...' variety."

From the raw BuzzFeed sample of 2,282 labeled Facebook news posts, we remove posts from "no factual content" category, posts where raters were unsure about their rating (i.e. debatable category), incomplete observations, and posts without any comments or reactions. We combine the category "mixture of true and false" and "mostly false" to the category "fake" and the remaining posts as "non-fake". We then randomly downsample the "non-fake" observation category to yield a balanced sample of 460 posts in our final sample. Using these posts, we update all metrics (e.g. number of shares) and download additional data (e.g. images used in the posts) as well as all 125,725 associated comments on January 18th, 2017.

Figure 2 provides a stylized version of a Facebook news posting of "Addicting Info" as well as a user comment. For easy reference, black circles denote specific data points (A-O). Figure 3 outlines how these data points are used to calculate specific features we use to train and test our machine learning classification models.

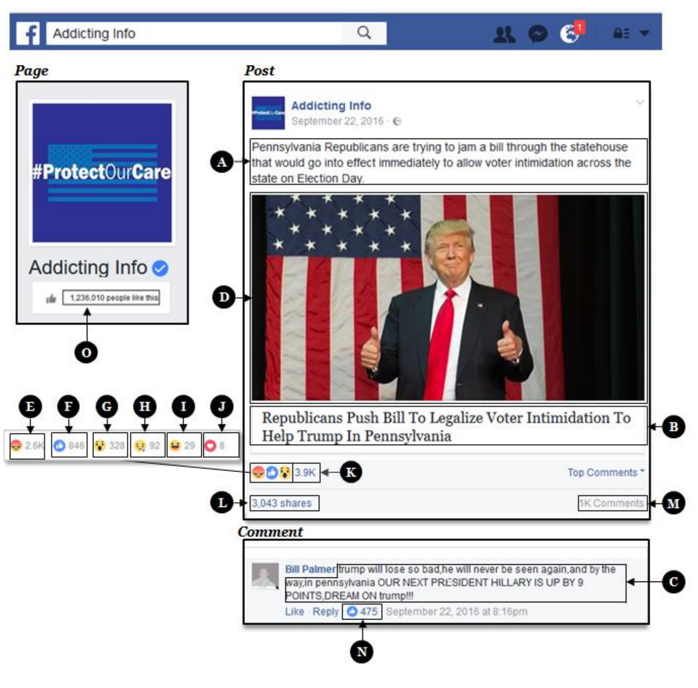

Figure 2. Stylized Facebook News Posting

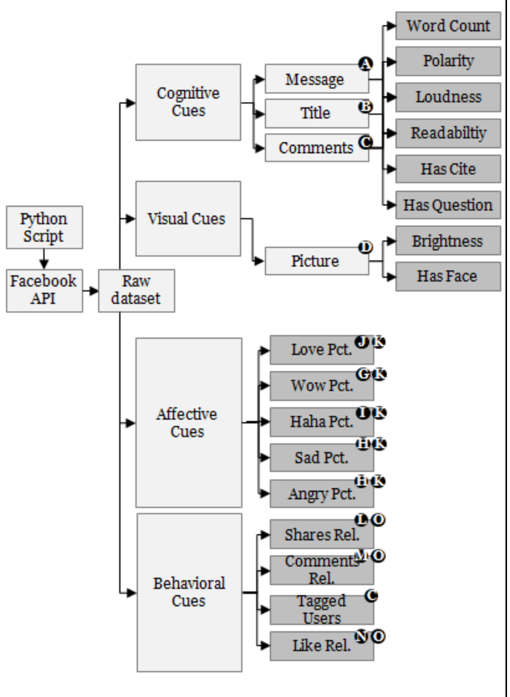

Figure 3. Feature Extraction

Cognitive Cues

We calculate a variety of metrics to measure *cognitive cues* based on the message (*mes_text*) and title (*tit_text*) of a given post as well as their associated comments (*c_text*). First, we calculate the word count (*wc*) of each observation *i* in *K={mes_text, tit_text, c_text}*. Second, we calculate the polarity *(pol*) via a dictionary-based approach as implemented in the R library "qdap". The word list is used in related studies (see for example M. Hu and Liu, 2004a, 2004b; B. Liu, Hu and Cheng, 2005) and entails 2,003 positive and 4,776 negative opinion as well as 23 negation words. Third, we calculate the loudness (*loud*) of the texts by dividing the number of capitalized characters *(cap_cc*) and the number of empathies characters (*emp)* used with *emp={!,*,_}* by the character count as shown in Equation 1. Fourth, we estimate the readability (*read*) by calculating the Flesch–Kincaid grade level (Kincaid, Fishburne Jr, Rogers and Chissom, 1975) as shown in Equation 2. Here, *sent* denotes to the sentence count, *syl* to the number of syllables and *wc* to the word count of a given text *i* in *K*.

$$loud_{i,K} = \frac{cap\_cc_{i,K} + emp\_cc_{i,K}}{cc_{i,K}} \quad (1). \qquad read_{i,K} = 0.39\,\frac{wc_{i,K}}{sent\_c_{i,K}} + 11.8\,\frac{syl_{i,K}}{wc_{i,K}} - 15.59 \quad (2).$$

Fifth, to determine and dummy code whether an observation contains a citation (*has_cit*) or a question (*has_que*) by checking whether it contains at least two quotation signs or at least one question mark with their absence as the reference category. Sixth, in case of the six textual cues calculated for the comment texts' (*wc, pol, loud, read, has_cit, has_que*), we subsequently calculate the arithmetic mean of all variables referring to the same post.

