Account Deletion Prediction on RuNet: A Case Study of Suspicious Twitter Accounts Active During the Russian-Ukrainian Crisis

Svitlana Volkova
Pacific Northwest National laboratory
902 Battelle Blvd
Richland, WA 99354
svitlana.volkova@pnnl.gov

Eric Bell
Pacific Northwest National laboratory
902 Battelle Blvd
Richland, WA 99354
Eric.Bell@pnnl.gov

Abstract

Social networks are dynamically changing over time e.g., some accounts are being created and some are being deleted or become private. This ephemerality at both an account level and content level results from a combination of privacy concerns, spam, and deceptive behaviors. In this study we analyze a large dataset of 180,340 accounts active during the Russian-Ukrainian crisis to discover a series of predictive features for the removal or shutdown of a suspicious account. We find that unlike previously reported profile and network features, lexical features form the basis for highly accurate prediction of the deletion of an account.

1 Introduction

Social media plays an important role in the life of millions of people. 1/7th of the world's population is using social media services such as Twitter, Facebook every day. There is no doubt that social media has positive effects on society by helping us to connect, communicate, access and spread information, and share our interests. Social media services have been effectively used to coordinate disaster responses (Sakaki et al., 2010), enhance emergency situational awareness (Yin et al., 2012) and coordinate crisis events (Bruno, 2011).

However, social media can potentially cause negative effects on our society. Social bots and spammers spread misinformation,[1] deceptive content, propaganda (Berger, 2015), manipulative campaigns over social networks on a large scale extremely fast e.g., several thousands retweets in a few minutes (Ferrara, 2015). Early detection of suspicious accounts that can potentially be spreading misinformation, manipulative and deceptive content is extremely important to ensure a safer and healthier environment in social media (Bamman et al., 2012b; Subrahmanian et al., 2016).

In this work we present an approach for automatically detecting deleted accounts in RuNet[2] collected during the Russian-Ukrainian crisis in 2014 - 2015. We focused on this data because news media reported several cases of misbehavior and deceptive content spread by suspended or allegedly deleted accounts on Twitter relevant to the crisis.[3] Unlike the existing work on social bot prediction (Ferrara et al., 2014), suspended account analysis (Thomas et al., 2011) and non-personal and spam user detection (Lin and Huang, 2013; Guo and Chen, 2014) we focus on a much harder task of automatically identifying fraudulent accounts (sometimes called trolls[4]). Unlike social bots or spam accounts, troll profiles on Twitter and other social networks e.g., LiveJournal, VKontakte are created to look like real users. Trolls have similar follower and friend counts as the legitimate users engage in communications with other users, express opinions etc. That's why they are very difficult to detect compared to social bots

[1] Syrian hackers claim AP hack that tipped stock market by $136 billion. Is it terrorism? (Fisher, 2013).

[2] RuNet – Russian-language community on the Internet.

[3] Inside Putin's Campaign Of Social Media Trolling And Faked Ukrainian Crimes (Gregory, 2015), Ukraine conflict: Inside Russia's 'Kremlin troll army' (Bugorkova, 2015).

[4] Europe's new cold war turns digital as Vladimir Putin expands media offensive (Boffey, 2016).

or spam accounts. Recent work on bot detection[5] analyzed 20,500 Twitter accounts that tweeted similar statements around key breaking news and events. The study suggested that bots follow many other bots, have no favorites and have no timezone, and never interact with other users through @replies and @mentions.

This is the first work that focuses on building predictive models and analyzing the effectiveness of different features to detect deleted accounts (including trolls[6]) on Twitter using deeper linguistic analysis of user-generated content in Russian and Ukrainian, sentiment and emotion features, text embeddings and topics, in addition to profile, network, and behavior clues.

2 Approach

2.1 Dataset

To collect our data we sampled Twitter accounts which used crisis-related keywords in Russian or Ukrainian[7] from the 1% Twitter feed from Mar 2014 to Mar 2015. For example, translated tweet with the crisis-relevant keywords (underlined) is: *A cache of rocket-propelled grenades was found in Kyiv which could be used for terrorist attacks*.

