Misinformation on Twitter During the Danish National Election: A Case Study

Leon Derczynski
ITU Copenhagen
ld@itu.dk

Torben Oskar Albert-Lindqvist
ITU Copenhagen
toal@itu.dk

Marius Venø Bendsen
ITU Copenhagen
mvbe@itu.dk

Nanna Inie
Lix Technologies
ITU Copenhagen
nanna@lix.com

Jens Egholm Pedersen
ITU Copenhagen
jegp@itu.dk

Viktor Due Pedersen
ITU Copenhagen
vipe@itu.dk

Abstract

Elections are a time when communication is important in democracies, including over social media. This paper describes a case study of applying NLP to determine the extent to which misinformation and external manipulation were present on Twitter during a national election. We use three methods to detect the spread of misinformation: analysing unusual spatial and temporal behaviours; detecting known false claims and using these to estimate the total prevalence; and detecting amplifiers through language use. We find that while present, detectable spread of misinformation on Twitter was remarkably low during the election period in Denmark.

1 Introduction

Misinformation is a constantly changing phenomenon (Rojecki and Meraz, 2016; Southwell et al., 2018; Harley, 2017), and detecting its spread between individuals online can be difficult (Derczynski et al., 2015a). Meanwhile, it undermines our trust in information and can distort or damage our democratic and other governance structures. Misinformation can also be produced at high speed and in large volumes, making manual checking impossible.

Danish poses a particularly interesting challenge to automatic detection of misinformation because of the low availability of NLP resources for the language (Kirkedal et al., 2019). Taking the Danish national election 2019 as a case study, we ask the following research questions: *How might we use natural language processing (NLP) to automatically identify spread of misinformation during an election campaign on Twitter?* and, consecutively, *is the spread of misinformation on Twitter a threat to the Danish political discourse?*

Detecting structured propaganda campaigns is difficult for a number of reasons, and likely to remain so. Firstly, the adversary is constantly adapting and developing new strategies. This may look like a new network spreading pattern, a new posting strategy, etc. Automated methods for detecting misinformation are therefore likely to decrease in efficacy over time. Secondly, many tools for propaganda detection use natural language processing, which is well-developed for English, but not nearly as advanced for other languages, making groups using other languages easier to infiltrate undetected. Thirdly, better deep learning generation of text, image and video means better quality synthetic evidence to "support" false claims – and greater quantities of it. Finally, the lines are blurred and subjective. It is not always clear that an actor is malicious, or that a claim is outright false, or that a social media account is an amplifier. This makes it difficult to train machine learning approaches.

Due to this, the general approach taken in this research has been to use NLP to detect lists of candidate manipulative or propaganda-spreading accounts and then have a human evaluate the results. While keeping a human in the loop may increase the risk of reduced automatic performance, a human evaluator is vital to preventing the machine from becoming the final arbiter of good and bad content, or in this context, of truth.

2 Background

Systematic interference in social media discourse has often sought to divide (Stewart et al., 2018; Marwick and Lewis, 2017; Bastos and Mercea, 2019; Deb et al., 2019), be that through political hyperpartisanship or through intensifying existing opinions. In relatively constrained environments such as news fora, users are even often aware of these behaviours (Mihaylov et al., 2015).

One way to find structured online propaganda

is to find the amplifiers spreading information into the area of study (Wu et al., 2016). Misinformation is spread through sources and *amplifiers* (Weedon et al., 2017; Alaphilippe et al., 2019). Amplifiers target specific groups of people, and feed them stories that should match their interests. One example could be a Facebook group called "Brøndby Fans Secret News", that would target a football fan demographic using paid advertisements. The stories are often supplied by amplifiers by sharing, retweeting, or copying other content. This means the manipulation can be automated. Such a set-up works well because people within a special-interest group do not expect group information to be verifiable. Furthermore, groups with narrow interest tend to have more trust, partly due to the effect of homophily (Tang et al., 2013).

Some previous work has investigated social media for misinformation campaigns. Gorwa (2017) shortlisted 500 accounts for manual examination, providing a characterisation of both bot and manual Facebook activity. Kušen and Strembeck (2018) found, through sentiment and network analysis, that misinformation spread in the 2018 Austrian elections came mostly from followers of the eventual winner. Gorrell et al. (2019) investigated manipulation on a range of political issues and discussions. This included two UK votes, where there was mild support from Russian accounts in just two of many misleading claims, and found that low-effort strategies of retweeting and spreading generally pleasing content had low impact. Rather, serious manipulation is a long-term campaign, where the manipulator engages through deeply-felt issues in the target country and builds sympathy and a following based on those. A related study (Narayanan et al.) agreed and found that Russian-origin manipulators were not prevalent on YouTube or Instagram. We examine cases of those issues for Denmark in Section 5.1.