Visual Cues

Next to cognitive cues, we extract *visual cues* from the images of the posts. First, we calculate the brightness of the picture (*img_brightness*) by greyscaling the image and calculating the arithmetic mean of the pixel level using the ImageStat module of the Python library Pillow. Example results are shown in Figure 4. Note that we did not include RGB color values as they are highly correlated with *img_brightness*.

Figure 4. Automatic Image Brightness Extraction

Second, we determine whether the posts' image contains a face (*img_has_face*). We do so by calculating the face count via a deep learning approach. Specifically, we utilize convolutional neural network (CNN) features in a max-margin object-detection (MMOD) algorithm as implemented in "DLIB" version 19.2. The face detection model is trained on 6,975 face images and achieves a recall rate of 87.91% on the FDDB unrestricted face detection benchmark sample (King, 2016). Thus, the approach achieves a state-of-the art performance in this task. Examples of the automatically detected faces are shown in Figure 5. We dummy code *img_has_face* as 1 if face count is equal or greater than 1 and 0 otherwise.

Figure 5. Automatic Face Detection (Figure Limited to 0-3 Faces)

Affective Cues

In terms of *affective cues*, we extract the number of votes each of the six reaction categories received (*p_like, p_love, p_wow, p_haha, p_sad, p_angry) f*rom each post. Furthermore, we calculate the percentage share (*pct*) of each reaction by dividing the number of specific reactions by the total number of reactions of the post (*p_num_reactions)* to yield the variables *like_pct, love_pct, wow_pct, haha_pct, sad_pct and angry_pct*.

Behavioral Cues

As *behavioral cues*, we extract the number of shares (*p_shares*) and comments (*p_ comments*) from each post received and subsequently normalize the data. Specifically, we divide the shares and comments of a given post by the number of fans of the Facebook page (*p_fans*) to yield the variables *p_shares_rel* and *p_comments_rel*. This normalization is necessary as we assume that posts of pages with a high number of fans have a higher reach and are therefore seen by a higher number of Facebook users. From each comment, we retrieve the number of Facebook users tagged within *c_text and store the results in c_tag_usrs* and the number of likes a specific posting received *c_like*. We normalize the latter again by dividing the number by *p_fans* to yield c*_like_rel* using the same rational from above. Then, we calculate the mean of *c_tag_usrs* and *c_like* for all comments of a given post to yield *c_tag_usrs_mean and c_like_mean*. All variables related to cognitive, visual, affective and behavioral cues described above are summarized in Table 2.

Table 2. Variable of a Facebook News Post

Variables	Description
fake	Variable equaling 1 if post contains fake news, and 0 otherwise.
mes_text, tit_text, c_text	String representing the text of a posts' message (mes), a posts title (tit), and a comment (c).
mes_wc, tit_wc, c_wc	Word count of mes_text, tit_text and c_text.
mes_pol, tit_pol, c_pol	Polarity of mes_text, tit_text and c_text.
mes_loud, tit_loud, c_loud	Loudness of mes_text, tit_text and c_text (see Equation 1).
mes_read, tit_read, c_read	Readabiltiy of mes_text, tit_text and c_text (see Equation 2).
mes_has_cit, tit_has_cit, c_has_cit	Variables equaling 1 if mes_text, tit_text, c_text contains a citation, and 0 otherwise.
mes_has_que, tit_has_que, c_has_que	Variables equaling 1 if mes_text, tit_text, c_text contains a question, and 0 otherwise.
img_brightness	Brightness of the picture of a post.
img_has_face	Binary variable yielding 1 if picture of a post contains a face, and 0 otherwise.
p_shares, p_comments, p_love, p_wow, p_haha, p_like, p_sad and p_angry	Absolute number of shares and comments as well love, wow, haha, like, sad and angry ratings a post received.
p_num_reactions	Sum of p_love, p_wow, p_haha, p_like, p_sad, and p_angry.

p_fans	Number of fans of the Facebook page of a post
c_tag_usrs	Number of Facebook users tagged in c_text.
c_like	Number of likes a comment received.
p_shares_rel, p_comments_rel	Relative number of shares and comments of a post as well as likes of a comment (p_shares, p_comments, c_like each divided by p_fans)
p_love_pct, p_wow_pct, p_haha_pct, p_sad_pct, p_angry_pct	Percentage share of p_love, p_wow, p_haha, p_sad, and p_angry by dividing the variables by p_num_reactions.
c_wc_mean, c_pol_mean, c_loud_mean, c_read_mean, c_has_cite_mean, c_has_que_mean, c_tag_usrs_mean, c_like_rel_mean	Mean of the variables c_wc, c_pol, c_loud, c_read, c_has_cite, c_has_que, c_tag_usrs, c_like_rel of all comments of a post.

Study Design and Evaluation Strategy

In our study we train a variety of machine learning classifiers suitable for our binary classification problem. Specifically, Logistic Regression (*LOG, see* Cox, 1958; Walker and Duncan, 1967)), Support Vector Machines (*SVM, see* Cortes and Vapnik, 1995), Decision Tree (*DTR, see* Quinlan, 1986), Random Forest (*RFO, see* Breiman, 2001)) and Extreme Gradient Boosting (*XGB,* see Chen and Guestrin, 2016). We train each machine learning classifier with the same set of features as described for the *LOG* specified in Equation 3. The outcome variable is fake which is binary coded as 1 if a post contains fake news and 0 otherwise.