The original dataset had 3.5 million users who used crisis-relevant keywords during this period. We then re-crawled a random sample of 1 million accounts within a couple of months (Jun 2015) of the initial data collection (Mar 2015). We discovered that 30% of previously active accounts have been deleted. We re-crawled these accounts in Dec 2015 to validate the accounts that have been deleted as of Mar 2015 and still remain deleted as of Dec 2015. We call this portion of the data *deleted accounts* $D = 94,170$. We then randomly sampled the same number of accounts that were still active e.g., not deleted as of Mar 2015 and still remain active as of Dec 2015. We call this portion of the data *non-deleted accounts* $\bar{D} = 94,170$. For each user $u \in \{D, \bar{D}\}$ we were able to access at least 20 tweets with crisis-relevant keywords as well as user profile metadata.

2.2 Models

We used scikit-learn (Pedregosa et al., 2011) to build models that can predict deleted accounts in social media. We prefer log-linear models over reasonable alternatives e.g., perceptron or SVM, following the practice of a range of previous work in related areas (Smith, 2004; Liu et al., 2005; Poon et al., 2009; Bamman et al., 2012a; Filippova, 2012; Volkova and Bachrach, 2015; Hovy, 2015).

In Table 1 we outline a comprehensive list of features we used to our build models. We significantly expanded the list of features that have been previously used for bot detection on Twitter (Ferrara et al., 2014). In addition to previously used account and behavior features our models rely on deeper linguistic analysis of content (tweets) generated by users, topics and embeddings, as well as visual and affect (sentiment and emotion) features. We outline the details on how we extracted lexical and affect features below.

BoW features Since Russian and Ukrainian are morphologically rich languages, to reduce sparsity and ensure better model generalization, we lemmatized words using pymorphy2 package.[8] We extracted bag-of-word (BoW) features from preprocessed lemmatized tweets; we also excluded all stopwords and words with frequency less than five; we run our experiments varying word ngram size (unigrams, bigrams and trigrams) for binary vs. normalized frequency-based features.

LSA features We performed linear dimensionality reduction on feature vectors extracted using BoW normalized frequency-based features as described above using Latent Semantic Analysis (Dumais, 2004) implemented as truncated Singular Value Decomposition (SVD) in scikit-learn.[9] Similarly, we

[5] Social Network Analysis Reveals Full Scale of Kremlin's Twitter Bot Campaign (Lawrence, 2015).

[6] We can not guarantee that these accounts might be potentially spreading deceptive content. However, after manual inspection of the tweets from 100 deleted accounts we found that all 100 accounts display characteristics and behavior shared by those involved in spreading deceptive content, for example, they only post/repost tweets relevant to crisis, there is high ngram/string similarity among their tweets.

[7] Our lexicon of crisis-related keywords has been built independently by three native speakers of Russian and Ukrainian. The final lexicon contains 53 keywords in both languages e.g., Crimea, revolution, Donetsk, ceasefire, NATO, EU etc.

[8] https://pypi.python.org/pypi/pymorphy2

[9] http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

performed linear dimensionality reduction on feature vectors extracted using hashtags and mentions. We varied the number of dimensions $c = [50, 100, 500]$ to get the best F1 and report the results for $c = 100$.

LDA features We learned topics using Latent Dirichlet Allocation (LDA)[10] (Blei et al., 2003) on 1 million tweets randomly sampled from the original 3.5 million tweets. We varied the number of topics $t = [50, 100, 250, 500, 1000]$, and optimized α and β priors by minimizing log-likelihood. We report the results for $t = 1000$, $\alpha = 0.1$ and $\beta = 0.005$.

Embeddings We learned word embeddings for Russian using Word2Vec's skip-gram and CBOW models (Mikolov et al., 2013) implemented in gensim package[11] with a layer size of 50. The embeddings are learned on the same corpus of 1 million tweets as LDA topics. After learning embeddings, we assign words to clusters by measuring cosine similarity between two word embeddings, and compute clusters using spectral clustering over a word-word similarity matrix.