3 Data

Danish national elections are declared typically by the prime minister, which begins a three-week campaign period ended by a vote. Thus, they make a good subject for case study, as election-related content has an easily-identifable start date (and ends on election day).

Data was drawn from the publicly-accessible Twitter API. We constructed three datasets to model politically engaged content and users.

General dataset We sampled from the #dkpol hashtag, which is about politics in Denmark, and #fv19, which was dominated by discussions relevant to the Danish national election (Folketingsvalg 2019). By sampling of this hashtag we identified 7005 unique user accounts that had posted on one of these hashtags during the election period. We retrieved each account's 200 most-recent tweets, yielding 1.1 million tweets in all.

Party-supporting dataset To build a dataset of party-supporting accounts, we first looked at official Twitter accounts of each party (Section A), capturing two hundred most-recent tweets. We make the assume retweets indicate support (boyd et al., 2010); Non-endorsing retweets tend to be quotes including a comment, and ironic native retweets tend to be used by a narrow segment of users unless there is a long lag before retweet (Guerra et al., 2017). While anecdotal counter-examples can be found, the majority of retweets suggest not only interest but also trust in a message (Metaxas et al., 2015). For each account we collect the 200 most recent tweets, giving a dataset of 196,000 tweets from supporters and 2,300 tweets from official party accounts.

Candidate dataset After all candidates standing in the election are officially registered in the election, we used a manually-constructed list detailing Twitter accounts for all 900 candidates. Of these, 614 were on Twitter, and 362 active during the election.[1] This dataset contains up the (up to) 200 most recent tweets from each candidate's account.

This paper refers occasionally to "Danish Twitter"; this is considered to be the union of tweets sent in Denmark, tweets sent in Danish, and tweets from accounts based in Denmark. We cannot make assumptions about the nationality of Twitter account users, or their true location, hence the above definition. Captures were performed between 28 May 2019 and 6 June 2019.

The statement (Bender and Friedman, 2018) for the Party-supporting and Candidate data is:

- **Curation Rationale** As above
- **Language Variety** Colloquial Danish, `da`
- **Speaker Demographic** Politically engaged Danish-speakers; demographic unknown
- **Annotator Demographic** Males aged 25-40; mixture of Danish, Swedish and British

[1] https://twitter.com/runello/status/1136931663873810432

- **Speech Situation** Captured online text, analysed while current
- **Text Characteristics** Social media text

As it bears personal political opinions, the data is sensitive under GDPR and cannot be shared.

4 Methods

To find manipulation and outside influence, we used three techniques – analysing account activity, looking for spread of known misinformation, and seeing what languages accounts use.

In every instance, the goal is to find accounts and messages that will then be investigated manually. Given the seriousness of the subject matter at hand, and the damage potential in e.g. mistakenly accusing user accounts of spreading propaganda, or of mistakenly declaring a sphere clear of propaganda, some human intervention in machine conclusions is required at each point – especially given the relatively young age of this problem in the digital, and the rate at which adversaries may be capable of adapting to avoid misinformation detection techniques.

4.1 Checking for known misinformation

By searching for instances of known misinformation on Twitter, we can establish a lower bound for misinformative content.

Misinformation can be found using fact-verification resources. For English, one might use Snopes[2] or Politifact.[3] In the case of Danish, there is a set of known-false claims and stories, curated and analysed by a Danish news desk named TjekDet.[4] We took the set of false and misleading claims found over the past three months (from 1 March 2019 until 3 June 2019). This amounted to 38 false claims found spread through Danish media.[5] We searched for evidence of these in social media activity in our party-supporting and candidate datasets. To search, we used both keyword-based search, following a claim quoted in the TjekDet report when present, and also by finding tweets with FastText sentence vectors (Bojanowski et al., 2016) that were similar to the misinformative claim's. Each candidate instance of misinformation spread was checked manually.

[2]https://www.snopes.com/

[3]https://www.politifact.com/

[4]https://www.mm.dk/tjekdet

[5]From https://www.mm.dk/tjekdet/artikel.aspx?type=106

Party supported	P(misinformative|party)
SK	0.04%
DF	0.04%
Enhedslisten	0.01%
Others	n/a

Table 1: Probability of an account's messages being misinformation given support for a particular party. From the 200 most-recent tweets from supporters and candidates of each party, compared to a set of known misinformation stories. Some data is available for other parties but at levels levels too low to be informative.