Regarding the information source, we add cognitive cues from the Facebook posts' message (*mes*) and title (*tit*) texts to our model. Specifically, their word count (*wc*), polarity (*pol*), loudness (*loud*), readability (*read*) as well as whether they contain a citation (*has_cit*) or a question (*has_que*). In addition, we add the visual cues brightness *(img_brightness)* and whether it contains at least one face *(img_has_face*) from the posts' image to the model. In terms of variables related to social judgement, we add affective cues, behavioral cues and cognitive cues. Affective cues entail the relative share of love *(p_love_pct*), wow (*wow_pct*), haha (*p_haha_pct*), sad (*p_sad_pct*) and angry (*p_angry_pct*) votes. Behavioral cues include the relative number of shares and comments a post received (*p_shares_rel* and *p_comments_rel*) as well as the mean number of users tagged in comments (*c_tag_usrs_mean*) as well as the mean number of relative likes (*c_like_rel_mean*) received by comments associated with the specific post. Cognitive terms include the same variables as in case of the information source.

$$\begin{aligned} fake = {} & \beta_1 mes\_wc + \beta_2 mes\_pol + \beta_3 mes\_loud + \beta_4 mes\_read + \beta_5 mes\_has\_cit + \beta_6 mes\_has\_que \\ & + \beta_7 tit\_wc + \beta_8 tit\_pol + \beta_9 tit\_loudness + \beta_{10} tit\_read + \beta_{11} tit\_has\_cit + \beta_{12} tit\_has\_que + \\ & \beta_{13} img\_brightness + \beta_{14} img\_has\_face + \beta_{15} p\_shares\_rel + \beta_{16} p\_comments\_rel + \\ & \beta_{17} p\_love\_pct + \beta_{18} p\_wow\_pct + \beta_{19} p\_haha\_pct + \beta_{20} p\_sad\_pct + \beta_{21} p\_angry\_pct + \\ & \beta_{22} c\_wc\_mean + \beta_{23} c\_pol\_mean + \beta_{24} c\_loud\_mean + \beta_{25} c\_read\_mean + \\ & \beta_{26} c\_has\_cite\_mean + \beta_{27} c\_has\_que\_mean + \beta_{28} c\_tag\_usrs\_mean + \beta_{29} c\_like\_rel\_mean + \varepsilon \end{aligned} \qquad (3).$$

We evaluate our classification models (*LOG, SVM, DTR, RFO and XGB*) via different metrics which are based on a stratified 10-fold cross validation approach. Specifically, we divide our data set of n=460 posts into 10 equally sized folds containing the same amount of fake and non-fake observations randomly selected from the total sample. Then, we take out one fold and train our models with the nine remaining folds. Subsequently, we use the model to predict the outcome variable fake of the left-out fold.

We then calculate a 2x2 confusion matrix where we assign examples where the predicted and actual outcome is fake or non-fake as true positives (*TP*) and true negatives (*TN*) respectively. Examples where the predicted outcome is fake and the actual outcome non-fake as false positives (*FP*) and examples where the predicted outcome is non-fake whereas the actual outcome is fake as false negatives (*FN*). Based on *TP, TN, FP* and *FN*, we calculate various evaluation metrics. First, we calculate the amount of correctly classified examples (*TP, TN*) by dividing their sum by the number of all observations (*TP, TN, FP, FN*). Second, we calculate the error rate by subtracting the accuracy score from 1. Third, we calculate the specificity (=*TN/(TN+FP)*). Fourth the sensitivity, which is also known as recall *(=TP/(TP+FN)*). Fifth, the precision (*TP/(TP+FP*) and lastly the F1-Score (=2*(Precision*Recall)/(Precision + Recall)). We repeat the procedure from above ten times, leaving out each fold one time. Subsequently we calculate the mean of the evaluation metrics grouped by the classifier used.

Empirical Study

Descriptive Statistics

Descriptive statistics of our data sample are shown in Table 3. Specifically, we provide information on the complete sample (n=460 posts) and the no fake (n=230 posts) and fake news (n=230 posts) postings separately.

