Mapping (Dis-)Information Flow about the MH17 Plane Crash

Mareike Hartmann
Dep. of Computer Science
University of Copenhagen
Denmark
hartmann@di.ku.dk

Yevgeniy Golovchenko
Dep. of Political Science
University of Copenhagen
Denmark
yg@ifs.ku.dk

Isabelle Augenstein
Dep. of Computer Science
University of Copenhagen
Denmark
augenstein@di.ku.dk

Abstract

Digital media enables not only fast sharing of information, but also disinformation. One prominent case of an event leading to circulation of disinformation on social media is the MH17 plane crash. Studies analysing the spread of information about this event on Twitter have focused on small, manually annotated datasets, or used proxys for data annotation. In this work, we examine to what extent text classifiers can be used to label data for subsequent content analysis, in particular we focus on predicting pro-Russian and pro-Ukrainian Twitter content related to the MH17 plane crash. Even though we find that a neural classifier improves over a hashtag based baseline, labeling pro-Russian and pro-Ukrainian content with high precision remains a challenging problem. We provide an error analysis underlining the difficulty of the task and identify factors that might help improve classification in future work. Finally, we show how the classifier can facilitate the annotation task for human annotators.

1 Introduction

Digital media enables fast sharing of information, including various forms of false or deceptive information. Hence, besides bringing the obvious advantage of broadening information access for everyone, digital media can also be misused for campaigns that spread disinformation about specific events, or campaigns that are targeted at specific individuals or governments. Disinformation, in this case, refers to intentionally misleading content (Fallis, 2015).

A prominent case of a disinformation campaign are the efforts of the Russian government to control information during the Russia-Ukraine crisis (Pomerantsev and Weiss, 2014). One of the most important events during the crisis was the crash of Malaysian Airlines (MH17) flight on July 17, 2014. The plane crashed on its way from Amsterdam to Kuala Lumpur over Ukrainian territory, causing the death of 298 civilians. The event immediately led to the circulation of competing narratives about who was responsible for the crash (see Section 2), with the two most prominent narratives being that the plane was either shot down by the Ukrainian military, or by Russian separatists in Ukraine supported by the Russian government (Oates, 2016). The latter theory was confirmed by findings of an international investigation team. In this work, information that opposes these findings by promoting other theories about the crash is considered disinformation. When studying disinformation, however, it is important to acknowledge that our fact checkers (in this case the international investigation team) may be wrong, which is why we focus on both of the narratives in our study.

MH17 is a highly important case in the context of international relations, because the tragedy has not only increased Western, political pressure against Russia, but may also continue putting the government's global image at stake. In 2020, at least four individuals connected to the Russian separatist movement will face murder charges for their involvement in the MH17 crash (Harding, 2019), which is why one can expect the waves of disinformation about MH17 to continue spreading. The purpose of this work is to develop an approach that may help both practitioners and scholars of political science, international relations and political communication to detect and measure the scope of MH17-related disinformation.

Several studies analyse the framing of the crash and the spread of (dis)information about the event in terms of pro-Russian or pro-Ukrainian framing. These studies analyse information based on manually labeled content, such as television transcripts (Oates, 2016) or tweets (Golovchenko

et al., 2018; Hjorth and Adler-Nissen, 2019). Restricting the analysis to manually labeled content ensures a high quality of annotations, but prohibits analysis from being extended to the full amount of available data. Another widely used method for classifying misleading content is to use distant annotations, for example to classify a tweet based on the domain of a URL that is shared by the tweet, or a hashtag that is contained in the tweet (Guess et al., 2019; Gallacher et al., 2018; Grinberg et al., 2019). Often, this approach treats content from uncredible sources as misleading (e.g. misinformation, disinformation or fake news). This methods enables researchers to scale up the number of observations without having to evaluate the fact value of each piece of content from low-quality sources. However, the approach fails to address an important issue: Not all content from uncredible sources is necessarily misleading or false and not all content from credible sources is true. As often emphasized in the propaganda literature, established media outlets too are vulnerable to state-driven disinformation campaigns, even if they are regarded as credible sources (Jowett and O'donnell, 2014; Taylor, 2003; Chomsky and Herman, 1988)[1].

In order to scale annotations that go beyond metadata to larger datasets, Natural Language Processing (NLP) models can be used to automatically label text content. For example, several works developed classifiers for annotating text content with frame labels that can subsequently be used for large-scale content analysis (Boydstun et al., 2014; Tsur et al., 2015; Card et al., 2015; Johnson et al., 2017; Ji and Smith, 2017; Naderi and Hirst, 2017; Field et al., 2018; Hartmann et al., 2019). Similarly, automatically labeling attitudes expressed in text (Walker et al., 2012; Hasan and Ng, 2013; Augenstein et al., 2016; Zubiaga et al., 2018) can aid the analysis of disinformation and misinformation spread (Zubiaga et al., 2016). In this work, we examine to which extent such classifiers can be used to detect pro-Russian framing related to the MH17 crash, and to which extent classifier predictions can be relied on for analysing information flow on Twitter.