Our results indicated a low minimum bound for misinformation on Danish Twitter. One in every 94,000 messages (roughly 0.001%) carried known misinformation. Many of the misinformative stories were not found at all in the data. While this is likely part due to Facebook being the dominant social media platform in Denmark, we would expect to see a higher degree of misinformative/propaganda stories being shared on Twitter. For comparison, in the US 2016 election, Facebook sharing of mainstream news and of misinformation reached roughly equal levels (Silverman, 2016). Facebook is too difficult to monitor automatically during emerging or mid-scale events, such as the case study in this paper; this is because of Facebook's restrictions, unless some kind of research agreement is already in place.

Using our party-supporting dataset (Section 3) also shows which party's supporters spread the most misinformation. Results are in Table 1. Figures are again generally low. The most significant rumours here were around police supply of fuel for a Koran burning[6] and about crime rising, especially in connection with foreigners.[7]

Based on these counts from our sample, it should be possible to estimate the amount of known misinformation on Twitter during the election period. The unobserved part can be estimated using smoothing methods, below.

Estimate Total Relevant Traffic Our data is a subsample of Twitter posts during the election. To accurately scale estimates of misinformation we need to estimate the total number of messages. The target is the number of messages on #dkpol from 7 May to 5 June 2019. We captured 1,046 of 1,788 hours in the voting period, yielding

[6]https://www.mm.dk/tjekdet/artikel/koeber-politiet-taendvaeske-til-paludans-koranafbraending

[7]https://www.mm.dk/tjekdet/artikel/kriminaliteten-bliver-vaerre

146,944 tweets. Our capture never hit Twitter's 50-messages-per-second threshold. As the dataset covers 59.2% of the time period, we estimate the total at 250K tweets.

As we will use data from user timelines to estimate misinformation spread, there are two potential sources of inaccuracy: firstly, we will assume that misinformation spread is the same before and during an election. However, we are calculating a lower bound. Second, we assume that misinformation will be shared on #dkpol; while this is not guaranteed, the skew shown in the TjekDet source is political, and it is fair to assume that misinformation during elections will be political in nature.

Discounting for Unobserved Events We apply Good-Turing discounting (Good, 1953) to estimate the scale of the unobserved misinformation. Good-Turing considers that there are a distinct number of categories, of which X have been observed; that a frequency vector $\bar{C}$ holds observation counts for each item $x \in X$, denoted C_x; and that the frequency of frequencies vector N_c holding how many times the frequency count c occurs in C. The goal is then to estimate the size of the unobserved mass, i.e. the number of events with observed frequency 0, N_0.

$$N_c = \sum_{x:count(x)=c} 1 \tag{1}$$

The total number of objects observed is:

$$N = \sum_{c=1}^{\infty} cN_c \tag{2}$$

The MLE count for N_c is c; under Good-Turing, this becomes the smoothed count, c^*:

$$c^* = (c+1)\frac{N_{c+1}}{N_c} \tag{3}$$

Thus, the extent of unobserved events can be estimated with N_0.

Determine Categories We consider two sets of categories: the first is groups of political supporters – the second, misinformative stories.

Estimating $N_0(supp)$ represents the amount of unobserved misinformation spreading by party supporters. The categories are parties.

Estimating $N_0(stories)$ represents the amount of unobserved misinformation spread by party supporters, where the categories are individual misinformative claims.

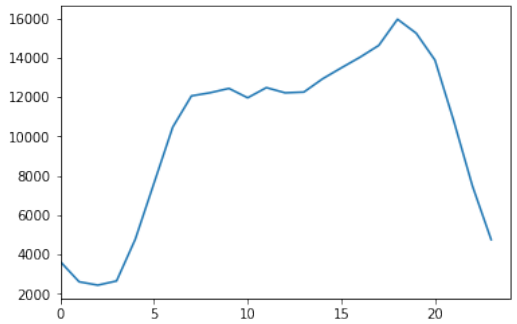


Figure 1: Time of day that politically-active users post on Danish Twitter.

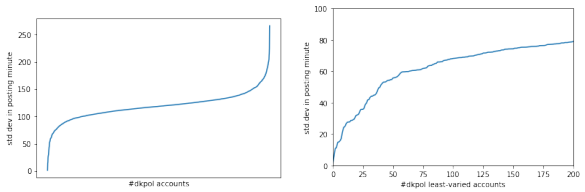
$N_0(stories)$ is a little more interesting, as Good-Turing is agnostic to what the actual categories mean, and so this term estimates the spread of any set of stories even those not already included by TjekDet. $N_{>1}$ counts for these unseen misinformative stories are missing from the total counts, and so it will be an underestimate.