Table 3. Descriptive Statistics

	Complete (N=460 posts)					*No Fake (N=230 posts)*					*Fake (N=230 posts)*				
	Min.	Max.	Mean	SD	Med.	Min.	Max.	Mean	SD	Med.	Min.	Max.	Mean	SD	Med.
mes_wc	1.0	63.0	16.9	11.9	14.0	1.0	63.0	19.5	11.2	18.0	1.0	60.0	14.3	12.1	11.0
mes_pol	-1.3	1.0	-0.1	0.3	0.0	-1.0	0.9	0.0	0.3	0.0	-1.3	1.0	-0.1	0.4	0.0
mes_loud	0.0	1.0	0.1	0.1	0.1	0.0	1.0	0.1	0.1	0.0	0.0	1.0	0.1	0.2	0.1
mes_read	-3.4	20.7	7.4	5.0	7.4	-3.4	20.7	8.7	5.0	8.4	-3.4	20.5	6.1	4.7	6.2
mes_has_cit	0.0	1.0	0.1	0.3	0.0	0.0	1.0	0.2	0.4	0.0	0.0	1.0	0.1	0.2	0.0
mes_has_que	0.0	1.0	0.1	0.3	0.0	0.0	1.0	0.1	0.3	0.0	0.0	1.0	0.2	0.4	0.0
tit_wc	1.0	22.0	11.2	3.6	11.0	1.0	22.0	9.7	3.3	9.0	1.0	21.0	12.6	3.4	13.0
tit_pol	-1.3	0.9	-0.1	0.4	0.0	-1.2	0.9	0.0	0.4	0.0	-1.3	0.9	-0.1	0.3	0.0
tit_loud	0.0	0.5	0.2	0.1	0.2	0.0	0.5	0.1	0.1	0.1	0.0	0.5	0.2	0.1	0.2
tit_read	-3.4	43.8	8.2	3.9	8.0	-3.4	19.4	7.8	3.6	7.4	0.5	43.8	8.6	4.2	8.4
tit_has_cit	0.0	1.0	0.1	0.2	0.0	0.0	1.0	0.0	0.2	0.0	0.0	1.0	0.1	0.3	0.0
tit_has_que	0.0	1.0	0.0	0.2	0.0	0.0	1.0	0.0	0.2	0.0	0.0	1.0	0.0	0.2	0.0
img_brightness	26.0	233.0	96.7	33.8	95.0	26.0	233.0	94.5	37.0	90.5	27.0	195.0	98.9	30.1	97.0
img_has_face	0.0	1.0	0.7	0.5	1.0	0.0	1.0	0.7	0.4	1.0	0.0	1.0	0.7	0.5	1.0
p_love_pct	0.0	0.2	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
p_wow_pct	0.0	0.3	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0
p_haha_pct	0.0	0.6	0.1	0.1	0.0	0.0	0.6	0.1	0.1	0.0	0.0	0.4	0.1	0.1	0.0
p_sad_pct	0.0	0.4	0.0	0.0	0.0	0.0	0.4	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.0
p_angry_pct	0.0	0.7	0.2	0.2	0.1	0.0	0.7	0.1	0.2	0.1	0.0	0.7	0.2	0.2	0.1
p_shares_rel	2.5E-07	7.1E-02	7.1E-04	3.5E-03	8.7E-05	2.5E-07	6.3E-03	2.2E-04	6.4E-04	2.9E-05	2.9E-06	7.1E-02	1.2E-03	4.9E-03	2.1E-04
p_comments_rel	2.9E-07	2.7E-03	1.2E-04	2.3E-04	5.3E-05	2.9E-07	4.9E-04	8.2E-05	9.5E-05	5.1E-05	2.9E-07	2.7E-03	1.5E-04	3.1E-04	5.9E-05
c_tag_usrs_mean	0.0	0.7	0.0	0.1	0.0	0.0	0.7	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.0
c_like_rel_mean	0.0E+00	5.1E-05	1.7E-06	3.6E-06	9.5E-07	0.0E+00	5.1E-05	1.6E-06	3.6E-06	9.3E-07	0.0E+00	3.8E-05	1.8E-06	3.7E-06	9.7E-07
c_wc_mean	4.0	398.3	37.0	41.6	21.5	4.0	398.3	52.2	52.3	34.6	5.7	209.9	21.9	16.5	18.0
c_pol_mean	-0.5	0.7	-0.1	0.1	-0.1	-0.4	0.7	0.0	0.1	0.0	-0.5	0.4	-0.1	0.1	-0.1
c_loud_mean	0.0	0.5	0.1	0.0	0.1	0.0	0.5	0.1	0.1	0.1	0.0	0.3	0.1	0.0	0.1
c_read_mean	1.7	19.1	10.0	2.1	10.0	2.0	19.1	10.3	2.3	10.4	1.7	16.5	9.6	1.7	9.6
c_has_cite_mean	0.0	1.0	0.1	0.1	0.0	0.0	1.0	0.1	0.1	0.1	0.0	0.5	0.0	0.0	0.0
c_has_que_mean	0.0	1.0	0.2	0.1	0.1	0.0	0.7	0.2	0.1	0.2	0.0	1.0	0.1	0.1	0.1

Results

Table 4 presents the results of the logistic regression, clustered by variables concerning the information source and the social judgment and trained on the complete sample. Note that we conduct 10-fold cross validations and other model diagnostics afterwards to evaluate the models.

Table 4. Results of Logistic Regression (N=460 posts & 125,725 comments)							
				Estimate	Std. Error	Z- Value	P-Value
Information Source	Cognitive Cues	Message	mes_wc	-1.94E-02	1.26E-02	-1.547	0.12191
			mes_pol	3.43E-01	4.24E-01	0.809	0.41857
			mes_loud	-9.02E-02	1.03E+00	-0.087	0.93032
			mes_read	-4.55E-03	3.04E-02	-0.15	0.88087
			mes_has_cit	-1.03E+00	4.69E-01	-2.199	0.02785**
			mes_has_que	-1.65E-01	4.25E-01	-0.389	0.69763
		Title	tit_wc	8.33E-02	4.17E-02	1.998	0.04576**
			tit_polarity	-3.89E-01	4.20E-01	-0.926	0.35455
			tit_loudness	8.36E+00	1.90E+00	4.398	1.09E-05***
			tit_read	8.08E-02	3.34E-02	2.423	0.0154**
			tit_has_cit	5.41E-01	5.75E-01	0.941	0.34686
			tit_has_que	1.82E-01	6.66E-01	0.274	0.78435
	Visual Cues		img_brightness	3.14E-03	4.00E-03	0.785	0.43268
			img_has_face	3.16E-01	3.04E-01	1.039	0.29863
Social Judgement	Affective Cues		p_love_pct	-1.62E+01	5.69E+00	-2.845	0.00445***
			p_wow_pct	-5.05E+00	3.33E+00	-1.514	0.13008
			p_haha_pct	-2.67E+00	1.58E+00	-1.687	0.09157*
			p_sad_pct	-9.77E+00	3.41E+00	-2.871	0.0041***
			p_angry_pct	-2.68E-01	9.85E-01	-0.272	0.78525
	Behavioral Cues		p_shares_rel	3.76E+02	1.67E+02	2.256	0.02408**
			p_comments_rel	3.29E+02	1.12E+03	0.294	0.76897
			c_tag_usrs_mean	-1.09E+00	3.20E+00	-0.34	0.73368
			c_like_rel_mean	2.84E+04	4.27E+04	0.665	0.5061
	Cognitive Cues		c_wc_mean	-1.07E-02	6.81E-03	-1.569	0.11666
			c_pol_mean	-2.90E+00	1.25E+00	-2.324	0.02014**
			c_loud_mean	-3.78E+00	2.93E+00	-1.29	0.19708
			c_read_mean	-6.80E-02	6.73E-02	-1.01	0.31238
			c_has_cite_mean	-3.46E+00	2.05E+00	-1.693	0.09039*
			c_has_que_mean	-2.98E+00	1.47E+00	-2.031	0.0423**
Other			(Intercept)	-2.45E-01	1.28E+00	-0.191	0.84881