MH17 Related (Dis-)Information Flow on Twitter We focus our classification efforts on a Twitter dataset introduced in Golovchenko et al. (2018), that was collected to investigate the flow of MH17-related information on Twitter, focusing on the question who is distributing (dis-)information. In their analysis, the authors found that citizens are active distributors, which contradicts the widely adopted view that the information campaign is only driven by the state and that citizens do not have an active role.

To arrive at this conclusion, the authors manually labeled a subset of the tweets in the dataset with pro-Russian/pro-Ukrainian frames and build a retweet network, which has Twitter users as nodes and edges between two nodes if a retweet occurred between the two associated users. An edge was considered as *polarized* (either pro-Russian or pro-Ukrainian), if at least one retweet between the two users connected by the edge was pro-Russian/pro-Ukrainian. Then, the amount of polarized edges between users with different profiles (e.g. citizen, journalist, state organ) was computed.

Labeling more data via automatic classification (or computer-assisted annotation) of tweets could serve an analysis as the one presented in Golovchenko et al. (2018) in two ways. First, more edges could be labeled.[2] Second, edges could be labeled with higher precision, i.e. by taking more tweets comprised by the edge into account. For example, one could decide to only label an edge as polarized if at least half of the retweets between the users were pro-Ukrainian/pro-Russian.

Contributions We evaluate different classifiers that predict frames for unlabeled tweets in Golovchenko et al. (2018)'s dataset, in order to increase the number of polarized edges in the retweet network derived from the data. This is challenging due to a skewed data distribution and the small amount of training data for the pro-Russian class. We try to combat the data sparsity using a data augmentation approach, but have to report a negative result as we find that data augmentation in this particular case does not improve classification results. While our best neural classifier clearly outperforms a hashtag-based baseline, generating high quality predictions for the pro-

---

[1]The U.S. media coverage of weapons of mass destruction in Iraq stands as one of the most prominent examples of how generally credible sources can be exploited by state authorities.

[2]Only 26% of the available tweets in Golovchenko et al. (2018)'s dataset are manually labeled.

Russian class is difficult: In order to make predictions at a precision level of 80%, recall has to be decreased to 23%. Finally, we examine the applicability of the classifier for finding new polarized edges in a retweet network and show how, with manual filtering, the number of pro-Russian edges can be increased by 29%. We make our code, trained models and predictions publicly available[3].

2 Competing Narratives about the MH17 Crash

We briefly summarize the timeline around the crash of MH17 and some of the dominant narratives present in the dataset. On July 17, 2014, the MH17 flight crashed over Donetsk Oblast in Ukraine. The region was at that time part of an armed conflict between pro-Russian separatists and the Ukrainian military, one of the unrests following the Ukrainian revolution and the annexation of Crimea by the Russian government. The territory in which the plane fell down was controlled by pro-Russian separatists.

Right after the crash, two main narratives were propagated: Western media claimed that the plane was shot down by pro-Russian separatists, whereas the Russian government claimed that the Ukrainian military was responsible. Two organisations were tasked with investigating the causes of the crash, the Dutch Safety Board (DSB) and the Dutch-led joint investigation team (JIT). Their final reports were released in October 2015 and September 2016, respectively, and conclude that the plane had been shot down by a missile launched by a BUK surface-to-air system. The BUK was stationed in an area controlled by pro-Russian separatists when the missile was launched, and had been transported there from Russia and returned to Russia after the incident. These findings are denied by the Russian government until now. There are several other crash-related reports that are frequently mentioned throughout the dataset. One is a report by Almaz-Antey, the Russian company that manufactured the BUK, which rejects the DSB findings based on mismatch of technical evidence. Several reports backing up the Dutch findings were released by the investigative journalism website Bellingcat.[4]

The crash also sparked the circulation of several alternative theories, many of them promoted in Russian media (Oates, 2016), e.g. that the plane was downed by Ukrainian SU25 military jets, that the plane attack was meant to hit Putin's plane that was allegedly traveling the same route earlier that day, and that the bodies found in the plane had already been dead before the crash.

3 Dataset

For our classification experiments, we use the MH17 Twitter dataset introduced by Golovchenko et al. (2018), a dataset collected in order to study the flow of (dis)information about the MH17 plane crash on Twitter. It contains tweets collected based on keyword search[5] that were posted between July 17, 2014 (the day of the plane crash) and December 9, 2016.

Golovchenko et al. (2018) provide annotations for a subset of the English tweets contained in the dataset. A tweet is annotated with one of three classes that indicate the framing of the tweet with respect to responsibility for the plane crash. A tweet can either be *pro-Russian* (Ukrainian authorities, NATO or EU countries are explicitly or implicitly held responsible, or the tweet states that Russia is not responsible), *pro-Ukrainian* (the Russian Federation or Russian separatists in Ukraine are explicitly or implicitly held responsible, or the tweet states that Ukraine is not responsible) or *neutral* (neither Ukraine nor Russia or any others are blamed). Example tweets for each category can be found in Table 2. These examples illustrate that the framing annotations do not reflect general polarity, but polarity with respect to responsibility to the crash. For example, even though the last example in the table is in general pro-Ukrainian, as it displays the separatists in a bad light, the tweet does not focus on responsibility for the crash. Hence the it is labeled as neutral.