Re-Estimate Lower Bound Based on these, we re-calculate the amount of known misinformation on Danish Twitter during the election, first by party. The data is sparse with low counts; the Good-Turing estimates of missed party shares come out at under one missed post ($p_{GT} * (unseen) = 0.03$), as do the Good-Turing estimates of missed stories ($p_{GT} * (unseen) = 0.07$), even when corrected for dataset size. Both point to a small lower bound for misinformation on Danish Twitter during the election.

4.2 Temporal Activity Patterns

People publish and send messages at the time that their target audience is active on the platform and likely to pay attention. Those that push out a burst of tweets rapidly in a short period of time, or that post uniformly throughout the day, can be suspicious. For example, a human will post sporadically and irregularly, and over time, is likely to send messages outside of their usual window, or send messages faster than they can be composed and typed. In contrast, we expect that an automatically-run account (e.g. a bot, a favoured tool of disinformers) will stick to rules, and that in general the rules will not implement a sophisticated timing regime. Exploiting timestamp-based information has previously helped identify bots.[8]

[8] https://www.indy100.com/article/brexit-party-nigel-farage-twitter-following-behaviour-fake-genuine-accounts-8920681



(a) Overall graph in order of user posting time variation

(b) Lowest temporal standard deviation accounts

Figure 2: Standard deviation in posting minute for #dkpol posters. Note small volume of low-σ posters.

We examined the following temporal facets of a user's timeline (i.e. 200 most-recent tweets):

- Activity at strange hours, e.g. 00.00 – 05.00, suggests overseas or automatic posting;
- The standard deviation σ of the minute-of-day that the posts are at; low variance can indicate one form of automatic activity
- Users who post in bursts, having the majority of their most-recent tweets be very close to the collection time.

Figure 1 shows diurnal activity in the general dataset. Overall, 1,215 of 7,005 users had been so active in the election period that their entire 200-tweet historical sample lay within it; the remainder sent fewer than 200 tweets during the election.

First, we looked at accounts posting at unusual times, namely 0am-5am (the least-active period). The filter was applied when over half the user's tweets fell in this window. There were only a handful, about 0.2% of all users. Most of these were in the USA or had US politics as their dominant topic; these mostly worked in English but a few tweeted in other languages. Being in the US would explain unusual (for Denmark) activity times. On manual inspection of this small volume, all accounts looked legitimate.

Another way of finding suspicious accounts is to see how irregular their posting pattern is. We measured deviation σ over the minute of day that each account posted at. Low variance indicates an automated posting pattern. In our data, 1% of users (73 in the sample of 7005) had a $\sigma < 60$ (one hour) indicating a tightly regular posting pattern. Results are shown in Figure 2.

The most unusual account is @folketingeterd, with posting minute $\sigma = 1.49$ (compare with the mean standard deviation of 119). This is an online version of Jakob Jakobsen's artwork containing statements that a politician has died (Figure 3).

Figure 3: Tweets from @folketingerd

This account *is* automated, and so the bot detection method has worked, finding a non-malicious bot.

Some low-σ accounts behaved like amplifiers, having a high degree of political content, a high proportion of retweeting, and also appearing in our time-of-day and posting intensity analyses below. However, none operated in Danish. Overall, looking at the 100 lowest-variance accounts that did not state they were from an organization, we found various unusual accounts: pro-Iran anti-Putin; Japanese cartoon pornography; automated general interest content retweeting; combined Greek and Turkish adverts; dedicated pro-Trump videos and news; and pro-Iran content in Farsi. Few accounts were run in Danish.

One account showing unusual temporal activity is that of Lars Løkke Rasmussen, prime minister at the time of the election. This account is most active late on Sunday nights (Figure 4) – perhaps to make sure that people who see interesting things there will discuss them at work the next day.

Finally we selected accounts that posted over 50 messages a day (or over 200 in a four-day window). This high temporal intensity sample comprised 47 accounts, 0.67% of the sample, many of which exhibited suspicious activity. Some were also present in the low-temporal-deviation list.

4.3 Spatial Account Activity

One way of finding an account posting from another country is to check the locations messages

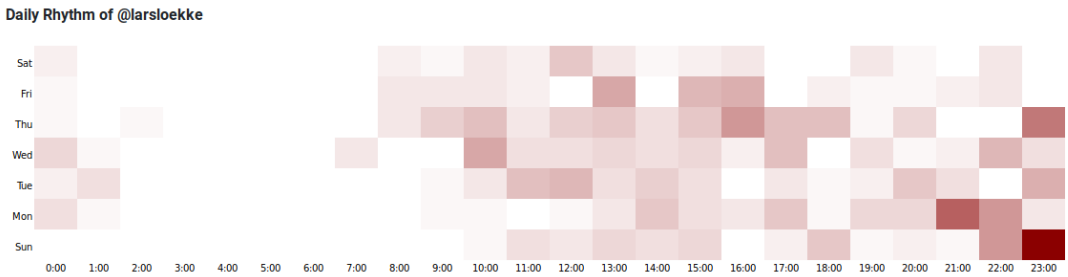


Figure 4: Tweets from the Danish prime minister's account (Lars Løkke Rasmussen). Note specific targeting of Sunday evenings. Via https://accountanalysis.lucahammer.com/