*Note: Dependent variable=fake news. * p <10%, ** p <5%, ***p <1%.*

Regarding the information source, four cognitive cues of the message and title of a Facebook post exhibit a statistically significant impact on the question whether a post contains fake news. First, messages containing a citation (*mes_has_cit*) reduce the possibility of fake news at the 5% significance level. Second, an increased word count (*tit_wc*), an increased loudness (*tit_wc*) as well as an increased readability *(tit_read*) of the title text is associated with an increased possibility of fake news. These effects are statistically significant at the 5%, 1% and 5% level respectively. Interestingly, neither the brightness *(img_brightness*) nor the question whether an image shows a face (img_has_face) of a Facebook post are associated with statistically significant effects on the question whether a post contains fake news.

Regarding the social judgment, three affective cues are statistically significant. Specifically, the relative number of loves *(p_love_pct),* haha *(p_haha_pct)* and sad *(p_haha_pct)* votes, which are all interpreted in relation to the left out reference category *p_like_pct* are associated with a decreased probability of fake news posts. These effects are statistically significant at the 1%, 10% and 1% level. Concerning the behavioral cues, an increased number of relative times a post was shared by Facebook users (*p_shares_rel*) is associated with a increased probability that the post in question contains fake news. This effect is statistically significant at the 5% level. Regarding the cognitive cues, three variables exhibit a statistically significant effect on the question whether a post contains fake news. The average polarity of the comments of a post (*c_pol_mean*) as well as the average number of comments that contain a citation (*c_has_cite_mean*) or a question (*c_has_que_mean*) all decrease the possibility that the associated post contains fake news. These effects are statistically significant at the 5%, 10% and 5% level.

Model Evaluation

Table 5 provides an overview on variable correlations as well as variance inflation factors (VIF). The unconditional associations among our variables represented by the Pearson product-moment correlation coefficient are observed to be moderate. Furthermore, looking at VIF scores reveal that our model is not subject to multicollinearity issues.

Table 5. Pearson Product Moment Correlations and Variance Inflation Factors

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	VIF
mes_wc	1																													1.39
mes_pol	2	.08																												1.27
mes_loud	3	-.28	-.13																											1.27
mes_read	4	.41	.00	-.13																										1.37
mes_has_cit	5	.27	.03	-.15	.17																									1.14
mes_has_que	6	-.05	.02	.00	-.17	-.02																								1.16
tit_wc	7	-.08	-.08	.16	-.23	-.16	.15																							1.30
tit_polarity	8	-.03	.29	.00	-.02	.00	.05	.00																						1.32
tit_loudness	9	-.23	-.14	.34	-.33	-.23	.17	.43	-.10																					1.47
tit_read	10	.00	.00	.07	.09	.02	.00	-.04	-.06	.01																				1.16
tit_has_cit	11	.00	.01	.02	-.12	.10	.09	.21	.07	.09	-.07																			1.14
tit_has_que	12	-.08	-.05	.08	.01	-.09	.08	-.01	-.04	.00	-.11	.09																		1.18
img_brightness	13	-.06	-.07	.14	-.01	-.04	-.03	.01	-.10	.02	.02	-.03	.00																	1.16
img_has_face	14	.07	.07	-.12	-.02	.10	.10	-.06	.09	-.08	-.04	-.01	-.03	-.25																1.24
p_shares_rel	15	.16	.01	.03	-.02	-.03	-.01	.08	.01	.07	-.02	-.01	-.02	-.03	.05															1.54
p_comments_rel	16	.04	-.06	.02	-.03	.00	.09	.19	.07	.05	.03	-.02	-.03	-.04	.07	.21														1.68
p_love_pct	17	.16	.17	-.13	.14	.15	-.06	-.20	.15	-.23	-.06	-.05	-.06	-.07	.14	.05	-.08													2.17
p_wow_pct	18	-.12	-.16	.18	-.03	-.09	-.05	.08	-.06	.09	.01	.04	.03	.10	-.07	.00	-.05	-.39												1.35
p_haha_pct	19	-.02	.16	-.04	-.07	.11	-.02	-.01	.17	-.08	-.05	.01	-.02	-.10	.18	-.06	.04	-.15	.01											1.59
p_sad_pct	20	-.03	-.09	-.05	.03	-.06	-.06	.03	-.15	.00	.05	-.01	-.04	.12	-.19	.00	.02	-.17	-.03	-.19										1.36
p_angry_pct	21	-.10	-.18	.06	-.04	-.10	.00	.15	-.06	.12	.04	.04	.02	.04	-.14	.02	.26	-.51	.18	-.20	.21									2.10
c_wc_mean	22	.21	.00	-.18	.27	.13	-.11	-.31	-.04	-.42	-.03	-.07	-.04	.02	.04	-.02	-.15	.14	-.10	-.06	.11	-.15								1.51
c_pol_mean	23	.06	.20	-.13	-.02	.02	-.09	-.17	.20	-.20	-.01	.03	-.09	-.03	.10	-.04	-.11	.33	-.11	-.04	-.15	-.33	.08							1.66
c_loud_mean	24	.00	.06	.05	.00	.01	.09	-.01	.11	.10	.04	-.03	.14	-.04	.04	-.07	.08	.05	-.13	-.06	.00	-.01	-.11	-.03						1.36
c_read_mean	25	.16	-.01	-.11	.09	.11	-.14	-.14	-.05	-.16	.00	-.01	-.12	-.07	-.05	.06	-.09	.10	-.08	.00	.13	.03	.30	-.02	-.29					1.27
c_has_cite_mean	26	.13	.02	-.13	.14	.14	-.04	-.18	.00	-.22	-.07	-.03	-.04	-.08	.07	.01	-.11	.05	.01	.09	.07	-.09	.48	.03	-.19	.21				1.22
c_has_que_mean	27	.11	-.08	-.14	.20	.14	-.13	-.23	-.02	-.31	-.01	-.06	.05	.02	.01	-.01	-.13	.04	.01	.01	.05	-.02	.39	-.16	-.15	.20	.23			1.36
c_tag_usrs_mean	28	-.08	.09	.06	-.01	.10	-.08	-.11	.04	-.08	-.08	.00	-.03	.05	-.03	.01	-.06	.25	-.04	-.02	-.09	-.17	-.03	.11	.06	.03	-.05	-.05		1.20
c_like_rel_mean	29	.06	-.02	-.04	.10	-.02	-.07	.03	.07	-.01	-.06	-.05	-.05	-.10	.07	.12	.00	-.03	.00	.18	-.01	.07	-.02	.02	-.16	.13	.16	.02	-.03	1.28