Table 1 shows the label distribution of the annotated portion of the data as well as the total amount of original tweets, and original tweets plus their retweets/duplicates in the network. A *retweet* is a repost of another user's original tweet, indicated by a specific syntax (RT @username:). We consider as *duplicate* a tweet with text that is identical to an original tweet after preprocessing (see Section 5.1). For our classification experiments,

---

[3]https://github.com/coastalcph/mh17
[4]https://www.bellingcat.com/

[5]These keywords were: MH17, Malazijskij [and] Boeing (in Russian), #MH17, #Pray4MH17, #PrayforMH17. The dataset was collected using the Twitter *Garden hose*, which means that it contains a 10% of all tweets within the specified period that matched the search criterion.

	Label	Original	All
Labeled	Pro-Russian	512	4,829
	Pro-Ukrainian	910	12,343
	Neutral	6,923	118,196
Unlabeled	-	192,003	377,679
Total	-	200,348	513,047

Table 1: Label distribution and dataset sizes. Tweets are considered *original* if their preprocessed text is unique. *All* tweets comprise original tweets, retweets and duplicates.

we exclusively consider original tweets, but model predictions can then be propagated to retweets and duplicates.

4 Classification Models

For our classification experiments, we compare three classifiers, a hashtag-based baseline, a logistic regression classifier and a convolutional neural network (CNN).

Hashtag-Based Baseline Hashtags are often used as a means to assess the content of a tweet (Efron, 2010; Godin et al., 2013; Dhingra et al., 2016). We identify hashtags indicative of a class in the annotated dataset using the pointwise mutual information (pmi) between a hashtag hs and a class c, which is defined as

$$\text{pmi}(hs, c) = \log\frac{\text{p}(hs, c)}{\text{p}(hs)\,\text{p}(c)} \tag{1}$$

We then predict the class for unseen tweets as the class that has the highest pmi score for the hashtags contained in the tweet. Tweets without hashtag (5% of the tweets in the development set) or with multiple hashtags leading to conflicting predictions (5% of the tweets in the development set) are labeled randomly. We refer to to this baseline as HS_PMI.

Logistic Regression Classifier As non-neural baseline we use a logistic regression model.[6] We compute input representations for tweets as the average over pre-trained word embedding vectors for all words in the tweet. We use fasttext embeddings (Bojanowski et al., 2017) that were pre-trained on Wikipedia.[7]

Convolutional Neural Network Classifier As neural classification model, we use a convolutional neural network (CNN) (Kim, 2014), which has previously shown good results for tweet classification (dos Santos and Gatti, 2014; Dhingra et al., 2016).[8] The model performs 1d convolutions over a sequence of word embeddings. We use the same pre-trained fasttext embeddings as for the logistic regression model. We use a model with one convolutional layer and a relu activation function, and one max pooling layer. The number of filters is 100 and the filter size is set to 4.

5 Experimental Setup

We evaluate the classification models using 10-fold cross validation, i.e. we produce 10 different datasplits by randomly sampling 60% of the data for training, 20% for development and 20% for testing. For each fold, we train each of the models described in Section 4 on the training set and measure performance on the test set. For the CNN and LOGREG models, we upsample the training examples such that each class has as many instances as the largest class (Neutral). The final reported scores are averages over the 10 splits.[9]

5.1 Tweet Preprocessing

Before embedding the tweets, we replace urls, retweet syntax (RT @user_name:) and @mentions (@user_name) by placeholders. We lowercase all text and tokenize sentences using the StandfordNLP pipeline (Qi et al., 2018). If a tweet contains multiple sentences, these are concatenated. Finally, we remove all tokens that contain non-alphanumeric symbols (except for dashes and hashtags) and strip the hashtags from each token, in order to increase the number of words that are represented by a pre-trained word embedding.

5.2 Evaluation Metrics

We report performance as F1-scores, which is the harmonic mean between precision and recall. As the class distribution is highly skewed and we

---

[6]As non-neural alternative, we also experimented with SVMs. These showed inferior performance to the regression model.

[7]In particular, with cross-lingual experiments in mind (see Section 7), we used embeddings that are pre-aligned

between languages available here https://fasttext.cc/docs/en/aligned-vectors.html

[8]We also ran intitial experiments with recurrent neural networks (RNNs), but found that results were comparable with those achieved by the CNN architecture, which runs considerably faster.

[9]We train with the same hyperparameters on all splits, these hyperparameters were chosen according to the best macro f score averaged over 3 runs with different random seeds on *one* of the splits.