Affect features Finally, to extract sentiment features we predict polarity score for every tweet for each user using the state-of-the-art sentiment classification system for Russian developed by Chetviorkin et al. (2014), Loukachevitch and Chetviorkin (2014). Polarity scores vary around 0 (neutral) between -2 (negative) and +2 (positive). We calculate mean polarity scores, and the proportions of positive, negative and neutral tweets for every user (Dickerson et al., 2014).

To extract emotion features, we predict one of six Ekman's emotions such as: sadness, joy, fear, disgust, surprise and anger for each tweet using an approach recently developed by Mohammad and Kiritchenko (2015) and Volkova and Bachrach (2015). Similar to sentiment features, we use six emotion proportions per user as features.

3 Experimental Results

3.1 Classification Results

In Table 3 we present account deletion classification results using individual feature types. We report our results using 10-fold cross validation on a balanced set of 188,340 deleted and non-deleted accounts.

Profile (account and behavior) features $\lvert f^{prof} \rvert = 12$
days since account creation, number of followers, number of friends, number of favorites, number of tweets, friend-to-follow ratio, name length in chars, bio in chars, screen name length in chars, screen name length in words, bio length words, avg. number of tweets per hour
Visual features $\lvert f^{vis} \rvert = 658$
bag-of-words (BoW) on profile background color, profile link color, text color, sidebar color, background tile, sidebar border color, default profile image
Syntactic features $\lvert f^{syn} \rvert = 14$
aver. tweet length in words, aver. tweet length in chars, retweet rate: prop. of RTs to tweets, uppercase word rate, elongated word rate, repeated mixed punctuation rate, prop. of tweets with links, tweets that are retweets (RTs), prop. of tweets with mentions, hashtags, punctuation, emoticons, mention, hashtag, url rate per word
Network features $\lvert f^{men} \rvert = 159,563, \lvert f^{ht} \rvert = 7,983$
mentioned and retweeted users (@mentions), LSA on @mentions with $c = [50, 100, 500]$ dimensions, BoW on hashtags, LSA on hashtags with $c = [50, 100, 500]$
Lexical features $\lvert f^{lex} \rvert = 110,302$
bag-of-words (BoW) on tweets, LSA on tweets, LDA on tweets with $t = [50, 100, 250, 500, 1000]$ topics embeddings with $d = [30, 50, \ldots, 2000]$ dimensions
Affect (sentiment and emotion) features $\lvert f^{affect} \rvert = 12$
number of emoticons, prop. of emotions, mean scores, prop. of tweets with positive, negative, neutral sentiment,

Table 1: Profile, visual, lexical, network and affect features used for account deletion prediction.

We found that lexical features are the most predictive yielding F1 as high as 0.87. Interestingly, we found that frequency-based features outperform binary features. *It means that for account deletion prediction it is not only important what the users say but how much they say it.* We also found that higher order ngrams only slightly outperform unigram features. When the dimensionality of the feature space is reduced from 110K to 1000 (Embeddings), 1,000 (LDA), and 100 (LSA), classification results drop by 0.11, 0.06 and 0.03, respectively. Syntactic features extracted using shallow linguistic analysis demonstrate lower F1 than lexical features, but higher F1 of 0.81 than the rest of non-lexical features.