Indeed, looking at Table 6, which presents a multitude of evaluation metrics of different machine learning classifiers calculated via stratified 10-fold cross-validation, reveals that the LOG classifier exhibits a predictive accuracy of 76.74%. This significantly outperforms the expected accuracy of 50% of guessing in a balanced stratified sample. In addition to that, DTR and XGB yield even higher accuracy of 78.26% and 78.70% respectively. Furthermore, the best performing classifiers, SVM and RFO both yield a predictive accuracy of 80.87% and diverge only slightly in terms of the remaining performance metrics. Looking at the specificity, which indicates the amount of correctly classified negative examples as well as the sensitivity, which represents the number of correctly classified positive examples, reveals that our models yield especially high numbers of correctly classified positive examples of 88.26%. Thus, our models are especially well suited in detecting fake news.

Table 6. Evaluation Results of Machine Learning Classifiers (Metrics Based on Stratified 10-Fold Cross-Valuation)

Classifier	Accuracy	Error Rate	Specificity	Sensitivity	Precision	F1-Score
LOG	0.7674	0.2326	0.7174	0.8174	0.7469	0.7782
DTR	0.7826	0.2174	0.7217	0.8435	0.7569	0.7957
XGB	0.7870	0.2130	0.7391	0.8348	0.7651	0.7964
RFO	0.8087	0.1913	0.7522	0.8652	0.7797	0.8193
SVM	**0.8087**	**0.1913**	**0.7348**	**0.8826**	**0.7712**	**0.8218**

Notes: SVM=Support Vector Machine, LOG= Logistic Regression, DTR=Decision Tree, RFO=Random Forest, XGB= XGBoost

Discussion and Limitations

The goal of this paper was to protect users and support platform providers by developing a method to automatically detect fake news. We assume that knowing whether the received information is fake news will reduce the recipients' susceptibility to the misguiding content. Therefore, we draw on the ELM, research on social judgment and related IS research on UGC information quality to identify metrics for detecting fake news. By applying a machine learning approach to a balanced sample of fake and non-fake news posted on Facebook during the U.S. presidential election 2016, we were able to create a model that correctly classifies more than 80% of news with a recall rate of almost 90%. Considering that 75% of US adults are not able to identify fake news as such from the headline (Silverman and Singer-Vine, 2016), this can be considered a major support for users. However, it needs to be critically noted that this accuracy comes at a cost of relatively lower specificity. Thus, future research should incorporate alternate metrics to improve the prediction, for example, by considering affective cues relative to the sites overall likes. Future research could also consider more source-centric or news related attributes. Source-centric metrics such as the overall number of Facebook likes or whether it is verified on Facebook can affect the contributor's trustworthiness on social media (Zhiming Liu, Liu and Li, 2012). Furthermore, fake news sites could falsely suggest probity by selecting name, profile pictures and logos similar to reliable sources. Thus, respective source-centric attributes should be considered in future. In the present study, we only considered the most apparent features of the news post, which are probably most influential due to their exposed position. However, characteristics of the actual fake news text should prospectively also be assessed to determine its status as being real or fake news. Beyond these considerations, it needs to be noted that we also excluded some seemingly relevant metrics like the percentage of post likes and the overall number of reactions due to multicollinearity. However, other limiting aspects concern the generalizability of our findings. The news detection in the present work only revolves around political topics. While these are currently of the predominant public interest, fake news can also target other areas like science, sports or economics, which are not part of the study's sample. Nevertheless, as we do not consider any topic specific features (e.g. term frequencies), we are confident in the generalizability of our results. Furthermore, we only considered messages from Facebook, which are structurally and functionally distinct from other social media platforms. While Facebook represents the social media platform where most news are consumed (Gottfried and Shearer, 2016) other platforms are also subject to fake news, which need individual means of detection. Next to this limitation, it is possible that future advances in the realm of natural language generation could potentially bypass our detection system by incorporating our findings to create fake news which are indistinguishable from non-fake news.

Beyond these practical deliberations, we also need to critically assess the theoretical assumptions. Firstly, we cannot guarantee that knowing something is fake news makes users actually disregard the respective opinions. Prominent studies from Jones and Harris (1967) or Ross, Amabile and Steinmetz (1977) demonstrate that people neglect contextual information regarding a source of information – in this case whether something is labeled "fake news" – when assessing the information they provide. Therefore, our method might not be sufficient to make people fully insusceptible to fake news. However, we enable platforms to swiftly block potential fake news (sites). Furthermore, our model considers reactions from the community, which is also applied in related context (e.g., detection of hate speech) (Bretschneider and Peters, 2017). However, if people were informed about the probability of something rather being fake news, their reactions might change and thus affect the model calculation. Thus, future models should also consider predicting fake news without social judgement characteristics.