Label	Example tweet
Pro-Ukrainian	Video - Missile that downed MH17 'was brought in from Russia' @peterlane5news RT @mashable: Ukraine: Audio recordings show pro-Russian rebels tried to hide #MH17 black boxes. Russia Calls For New Probe Into MH17 Crash. Russia needs to say, ok we fucked up.. Rather than play games @IamMH17 STOP LYING! You have ZERO PROOF to falsely blame UKR for #MH17 atrocity. You will need to apologize.
Pro-Russian	Why the USA and Ukraine, NOT Russia, were probably behind the shooting down of flight #MH17 RT @Bayard_1967: UKRAINE Eyewitness Confirm Military Jet Flew Besides MH17 Airliner: BBC ... RT @GrahamWP_UK: Just read through #MH17 @bellingcat report, what to say - written by frauds, believed by the gullible. Just that.
Neutral	#PrayForMH17 :(RT @deserto_fox: Russian terrorist stole wedding ring from dead passenger #MH17

Table 2: Example tweets for each of the three classes.

are mainly interested in accurately classifying the classes with low support (pro-Russian and pro-Ukrainian), we report macro-averages over the classes. In addition to F1-scores, we report the area under the precision-recall curve (AUC).[10] We compute an AUC score for each class by converting the classification task into a one-vs-all classification task.

6 Results

The results of our classification experiments are presented in Table 3. Figure 1 shows the per-class precision-recall curves for the LOGREG and CNN models as well as the confusion matrices between classes.[11]

Comparison Between Models We observe that the hashtag baseline performs poorly and does not improve over the random baseline. The CNN classifier outperforms the baselines as well as the LOGREG model. It shows the highest improvement over the LOGREG for the pro-Russian class. Looking at the confusion matrices, we observe that for the LOGREG model, the fraction of True Positives is equal between the pro-Russian and the pro-Ukrainian class. The CNN model produces a higher amount of correct predictions for the pro-Ukrainian than for the pro-Russian class. The absolute number of pro-Russian True Positives is lower for the CNN, but so is in return the amount of misclassifications between the pro-Russian and pro-Ukrainian class.

Per-Class Performance With respect to the per class performance, we observe a similar trend across models, which is that the models perform

best for the neutral class, whereas performance is lower for the pro-Ukrainian and pro-Russian classes. All models perform worst on the pro-Russian class, which might be due to the fact that it is the class with the fewest instances in the dataset.

Considering these results, we conclude that the CNN is the best performing model and also the classifier that best serves our goals, as we want to produce accurate predictions for the pro-Russian and pro-Ukrainian class without confusing between them. Even though the CNN can improve over the other models, the classification performance for the pro-Russian and pro-Ukrainian class is rather low. One obvious reason for this might be the small amount of training data, in particular for the pro-Russian class.

In the following, we briefly report a negative result on an attempt to combat the data sparseness with cross-lingual transfer. We then perform an error analysis on the CNN classifications to shed light on the difficulties of the task.

7 Data Augmentation Experiments using Cross-Lingual Transfer

The annotations in the MH17 dataset are highly imbalanced, with as few as 512 annotated examples for the pro-Russian class. As the annotated examples were sampled from the dataset at random, we assume that there are only few tweets with pro-Russian stance in the dataset. This observation is in line with studies that showed that the amount of disinformation on Twitter is in fact small (Guess et al., 2019; Grinberg et al., 2019). In order to find more pro-Russian training examples, we turn to a resource that we expect to contain large amounts of pro-Russian (dis)information. The *Elections integrity dataset*[12] was released by Twitter in 2018 and contains the tweets and ac-

---

[10] The AUC is computed according to the trapezoidal rule, as implemented in the sklearn package (Pedregosa et al., 2011)

[11] Both the precision-recall curves and the confusion matrices were computed by concatenating the test sets of all 10 datasplits

[12] https://about.twitter.com/en_us/values/elections-integrity.html#data

Model	Macro-avg		Pro-Russian		Pro-Ukrainian		Neutral	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC
RANDOM	0.25	-	0.10	-	0.16	-	0.47	-
HS_PMI	0.25	-	0.10	-	0.16	-	0.48	-
LOGREG	0.59	0.53	0.38	0.34	0.51	0.41	0.88	0.86
CNN	**0.69**	**0.71**	**0.55**	**0.57**	**0.59**	**0.60**	**0.93**	**0.94**

Table 3: Classification results on the English MH17 dataset measured as F1 and area under the precision-recall curve (AUC).

count information for 3,841 accounts that are believed to be Russian trolls financed by the Russian government. While most tweets posted after late 2014 are in English language and focus on topics around the US elections, the earlier tweets in the dataset are primarily in Russian language and focus on the Ukraine crisis (Howard et al., 2018). One feature of the dataset observed by Howard et al. (2018) is that several hashtags show high peakedness (Kelly et al., 2012), i.e. they are posted with high frequency but only during short intervals, while others are persistent during time.