are tagged with. Criminals have in the past been caught out by letting their real location slip out (Lee, 2013), or so-called "location leaking". We will use this information source to build a short-list of potential disinformers and manipulators. For example, an automated account may be set to post everything from Copenhagen claiming to be using an iPhone, but the account's operator may post from their real location and forget to set "Copenhagen" when e.g. replying to a comment.

We monitor account locations using user-supplied locations and, when present, GPS. The target pattern is where most of the tweets have one location, but a small number have somewhere quite different. For example, where e.g. 99% of posts are labelled "Copenhagen" and 1% are from e.g. Moldova, where the account operator has stepped in to manually place a post.

User-supplied locations were sometimes unreliable, and will contain fictitious locations such as "mitten im Leben", "the void", or "Planet Earth", matching expectations from earlier studies (Hecht et al., 2011). The fact that these non-locations are present is not problematic: the goal of this exercise is to detect anomalous location behaviour.

- Copenhagen (481)
- København, Danmark (427)
- Denmark (395)
- Copenhagen, Denmark (342)
- Danmark (309)
- København (203)
- Aarhus, Danmark (85)
- København, Hovedstaden (68)
- Hovedstaden, Danmark (59)
- Odense, Danmark (52)
- Aarhus, Denmark (46)
- Frederiksberg (41)
- Aalborg, Danmark (40)
- Frederiksberg, Danmark (40)
- Aalborg (36)
- Odense (36)
- Aarhus (36)
- Arhus (35)
- Midtjylland, Danmark (17)
- Sjælland, Danmark (17)
- Odense, Denmark (16)
- Arhus, Denmark (15)
- Brussels, Belgium (15)
- Syddanmark, Danmark (15)
- Fredericia, Danmark (15)

We see a reasonably clean set of data. The top 22 locations are all within Denmark, and only three locations in the top 50 are outside: Brussels, London, and Europe. This makes an ostensibly homogeneous, Danish dataset.

There are many accounts that post almost always from "Denmark" but sometimes from a city in Denmark (e.g. 199 tweets from "Danmark" and one from "Aarhus"). This pattern of mostly cities in one country and a few mentions of another city in the same country continues in cases of foreign locations. There are a few cases of accounts that post 99% from Denmark and 1% from e.g. Chiang Mai or Mauritius. We inspected 90 accounts that had a regular posting location with $< 2\%$ of the tweets from unusual locations. All these appeared to be people on holiday or making occasional trips; unusually-located content often were about the location, e.g. a user checking in to a new place and drawing attention to it.

4.4 Multi-language Accounts

A further way of finding foreign interference is the see what other languages people active on #dkpol use. This can reveal accounts that primarily act in foreign languages. An automatic language detection tool called langid.py (Lui and Baldwin, 2012), which performs well on social media text in general (Derczynski et al., 2015b), does this.

Over the same users, we are interested in those who use a non-Danish language more than seldom (in more than 20 of their posts). Out of the 7,005

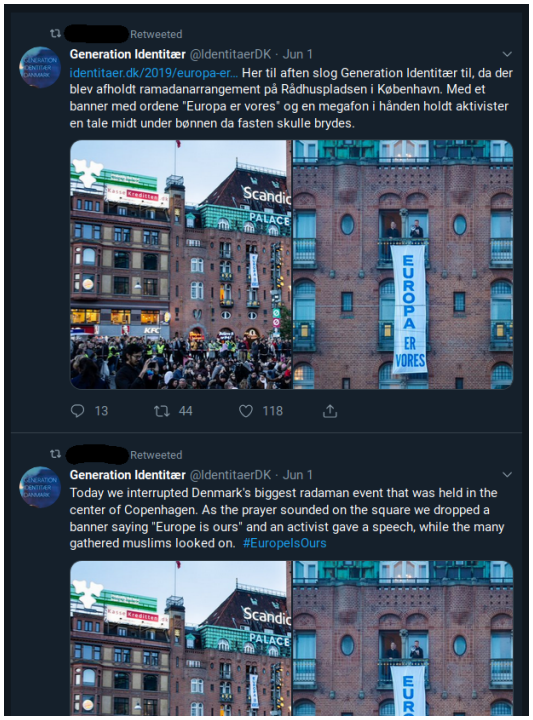


Figure 5: Amplification of a bilingual Danish/English propaganda source on #dkpol

users, 3,126 are active in English, 115 also use Swedish, 65 German, 50 Spanish, and 48 Farsi. Other languages exist but at low counts.