Conclusion

People increasingly rely on social media platforms as a news source. In a recent survey, 23% of the respondents indicate to use Facebook as their major- and 27% as their minor news source. According to the same survey, 75% of adult in the United States are unable to identify fake news (Silverman and Singer-Vine, 2016). Similar to traditional hoaxes and propaganda, fake news contain fabricated misinformation which are devoid of supportive facts and designed to mislead recipients. Unlike traditional hoaxes and propaganda, fake news shared on social media platforms might have a far greater impact because of its sheer speed, reach and personalization. Thus, fake news shared on social media platforms substantially transform society. For example by changing the political landscape as indicated by high-profile cases such as the United States presidential election 2016 (Mozur and Scott, 2016).

Because of the societal transformation induced by fake news and the difficulties people have when asked to identify them, our explorative study investigates *how to fully automatically identify fake news using information immediately apparent on social media platforms*. Specifically, building on the ELM and existing works in the realm of UGC and social psychology, we design an exploratory research model to study how cognitive, visual, affective and behavioral cues of a Facebook news posting as well as the associated comments allow for the prediction of fake news using machine learning classifiers.

Utilizing labeled ground-truth data covering human fact-checked fake and non-fake news articles of the U.S. presidential election 2016 shared by left-wing, right-wing and mainstream media outlets on Facebook, we are able to identify cues to reliably predict fake news fully automatically. Specifically, the best performing algorithmic approach achieves a predictive accuracy of more than 80%, and even more importantly, a recall rate of almost 90% (share of correctly classified fake news) in a stratified 10-fold cross-validation using a balanced sample.

Next to the automatic classification of non-fake and fake news, we provide insights on how to heuristically spot fake news. First, regarding the information source, we provide statistically significant evidence that posts whose message title contains a citation reduces the possibility of that post containing fake news, whereas an increase word count, an increased loudness as well as an increased readability of the posts title increases the possibility of fake news. Furthermore, we find no evidence of a predictive power of visual cues from the image attached to a posting. Second, regarding social judgment, we find that an increase of specific affective cues (love, haha and sad votes) relative to the likes, exhibit a statistically significant decrease on the probability that a posting contains fake news. Furthermore, regarding behavioral cues, an increased number of relative times a post was shared is associated with an increased probability that the post contains fake news. Additionally and focusing on cognitive cues, an increased average polarity of the comments of a post, an increased average number of comments with a citation and an increased number of comments containing a question significantly decrease the possibility of fake news.

Considering the alleged substantial effects of fake news on recent political events, the automatic detection of fake news has important practical consequences. For future research, the present study provides a starting point to identify other potentially relevant features in order to further improve the detection of fake news, which could also be expanded to other (nonpolitical) topics and tested using data from additional social media platforms. Current efforts of major platform operators to manually tag fake news could allow for such additional research in the near future.

References

Allcott, H. and M. Gentzkow. (2017). *Social Media and Fake News in the 2016 Election*. National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w23089

Ayeh, J. K. (2015). "Travellers' acceptance of consumer-generated media: An integrated model of technology acceptance and source credibility theories." *Computers in Human Behavior*, *48*, 173–180.

Boyd, D., S. Golder and G. Lotan. (2010). "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter." In: *43rd Hawaii International Conference on System Sciences (HICSS)* (pp. 1–10). IEEE.

Breiman, L. (2001). "Random forests." *Machine Learning*, *45*(1), 5–32.

Bretschneider, U. and R. Peters. (2017). "Detecting Offensive Statements towards Foreigners in Social Media." In: *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Bruns, A. and S. Stieglitz. (2014). "Metrics for understanding communication on Twitter." In: *Twitter and society* (Vol. 89, pp. 69–82). Peter Lang.

Cao, Q., W. Duan and Q. Gan. (2011). "Exploring determinants of voting for the 'helpfulness' of online user reviews: A text mining approach." *Decision Support Systems*, *50*(2), 511–521.

Carvalho, C., N. Klagge and E. Moench. (2011). "The persistent effects of a false news shock." *Journal of Empirical Finance*, *18*(4), 597–615.

Chen, T. and C. Guestrin. (2016). "XGBoost: A Scalable Tree Boosting System" (pp. 785–794). ACM Press.

Cheng, Y.-H. and H.-Y. Ho. (2015). "Social influence's impact on reader perceptions of online reviews." *Journal of Business Research*, *68*(4), 883–887.

Cortes, C. and V. Vapnik. (1995). "Support-vector networks." *Machine Learning*, *20*(3), 273–297.

Cox, D. (1958). "The regression analysis of binary sequences (with discussion)." *J Roy Stat Soc B*, *20*, 215–242.

DuBay, W. H. (2004). "The Principles of Readability." *Online Submission*.

Fang, B., Q. Ye, D. Kucukusta and R. Law. (2016). "Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics." *Tourism Management*, *52*, 498–506.

Ghose, A. and P. G. Ipeirotis. (2011). “Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics.” *IEEE Transactions on Knowledge and Data Engineering*, *23*(10), 1498–1512.

Gilovich, T., D. Keltner and R. E. Nisbett. (2010). *Social psychology* (1st ed.). New York: Norton & Compan.

Gottfried, J. and E. Shearer. (2016, May 26). “News Use Across Social Media Platforms 2016.”

Honey, C. and S. C. Herring. (2009). “Beyond microblogging: Conversation and collaboration via Twitter.” In: *System Sciences, 2009. HICSS’09. 42nd Hawaii International Conference on* (pp. 1–10). IEEE.