We find two hashtags in the Elections integrity dataset with high peakedness that were exclusively posted within 2 days after the MH17 crash and that seem to be pro-Russian in the context of responsibility for the MH17 crash: #КиевСкажиПравду (*Kiew tell the truth*) and #Киевсбилбоинг (*Kiew made the plane go down*). We collect all tweets with these two hashtags, resulting in 9,809 Russian tweets that we try to use as additional training data for the pro-Russian class in the MH17 dataset. We experiment with cross-lingual transfer by embedding tweets via aligned English and Russian word embeddings.[13] However, so far results for the cross-lingual models do not improve over the CNN model trained on only English data. This might be due to the fact that the additional Russian tweets rather contain a general pro-Russian frame than specifically talking about the crash, but needs further investigation.

8 Error Analysis

In order to integrate automatically labeled examples into a network analysis that studies the flow of polarized information in the network, we need to produce high precision predictions for the pro-

---

[13]We use two sets of monolingual fasttext embeddings trained on Wikipedia (Bojanowski et al., 2017) that were aligned relying on a seed lexicon of 5000 words via the RC-SLS method (Joulin et al., 2018)

Russian and the pro-Ukrainian class. Polarized tweets that are incorrectly classified as neutral will hurt an analysis much less than neutral tweets that are erroneously classified as pro-Russian or pro-Ukrainian. However, the worst type of confusion is between the pro-Russian and pro-Ukrainian class. In order to gain insights into why these confusions happen, we manually inspect incorrectly predicted examples that are confused between the pro-Russian and pro-Ukrainian class. We analyse the misclassifications in the development set of all 10 runs, which results in 73 False Positives of pro-Ukrainian tweets being classified as pro-Russian (referred to as *pro-Russian False Positives*), and 88 False Positives of pro-Russian tweets being classified as pro-Ukrainian (referred to as *pro-Ukrainian False Positives*). We can identify three main cases for which the model produces an error:

1. the correct class can be directly inferred from the text content easily, even without background knowledge

2. the correct class can be inferred from the text content, given that event-specific knowledge is provided

3. the correct class can be inferred from the text content if the text is interpreted correctly

For the pro-Russian False Positives, we find that 42% of the errors are category I and II errors, respectively, and 15% of category III. For the pro-Ukrainian False Positives, we find 48% category I errors, 33% category II errors and and 13% category III errors. Table 4 presents examples for each of the error categories in both sets which we will discuss in the following.

Category I Errors Category I errors could easily be classified by humans following the annotation guidelines (see Section 3). One difficulty can
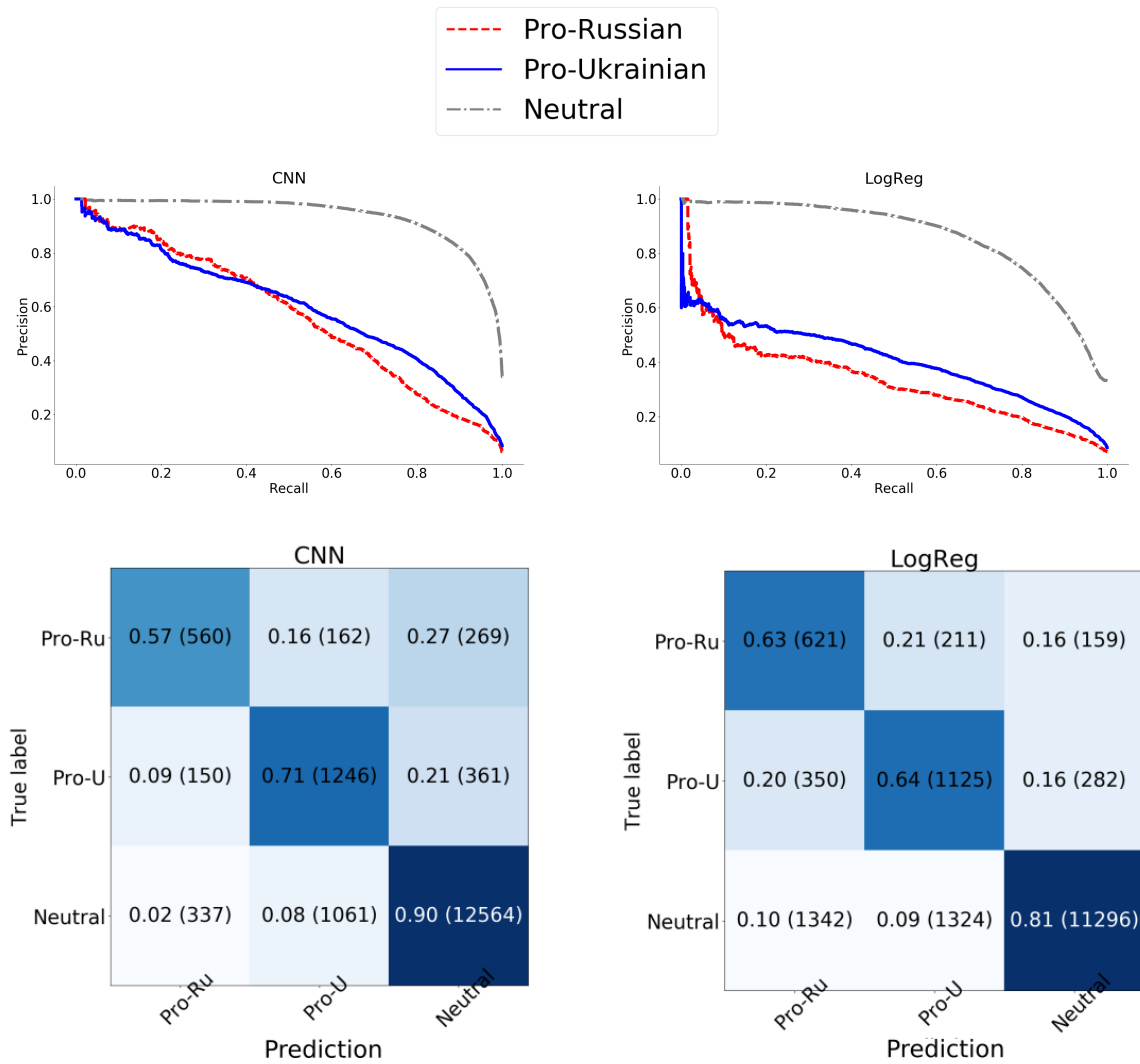
Figure 1: Confusion matrices for the CNN (left) and the logistic regression model (right). The y-axis shows the true label while the x-axis shows the model prediction.