Similar to earlier work, we found that profile features have high predictive power for detecting deleted accounts yielding F1 as high as 0.85. Network features have moderate predictive power,

[10] https://pypi.python.org/pypi/lda

[11] https://radimrehurek.com/gensim/models/word2vec.html

Feature Type	Example features sorted by predictive power for deleted D and non-deleted $\bar{D}$ accounts
Lexical	D: end, cressid, sokrin, alphabet, web money, haim, master, video segment, klyati, forest restoration $\bar{D}$: arbi, mes, venta, lambesis, cozy, nikolay, restrict, agreement, perl, chubais, ethernet, insulation
Hashtags	D: #volkswagen, #win, #meat, #slovenia, #therewillneverbeanotheronedirection, #crishtian, #kebab $\bar{D}$: #brent, #novorussia, #gromaidan, #leg, #hydroelectric, #media, #plantyourowntree, #underwater
Mentions	D: @newskazru, @volumesocial, @whafar, @max_7korolei, @chernyj1974, @dreamknoxville $\bar{D}$: @agnfkvvaalena, blascepna72, @chico6, @xagiqasez, @kathrynbruscobk, @deanarianda
Topics	D: 337: beat up, resolve, press office, parliamentarian, intimidation; 376: accountability, position $\bar{D}$: 792: reach, captain, fluffy, quit the job, shoot, satellite; 310: quarter, hitchcock, pitting, ensue

Table 2: The most discriminative unigrams, hashtags, mentions and topics (translated) for account deletion prediction.

Feature Type	F1	P	R
Profile			
Account + behavior	**0.85**	0.84	0.86
Visual	**0.73**	0.65	0.83
Language			
Syntactic	0.81	0.77	0.85
BoW tweets	**0.87**	**0.89**	**0.86**
LSA tweets	0.84	0.89	0.79
LDA tweets	0.81	0.85	0.78
Embeddings	0.76	0.68	0.85
Network			
Hashtags	0.76	0.63	0.96
LSA hashtags	0.73	0.59	0.97
Mentions	**0.78**	0.66	0.96
LSA mentions	0.72	0.60	0.91
Affect			
Sentiment + emotion	**0.72**	0.64	0.81
ALL	**0.82**	**0.79**	**0.88**

Table 3: Classification results in terms of F1, precision (P), and recall (R) based on individual feature types.

with mentions demonstrating F1=0.78 and hashtags F1=0.76. Interestingly, unlike lexical features, binary and frequency-based mention and hashtag features demonstrate equal classification results. *It means that for account deletion prediction it is not important how much the users use some hashtags or @mentions, but whether they use them or not.* Finally, sentiment and emotion features yield comparable F1 of 0.72 to visual features.

3.2 Feature Analysis

To show that the differences between deleted and non-deleted accounts are statistically significant we performed a Mann-Whitney U-test on account, affect and syntactic features (Mann and Whitney, 1947). We found all differences to be significant (p-value ≤ 0.001). We outline our key findings below.

Profile differences Deleted accounts have less followers than non-deleted accounts, but they have more friends. They have less favorites than non-deleted, as well as the tweets, and significantly lower friend-to-follower ratio. Deleted account have significantly shorter bios, but longer user names.

Syntactic differences Deleted accounts generate shorter tweets, use less elongated words, capitalized words and repeated punctuation. They have lower hashtag, mention and url per word ratios. They produce significantly less retweets, tweets with hashtags, urls and mentions, tweets with punctuations and emoticons than non-deleted accounts.

Sentiment and emotion differences Deleted accounts produce *less positive tweets, more negative and more neutral tweets* compared to non-deleted accounts. Deleted accounts *express less anger, but significantly more sadness and fear in their tweets*. Both account types produce comparable amounts of joy, disgust and surprise emotions.

We present the examples of the most discriminative ngram, mention, hashtag and topic features learned by our models in Table 2.

4 Conclusion

We presented the first work on suspicious account deletion prediction in RuNet. We analyzed the predictive power of a variety of previously unexplored features including lexical, topics, hashtags, mentions, sentiments and emotions, in addition to the existing profile and behavior features. We found that deleted and non-deleted accounts on Twitter not only have different profiles, but also express significant differences in topics, hashtags and lexical terms they mention, the ways they generate tweets (syntactic differences), as well as sentiments and emotions they express. All of these differences allow building highly accurate models for detecting suspicious accounts in social media.

References

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2012a. Gender in twitter: Styles, stances, and social networks. *CoRR abs/1210.4567*.