Hu, M. and B. Liu. (2004a). “Mining and summarizing customer reviews.” In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168–177). ACM.

Hu, M. and B. Liu. (2004b). “Mining opinion features in customer reviews.” In: *AAAI* (Vol. 4, pp. 755–760).

Hu, N., I. Bose, N. S. Koh and L. Liu. (2012). “Manipulation of online reviews: An analysis of ratings, readability, and sentiments.” *Decision Support Systems*, *52*(3), 674–684.

Jones, E. E. and V. A. Harris. (1967). “The attribution of attitudes.” *Journal of Experimental Social Psychology*, *3*(1), 1–24.

Kincaid, J. P., R. P. Fishburne Jr, R. L. Rogers and B. S. Chissom. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA006655

King, D. (2016, October 11). “Easily Create High Quality Object Detectors with Deep Learning.”

Korfiatis, N., E. García-Bariocanal and S. Sánchez-Alonso. (2012). “Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content.” *Electronic Commerce Research and Applications*, *11*(3), 205–217.

Kosinski, M., D. Stillwell and T. Graepel. (2013). “Private traits and attributes are predictable from digital records of human behavior.” *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805.

Kragh, M. and S. Åsberg. (2017). “Russia’s strategy for influence through public diplomacy and active measures: the Swedish case.” *Journal of Strategic Studies*, 1–44.

Krüger, N., S. Stieglitz and T. Potthoff. (2012). “Brand Communication In Twitter-A Case Study On Adidas.” In: *PACIS* (p. 161).

Liu, B., M. Hu and J. Cheng. (2005). “Opinion observer: analyzing and comparing opinions on the web.” In: *Proceedings of the 14th international conference on World Wide Web* (pp. 342–351). ACM.

Liu, Zhiming, L. Liu and H. Li. (2012). “Determinants of information retweeting in microblogging.” *Internet Research*, *22*(4), 443–466.

Liu, Zhiwei and S. Park. (2015). “What makes a useful online review? Implication for travel product websites.” *Tourism Management*, *47*, 140–151.

Mozur, P. and M. Scott. (2016, November 17). “Fake News in U.S. Election? Elsewhere, That’s Nothing New.” *The New York Times*.

Mudambi, S. M. and D. Schuff. (2010). “What makes a helpful review? A study of customer reviews on Amazon. com.” *MIS Quarterly*, *34*(1), 185–200.

Pan, Y. and J. Q. Zhang. (2011). “Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews.” *Journal of Retailing*, *87*(4), 598–612.

Park, S. and J. L. Nicolau. (2015). “Asymmetric effects of online consumer reviews.” *Annals of Tourism Research*, *50*, 67–83.

Petty, R. E. and J. T. Cacioppo. (1986). “The elaboration likelihood model of persuasion.” In: *Communication and persuasion* (pp. 1–24). Springer.

Qazi, A., K. B. Shah Syed, R. G. Raj, E. Cambria, M. Tahir and D. Alghazzawi. (2016). “A concept-level approach to the analysis of online review helpfulness.” *Computers in Human Behavior*, *58*, 75–81.

Quinlan, J. R. (1986). “Induction of decision trees.” *Machine Learning*, *1*(1), 81–106.

Risius, M., F. Akolk and R. Beck. (2015). “Differential Emotions and the Stock Market-The Case of Company-Specific Trading.” In: *European Conference on Information Systems*.

Risius, M. and R. Beck. (2015). “Effectiveness of corporate social media activities in increasing relational outcomes.” *Information & Management*, *52*(7), 824–839.

Ross, L. D., T. M. Amabile and J. L. Steinmetz. (1977). “Social roles, social control, and biases in social-perception processes.” *Journal of Personality and Social Psychology*, *35*(7), 485–494.

Salehan, M. and D. J. Kim. (2016). "Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics." *Decision Support Systems*, *81*, 30–40.

Seebach, C. (2012). "Searching for Answers–Knowledge Exchange through Social Media in Organizations." In: *45th Hawaii International Conference on System Science (HICSS)* (pp. 3908–3917). IEEE.

Shan, Y. (2016). "How credible are online product reviews? The effects of self-generated and system-generated cues on source credibility evaluation." *Computers in Human Behavior*, *55*, 633–641.

Siering, M., K. Zimmermann and M. Haferkorn. (2014). "Read This! How to Boost the Interest towards Research Articles-A Study on SSRN Research Impact."

Silverman, C. and J. Singer-Vine. (2016, December 7). "Most Americans Who See Fake News Believe It, New Survey Says." Retrieved from https://www.buzzfeed.com/craigsilverman/fake-news-survey

Silverman, C., L. Strapagie, H. Shaban, E. Hall and J. Singer-Vine. (2016). "Hyperpartisan Facebook Pages Are Publishing False And Misleading Information At An Alarming Rate." Retrieved from https://www.buzzfeed.com/craigsilverman/partisan-fb-pages-analysis

Singer-Vine, J. (2017). "BuzzFeedNews/2016-10-facebook-fact-check." Retrieved from https://github.com/BuzzFeedNews/2016-10-facebook-fact-check

Walker, S. H. and D. B. Duncan. (1967). "Estimation of the probability of an event as a function of several independent variables." *Biometrika*, *54*(1–2), 167–179.

Yin, D., S. Mitra and H. Zhang. (2016). "Research Note—When Do Consumers Value Positive vs. Negative Reviews? An Empirical Investigation of Confirmation Bias in Online Word of Mouth." *Information Systems Research*, *27*(1), 131–144.

Zhang, K. Z. K., S. J. Zhao, C. M. K. Cheung and M. K. O. Lee. (2014). "Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model." *Decision Support Systems*, *67*, 78–89.