be seen in example f). Even though no background knowledge is needed to interpret the content, interpretation is difficult because of the convoluted syntax of the tweet. For the other examples it is unclear why the model would have difficulties with classifying them.

Category II Errors Category II errors can only be classified with event-specific background knowledge. Examples g), i) and k) relate to the theory that a Ukrainian SU25 fighter jet shot down the plane in air. Correct interpretation of these tweets depends on knowledge about the SU25 fighter jet. In order to correctly interpret example j) as pro-Russian, it has to be known that the bellingcat report is pro-Ukrainian. Example l) relates to the theory that the shoot down was a false flag operation run by Western countries and the bodies in the plane were already dead before the

crash. In order to correctly interpret example m), the identity of *Kolomoisky* has to be known. He is an anti-separatist Ukrainian billionaire, hence his involvement points to the Ukrainian government being responsible for the crash.

Category III Errors Category III errors occur for examples that can only be classified by correctly interpreting the tweet authors' intention. Interpretation is difficult due to phenomena such as irony as in examples n) and o). While the irony is indicated in example n) through the use of the hashtag *#LOL*, there is no explicit indication in example o).
Interpretation of example q) is conditioned on world knowledge as well as the understanding of the speakers beliefs. Example r) is pro-Russian as it questions the validity of the assumption AC360 is making, but we only know that because we

Error cat.	True class	Model prediction	id	Tweet
I	Pro-U	Pro-R	a)	RT @ChadPergram: Hill intel sources say Russia has the capability to potentially shoot down a #MH17 but not Ukraine.
			b)	RT @C4ADS: .@bellingcat's new report says #Russia used fake evidence for #MH17 case to blame #Ukraine URL
			c)	The international investigation blames Russia for MH17 crash URL #KievReporter #MH17 #Russia #terror #Ukraine #news #war
	Pro-R	Pro-U	d)	RT @RT_com: BREAKING: No evidence of direct Russian link to #MH17 - US URL URL
			e)	RT @truthhonour: Yes Washington was behind Eukraine jets that shot down MH17 as pretext to conflict with Russia. No secrets there
			f)	Ukraine Media Falsely Claim Dutch Prosecutors Accuse Russia of Downing MH17: Dutch prosecutors de URL #MH17 #alert
II	Pro-U	Pro-R	g)	@Werteverwalter @Ian56789 @ClarkeMicah no SU-25 re #MH17 believer has ever been able to explain it,facts always get in their way
			h)	Rebel theories on #MH17 "total nonsense", Ukrainian Amb to U.S. Olexander Motsyk interviewed by @jaketapper via @cnn
			i)	Ukrainian Pres. says it's false "@cnnbrk: Russia says records indicate Ukrainian warplane was flying within 5 km of #MH17 on day of crash.
	Pro-R	Pro-U	j)	Russia has released some solid evidence to contradict @EliotHiggins + @bellingcat's #MH17 report. http://t.co/3leYfSoLJ3
			k)	RT @masamikuramoto: @MJoyce2244 The jets were seen by Russian military radar and Ukrainian eyewitnesses. #MH17 @Fossibilities @irina
			l)	RT @katehodal: Pro-Russia separatist says #MH17 bodies "weren't fresh" when found in Ukraine field,suggesting already dead b4takeoff
			m)	RT @NinaByzantina: #MH17 redux: 1) #Kolomoisky admits involvement URL 2) gets $1.8B of #Ukraine's bailout funds
III	Pro-U	Pro-R	n)	#Russia again claiming that #MH17 was shot down by air-to-air missile, which of course wasn't russian-made. #LOL URL
			o)	RT @20committee: New Moscow line is #MH17 was shot down by a Ukrainian fighter. With an LGBT pilot, no doubt.
	Pro-R	Pro-U	q)	RT @merahza: If you believe the pro Russia rebels shot #MH17 then you'll believe Justine Bieber is the next US President and that Coke is a
			q)	So what @AC360 is implying is that #US imposed sanctions on #Russia, so in turn they shot down a #Malaysia jet carrying #Dutch people? #MH17
			r)	RT @GrahamWP_UK: #MH17 1. A man on sofa watching YouTube thinks it was a 'separatist BUK'. 2. Man on site for over 25 hours doesn't.