A higher threshold ($> 50\%$ in non-Danish) indicates accounts that are primarily multilingual or non-Danish. This matches 1,806 accounts (25.8%), the primary language of which is overwhelmingly English (1,483), or Swedish (77), German (38), French (30), Italian (28), Faroese (26) or Spanish (25). The dominance of English may be partly attributed to Denmark being small and so having nationals who also use English to reach the broader world. Overall this does not remove foreign manipulation through brigading, which is when people organise themselves ad-hoc to influence an online discussion using their personal accounts. Random examination of accounts did not reveal much active #dkpol interaction. Of course this does not rule out domestic brigading, which would not be detected through language ID.

There is some unusual, manipulator-like activity in the English tweets. Some accounts push a political point of view by retweeting very organised sources. These sources use multiple languages. We see bilingual activity, with accounts that use #dkpol spreading news in either both English and Danish, English and Swedish, or English and German (e.g. Figure 5). Social media manipulation is well-organised in Britain, Sweden, and Germany (Howard and Kollanyi, 2016; Gorrell et al., 2018; Bradshaw and Howard, 2018; Nygren et al., 2019; Neudert, 2017), countries having languages that Danes are likely to understand.

Propaganda narrative
Arabic teaching in school
Arabic writing on election poster
Syrians propagating
Muslims support a left-wing minister
Muslims take over foreign political party
Muslims take over domestic political party
Muslims cause crime abroad
We cannot build a country together with muslims
There have been attacks on nationalists abroad
Foreigners are committing crimes
Foreigners are expensive to integrate
Foreigners are draining pension funds
Anti semitism
US hyperpartisanship
There will be major pension cuts
Foreigners are draining pension funds
Police are supporting extremists
Police are ignoring hate crimes in the name of jihad
5G is very harmful to humans

Table 2: Groups of topics discussed by misinformation spreaders engaged on #dkpol

We filtered for accounts that had been active on #dkpol and used Danish in under half their tweets, finding 1806 accounts. From a sample, about 30% appeared to be mixed-language accounts working as amplifiers. They retweet stories selecting those that have an anti-EU, anti-feminism, pro-Trump agenda. They generally retweet the same story in multiple languages and stick to a small number of sources, including cartoons (often the same one across all accounts) and shock stories. Some of these accounts were also found in the high temporal intensity set described above.

From this, it seems that about 4-5% of #dkpol-active accounts sometime engage in propaganda spread. However, they do not usually do it in Danish. This corresponds to a low-effort campaign which as Gorrell et al. (2019) note typically has little effect on views.

5 Discussion

In general, we found that Danish Twitter is relatively free from misinformation – more so than Twitter in e.g. the UK, even controlling for size.

5.1 Propaganda topics: Islam and pensions

We segment the misinformation propagated on Twitter into topics (Table 2), taken from the 10

most recent posts by 50 suspicious accounts. As accounts tweet at different rates, directly comparing volumes between accounts did not make sense.

Anti-Islam material is by far the most dominant. Material about problems with foreigners also feature, which matches findings around typical party discussion topics (Derczynski et al., 2019). Pensions make a cursory appearance, perhaps playing on the finding that over-65s are more susceptible to false news (Guess et al., 2019).[9]

5.2 Limitations

The current case study is limited to Twitter. This is primarily due to Facebook's restrictions on data access for research, compared to Twitter's.

Facebook is essentially a black box of information when it comes to monitoring and data collection in emergencies; on the other hand, it offers user privacy by default. Facebook is the dominant platform in Denmark, with over 80% of residents over age 17 using Facebook every single day (Runge, 2019) – Twitter is used by fewer than 5% of residents.[10] Other social media platforms like Instagram and Snapchat, and local sites like Hestenettet and Slyngebarn (a forum that is like Danish Mumsnet), do not make up nearly as much of Danish social media consumption as Facebook. This means Denmark must rely on Facebook, itself a huge foreign corporation that is difficult to regulate (Vaidhyanathan, 2019), to manage misinformation detection and to spot domestic manipulation. We know, however, that Danish Twitter and Danish Facebook share roughly similar political views (Derczynski et al., 2019).

6 Conclusion

We examined misinformation on Danish Twitter during the 2019 national elections, analysing over 1.5 million tweets over the three-week process. A lower bound was established, and various behavioural and content-based techniques applied to help find suspicious accounts and activity. Twitter bore less misinformation than expected, and while signs of low-effort manipulation were present, the manipulation was not customised to Denmark.