David Bamman, Brendan O'Connor, and Noah Smith. 2012b. Censorship and deletion practices in chinese social media. *First Monday*, 17(3).

J.M. Berger. 2015. The evolution of terrorist propaganda: The paris attack and social media. http://www.brookings.edu/research/testimony/2015/01/27-terrorist-propaganda-/social-media-berger.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Daniel Boffey. 2016. Europe's new cold war turns digital as vladimir putin expands media offensive. http://www.theguardian.com/world/2016/mar/05/europe-vladimir-putin-russia-/social-media-trolls.

Nicola Bruno. 2011. Tweet first, verify later? How real-time information is changing the coverage of worldwide crisis events. *Oxford: Reuters Institute for the Study of Journalism, University of Oxford. Retrieved June*, 10(2011):2010–2011.

Olga Bugorkova. 2015. Ukraine conflict: Inside russia's 'Kremlin troll army'. http://www.bbc.com/news/world-europe-31962644.

Ilia Chetviorkin, Leninskiye Gory Moscow, and Natalia Loukachevitch. 2014. Two-step model for sentiment lexicon extraction from Twitter streams. *Proceedings of ACL*, pages 67–72.

John P Dickerson, Vadim Kagan, and VS Subrahmanian. 2014. Using sentiment to detect bots on Twitter: Are humans more opinionated than bots? In *Proceedings of ASONAM*, pages 620–627.

Susan T Dumais. 2004. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2014. The rise of social bots. *arXiv preprint arXiv:1407.5225*.

Emilio Ferrara. 2015. Manipulation and abuse on social media. *arXiv preprint arXiv:1503.03752*.

Katja Filippova. 2012. User demographics and language in an implicit social network. In *Proceedings of EMNLP-CoNLL*, pages 1478–1488.

Max Fisher. 2013. Syrian hackers claim ap hack that tipped stock market by $136 billion. Is it terrorism? https://www.washingtonpost.com/news/worldviews/wp/2013/04/23/syrian-hackers-claim-ap-hack-that-tipped-stock-market-by-136-billion-is-it-terrorism/.

Paul Roderick Gregory. 2015. Inside putin's campaign of social media trolling and faked ukrainian crimes. http://www.forbes.com/sites/paulroderickgregory/2014/05/11/inside-putins-campaign-of-social/-media-trolling-and-faked-ukrainian/-crimes/#97c1e9629db0.

Diansheng Guo and Chao Chen. 2014. Detecting non-personal and spam users on geo-tagged Twitter network. *Transactions in GIS*, 18(3):370–384.

Dirk Hovy. 2015. Demographic factors improve classification performance. *Proceedings of ACL*, pages 752–762.

Alexander Lawrence. 2015. Social network analysis reveals full scale of Kremlin's Twitter bot campaign. https://globalvoices.org/2015/04/02/analyzing-kremlin-twitter-bots/.

Po-Ching Lin and Po-Min Huang. 2013. A study of effective features for detecting long-surviving twitter spam accounts. In *Proceedings of ICACT*, pages 841–846.

Yang Liu, Qun Liu, and Shouxun Lin. 2005. Log-linear models for word alignment. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, pages 459–466.

Natalia Loukachevitch and Ilia Chetviorkin. 2014. Open evaluation of sentiment-analysis systems based on the material of the russian language. *Scientific and Technical Information Processing*, 41(6):370–376.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif M Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Douard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of NAACL*, pages 209–217.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of WWW*, pages 851–860. ACM.

Noah A. Smith. 2004. Log-linear models.

VS Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong

Zhu, Emilio Ferrara, Alessandro Flammini, Filippo Menczer, et al. 2016. The DARPA Twitter Bot Challenge. *arXiv preprint arXiv:1601.05140*.

Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. 2011. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference*, pages 243–258. ACM.

Svitlana Volkova and Yoram Bachrach. 2015. On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. *Cyberpsychology, Behavior, and Social Networking*, 18(12):726–736.

Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. 2012. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, (6):52–59.