Table 4: Examples for the different error categories. Error category I are cases where the correct class can easily be inferred from the text. For error category II, the correct class can be inferred from the text with event-specific knowledge. For error category III, it is necessary to resolve humour/satire in order to infer the intended meaning that the speaker wants to communicate.

know that the assumption is absurd. Example s) requires to evaluate that the speaker thinks people on site are trusted more than people at home.

From the error analysis, we conclude that category I errors need further investigation, as here the model makes mistakes on seemingly easy instances. This might be due to the model not being able to correctly represent Twitter specific language or unknown words, such as *Eukraine* in example e). Category II and III errors are harder to avoid and could be improved by applying reasoning (Wang and Cohen, 2015) or irony detection methods (Van Hee et al., 2018).

9 Integrating Automatic Predictions into the Retweet Network

Finally, we apply the CNN classifier to label new edges in Golovchenko et al. (2018)'s retweet network, which is shown in Figure 2. The retweet network is a graph that contains users as nodes and an edge between two users if the users are retweeting each other.[14] In order to track the flow of polarized information, Golovchenko et al. (2018) label an edge as polarized if at least one tweet contained in the edge was manually annotated as

---

[14]Golovchenko et al. (2018) use the k10 core of the network, which is the maximal subset of nodes and edges, such that all included nodes are connected to at least k other nodes (Seidman, 1983), i.e. all users in the network have interacted with at least 10 other users.
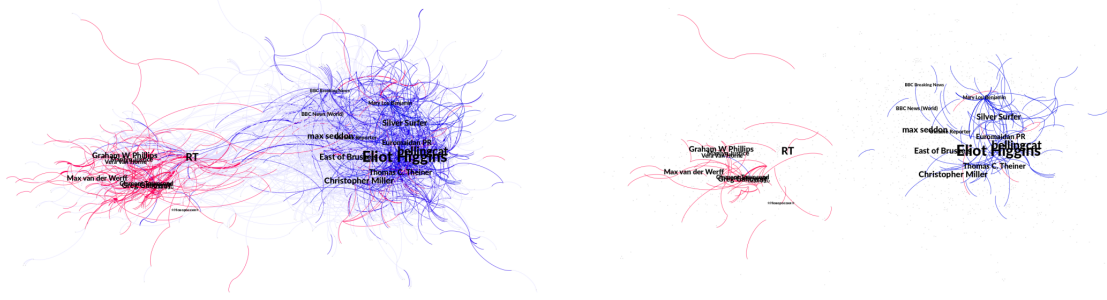
Figure 2: The left plot shows the original k10 retweet network as computed by Golovchenko et al. (2018) together with the new edges that were added after manually re-annotating the classifier predictions. The right plot only visualizes the new edges that we could add by filtering the classifier predictions. Pro-Russian edges are colored in red, pro-Ukrainian edges are colored in dark blue and neutral edges are colored in grey. Both plots were made using The Force Atlas 2 layout in gephi (Bastian et al., 2009).

pro-Russian or pro-Ukrainian. While the network shows a clear polarization, only a small subset of the edges present in the network are labeled (see Table 5).

Automatic polarity prediction of tweets can help the analysis in two ways. Either, we can label a previously unlabeled edge, or we can verify/confirm the manual labeling of an edge, by labeling additional tweets that are comprised in the edge.

9.1 Predicting Polarized Edges

In order to get high precision predictions for unlabeled tweets, we choose the probability thresholds for predicting a pro-Russian or pro-Ukrainian tweet such that the classifier would achieve 80% precision on the test splits (recall at this precision level is 23%). Table 5 shows the amount of polarized edges we can predict at this precision level. Upon manual inspection, we however find that the quality of predictions is lower than estimated. Hence, we manually re-annotate the pro-Russian and pro-Ukrainian predictions according to the official annotation guidelines used by (Golovchenko et al., 2018). This way, we can label 77 new pro-Russian edges by looking at 415 tweets, which means that 19% of the candidates are hits. For the pro-Ukrainian class, we can label 110 new edges by looking at 611 tweets (18% hits). Hence even though the quality of the classifier predictions is too low to be integrated into the network analysis right away, the classifier drastically facilitates the annotation process for human annotators compared to annotating unfiltered tweets (from the original labels we infer that for unfiltered tweets, only 6% are hits for the pro-Russian class, and 11% for the pro-Ukrainian class).