So, to answer the research question posed in the introduction: there is mild evidence of some manipulation on Danish Twitter, and low evidence for misinformation. The easiest to detect is from cross-border accounts working in multiple languages, sometimes not using Danish at all. It remains hard to precisely locate sources of misinformation and manipulation, but it appears that they are either unmotivated foreign actors, or – perhaps more likely, given the content of the known misinformation – native-speaking domestic actors.

Despite this, one should not paint a rosy picture of social media in Denmark. We have determined in some cases only lower bounds, i.e. minimum levels, of misinformation spread. We know that propaganda targets Denmark (even if the Russian ambassador declared that meddling in Danish politics "makes no sense" [11]).

We believe that protecting Denmark from misinformation is best achieved through understanding the threat better, in three ways: through modelling relevant misinformation networks in detail; with better NLP tools for Danish; and through getting better access to Facebook, the dominant platform here. We discovered efforts to disturb Danish political discourse; these should not be ignored, but instead pursued and used to develop defences.

Synthetic propaganda is unlikely to perturb the Danish web until someone invests in training models like Grover (Zellers et al., 2019) and GPT-2 (Radford et al., 2019) for Danish. Given the current misinformation effort here, which one would expect to be most visible during election times, this is not a current major threat for Denmark.

On the other hand, Facebook's policies make it hard for others to develop social media propaganda defences. There are few tools for defending against misinformation on Facebook. Also, most misinformation research is on English; there are few tools for other languages. Denmark is Facebook-heavy and Danish-speaking. So, while propaganda continues to target Denmark, there is simultaneously little defence available.

Acknowledgments

We thank the reviewers for their comments. Leon Derczynski is part of the EU Center of Excellence for Research in Social Media and Information Disorder (EU REMID) research network. This research was partly supported from a project funded by TjekDet at Mandag Morgen.[12]

[9] An important counterpoint to this finding is that younger internet users may be better at spotting badly-designed pages but are in fact worse at discriminating between well-written true and manipulative content (Nygren and Guath, 2019).

[10] http://gs.statcounter.com/social-media-stats/all/denmark

[11] https://twitter.com/RusEmbDK/status/1021688735677734912

[12] http://mm.dk/tjekdet

References

Alexandre Alaphilippe, Alexis Gizikis, Clara Hanot, and Kalina Bontcheva. 2019. Automated tackling of disinformation - Major challenges ahead. European Parliament Think Tank.

Marco T Bastos and Dan Mercea. 2019. The Brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, 37(1):38–54.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint*, abs/1607.04606.

danah boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE.

Samantha Bradshaw and Philip N Howard. 2018. Challenging truth and trust: A global inventory of organized social media manipulation. *The Computational Propaganda Project*.

Ashok Deb, Luca Luceri, Adam Badawy, and Emilio Ferrara. 2019. Perils and challenges of social media and election manipulation analysis: The 2018 US midterms. *arXiv preprint*, abs/1902.00043.

Leon Derczynski, Torben Oskar Albert-Lindqvist, Marius Ven Bendsen, Nanna Inie, Jens Egholm Pedersen, Viktor Due Pedersen, and Troels Runge. 2019. Politikerne og vælgere har hver deres valgkamp på nettet. *Mandag Morgen*.

Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kolliakou, Robert Stewart, Sara-Jayne Terp, Geraldine Wong, Christian Burger, Arkaitz Zubiaga, Rob Procter, and Maria Liakata. 2015a. PHEME: Computing Veracity–the Fourth Challenge of Big Social Data. In *Proceedings of the Extended Semantic Web Conference EU Project Networking session*.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke Van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015b. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

Irving J Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

Genevieve Gorrell, Mehmet E Bakir, Ian Roberts, Mark A Greenwood, Benedetta Iavarone, and Kalina Bontcheva. 2019. Partisanship, Propaganda and Post-Truth Politics: Quantifying Impact in Online Debate. *arXiv preprint arXiv:1902.01752*.

Genevieve Gorrell, Ian Roberts, Mark A Greenwood, Mehmet E Bakir, Benedetta Iavarone, and Kalina Bontcheva. 2018. Quantifying media influence and partisan attention on Twitter during the UK EU referendum. In *International Conference on Social Informatics*, pages 274–290. Springer.

Robert Gorwa. 2017. Computational propaganda in Poland: False amplifiers and the digital public sphere. *Computational Propaganda Research Project Working Paper*, 4(2017).

Pedro Calais Guerra, Roberto Nalon, Renato Assunção, and Wagner Meira Jr. 2017. Antagonism also flows through retweets: The impact of out-of-context quotes in opinion polarization analysis. In *Proceedings of the International Conference on Web and Social Media*. AAAI.

Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1):eaau4586.

David Harley. 2017. Origin of the Suspicious: The Evolution of Misinformation. Technical report, ESET United Kingdom.

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of CHI*, pages 237–246. ACM.

Philip N Howard and Bence Kollanyi. 2016. Bots, #strongerin, and #brexit: computational propaganda during the uk-eu referendum. *Available at SSRN 2798311*.

Andreas Kirkedal, Barbara Plank, Leon Derczynski, and Natalie Schluter. 2019. The Lacunae of Danish Natural Language Processing. In *Proceedings of the Nordic Conference on Computational Linguistics (NODALIDA)*.

Ema Kušen and Mark Strembeck. 2018. Politics, sentiments, and misinformation: an analysis of the Twitter discussion on the 2016 Austrian presidential elections. *Online Social Networks and Media*, 5:37–50.

Dave Lee. 2013. Silk Road: How FBI closed in on suspect Ross Ulbricht. *BBC*.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Alice Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. *New York: Data & Society Research Institute*.

Panagiotis Metaxas, Eni Mustafaraj, Kily Wong, Laura Zeng, Megan O'Keefe, and Samantha Finn. 2015. What do retweets indicate? Results from user survey and meta-review of research. In *Ninth International AAAI Conference on Web and Social Media*.

Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015. Finding opinion manipulation trolls in news community forums. In *Proceedings of CoNLL*, pages 310–314.

Vidya Narayanan, Philip N Howard, Bence Kollanyi, and Mona Elswah. Russian involvement and junk news during Brexit. *The Computational Propaganda Research Project. Algorithms, automation and digital politics*.

Lisa-Maria N Neudert. 2017. Computational propaganda in Germany: A cautionary tale. *The Computational Propaganda Research Project*, 7:2017.

Thomas Nygren and Mona Guath. 2019. Swedish teenagers difficulties and abilities to determine digital news credibility. *Nordicom Review*, 40(1):23–42.

Thomas Nygren, Mona Guath, and Anton Axelsson. 2019. Defining and measuring abilities to debunk disinformation among Swedish adults. In *Online disinformation: an integrated view – Aarhus*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Andrew Rojecki and Sharon Meraz. 2016. Rumors and factitious informational blends: The role of the web in speculative politics. *New Media & Society*, 18(1):25–43.

Troels Runge. 2019. Den digitale valgkamp: Hvordan politikere fisker efter vælgere på de sociale medier. *Videnskab.dk*.

Craig Silverman. 2016. This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed*.

Brian G Southwell, Emily A Thorson, and Laura Sheble. 2018. *Misinformation and mass audiences*. University of Texas Press.

Leo G Stewart, Ahmer Arif, and Kate Starbird. 2018. Examining trolls and polarization with a retweet network. In *Proc. ACM WSDM, Workshop on Misinformation and Misbehavior Mining on the Web*.

Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. 2013. Exploiting homophily effect for trust prediction. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 53–62. ACM.

Siva Vaidhyanathan. 2019. Regulating Facebook will be one of the greatest challenges in human history. *The Guardian*.

Jen Weedon, William Nuland, and Alex Stamos. 2017. Information operations and Facebook. Facebook.

Liang Wu, Fred Morstatter, Xia Hu, and Huan Liu. 2016. Mining misinformation in social media. *Big Data in Complex and Social Networks*, pages 123–152.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *arXiv preprint arXiv:1905.12616*.

A Appendix: Party acronyms

Party	Acronym	Official letter	Twitter account
Alternativet	-	Å	@alternativet_
Dansk Folkeparti	DF	O	@DanskDf1995
Det Konservative Folkeparti	K	C	@konservativedk
Enhedslisten – De Rød-Grønne	EL	Ø	@Enhedslisten
Klaus Riskær Pedersen	KRP	E	@KlausRiskaer
Kristendemokraterne	KD	K	@KDDanmark
Liberal Alliance	LA	I	@liberalalliance
Nye Borgerlige	NB	D	@NyeBorgerlige
Radikale Venstre	RV/R	B	@radikale
Slesvigsk Parti	-	S	@SlesvigskParti
SF – Socialistisk Folkeparti	SF	F	@sfpolitik
Socialdemokratiet	SD	A	Spolitik
Stram Kurs	SK	P	@RasmusPaludan^{A1}
Venstre, Danmarks Liberale Parti	V	V	@Venstredk

Parties surveyed, including acronym, the one-letter code that Danish parties use in campaigning, and the Twitter account monitored. Ordering taken from the official parliament website at http://ft.dk/.
$A1$: The prior official account for this party was banned by Twitter.