	Pro-R	Pro-U	Neutral	Total
# labeled edges in k10	270	678	2193	3141
# candidate edges	349	488	-	873
# added after filtering predictions	**77**	**110**	-	187

Table 5: Number of labeled edges in the k10 network before and after augmentation with predicted labels. Candidates are previously unlabeled edges for which the model makes a confident prediction. The total number of edges in the network is 24,602.

10 Conclusion

In this work, we investigated the usefulness of text classifiers to detect pro-Russian and pro-Ukrainian framing in tweets related to the MH17 crash, and to which extent classifier predictions can be relied on for producing high quality annotations. From our classification experiments, we conclude that the real-world applicability of text classifiers for labeling polarized tweets in a retweet network is restricted to pre-filtering tweets for manual annotation. However, if used as a filter, the classifier can significantly speed up the annotation process, making large-scale content analysis more feasible.

Acknowledgements

References

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of EMNLP*.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Amber E. Boydstun, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2014. Tracking the Development of Media Frames within and across Policy Issues. In *Proceedings of APSA*.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The Media Frames Corpus: Annotations of Frames Across Issues. In *Proceedings of ACL*, pages 438–444.

Noam Chomsky and Edward Herman. 1988. Manufacturing Consent New York. *Pantheon*.

Bhuwan Dhingra, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William Cohen. 2016. Tweet2Vec: Character-Based Distributed Representations for Social Media. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 269–274, Berlin, Germany. Association for Computational Linguistics.

Miles Efron. 2010. Hashtag retrieval in a microblogging environment. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788. ACM.

Don Fallis. 2015. What is disinformation? *Library Trends*, 63(3):401–426.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies. In *Proceedings of EMNLP*, pages 3570–3580. Association for Computational Linguistics.

John D. Gallacher, Vlad Barash, Philip N. Howard, and John Kelly. 2018. Junk news on military affairs and national security: Social media disinformation campaigns against us military personnel and veterans. *ArXiv*, abs/1802.03572.

Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. 2013. Using topic models for twitter hashtag recommendation. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 593–596. ACM.

Yevgeniy Golovchenko, Mareike Hartmann, and Rebecca Adler-Nissen. 2018. State, media and civil society in the information warfare over Ukraine: citizen curators of digital disinformation. *International Affairs*, 94(5):975–994.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire, and David Lazer. 2019. Fake news on twitter during the 2016 u.s. presidential election. *Science*, 363:374–378.

Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on facebook. In *Science advances*.

Like Harding. 2019. Three Russians and one Ukrainian to face MH17 murder charges. *The Guardian*.

Mareike Hartmann, Tallulah Jansen, Isabelle Augenstein, and Anders Søgaard. 2019. Issue Framing in Online Discussion Fora. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1401–1407.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance Classification of Ideological Debates: Data, Models, Features, and Constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan. Asian Federation of Natural Language Processing.

Frederik Hjorth and Rebecca Adler-Nissen. 2019. Ideological Asymmetry in the Reach of Pro-Russian Digital Disinformation to United States Audiences. *Journal of Communication*, 69(2):168–192.

Philip N Howard, Bharath Ganesh, Dimitra Liotsiou, John Kelly, and Camille François. 2018. *The IRA, social media and political polarization in the United States, 2012-2018*. University of Oxford.

Yangfeng Ji and Noah Smith. 2017. Neural Discourse Structure for Text Categorization. In *Proceedings of ACL*.

Kristen Johnson, Di Jin, and Dan Goldwasser. 2017. Leveraging Behavioral and Social Information for Weakly Supervised Collective Classification of Political Discourse on Twitter. In *Proceedings of ACL*.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Garth S. Jowett and Victoria O'donnell. 2014. *Propaganda & persuasion*. Sage.

John Kelly, Vladimir Barash, Karina Alexanyan, Bruce Etling, Robert Faris, Urs Gasser, and John G Palfrey. 2012. Mapping Russian Twitter. *Berkman Center Research Publication*, (2012-3).

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Nona Naderi and Graeme Hirst. 2017. Classifying Frames at the Sentence Level in News Articles. In *Proceedings of RANLP*, pages 536–542.

Sarah Oates. 2016. Russian media in the digital age: Propaganda rewired. *Russian Politics*, 1(4):398–417.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peter Pomerantsev and Michael Weiss. 2014. *The menace of unreality: How the Kremlin weaponizes information, culture and money*.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D Manning. 2018. Universal Dependency Parsing from Scratch. In *Proceedings of the {CoNLL} 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Cícero dos Santos and Maíra Gatti. 2014. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Stephen B. Seidman. 1983. Network structure and minimum degree. *Social Networks*, 5(3):269 – 287.

Philip M. Taylor. 2003. Munitions of the Mind. *A history of propaganda from the ancient world*.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns. In *Proceedings of ACL-IJCNLP*, pages 1629–1638. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012. Stance Classification using Dialogic Properties of Persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596, Montréal, Canada. Association for Computational Linguistics.

William Yang Wang and William W Cohen. 2015. Joint information extraction and reasoning: A scalable statistical relational learning approach. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 355–364.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Inf. Process. Manage.*, 54(2):273–290.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.