Predicting Popularity of Polish Facebook Posts Using Author and Content Features[*]

Nataliya Stepanova,[1] C. Anton Rytting,[1] Ewa Golonka,[1] Susannah B. F. Paletz[1]

[1] University of Maryland, College Park MD 20742
nstepan1@terpmail.umd.edu, {crytting,egolonka,paletz}@umd.edu

Abstract. Social and political discourse on social media platforms has been a persistent topic of academic study in the past decade. The question of why certain content is shared is increasingly important for understanding the impact of social media on the democratic process. In the present study we determined how well a Facebook post's semantic content can distinguish "popular" from "unpopular" posts. Using a corpus of roughly half a million Facebook posts by Polish socio-political influencers, we conducted two experiments: the first compared the performance of topic features (obtained via topic modeling) against author features, while the second compared topic features against a bag-of-words feature set. Our first experiment showed that previous findings from Twitter cascades were replicated on our Facebook dataset: although topic features predicted popularity better than chance, they performed worse than author features. Our second experiment showed that topic features performed worse than bag-of-words features in predicting popularity. Still, in the analysis of individual (weighted) features, we found that the topic features were more interpretable than an ordered list of individual words, as they yielded interesting insights into the kinds of topics most likely to be shared.

Keywords: topic modeling, social media, Facebook, natural language processing, sharing behavior, Poland, complex modeling, logistic regression

1 Introduction

Social and political discourse increasingly occurs on social media platforms, reflecting a digital realm where social and political influencers can reach and influence their audiences [1-2]. The question of why certain content is shared (and in the extreme case, goes viral) is arguably important for understanding how social media can impact the democratic process. For example, Chmielewska-Szlajfer [3] explored the role that

[*] This work is supported by the Minerva Research Initiative / Office of Naval Research (ONR) Grant #N00014-19-1-2506. The views and conclusions in this work are the authors' and should not be construed as representing the views of the Minerva Research Initiative, the Office of Naval Research, or the United States Government. The authors would like to thank Dr. Rebecca Goolsby for supporting the research and Dr. Cody Buntain for collecting the data from Facebook and for helpful comments on earlier drafts of this paper.

Facebook may have had in the Polish 2015 election. She pointed out that, insofar as Facebook's algorithm uses (or has used) shares and likes to determine what to show to others, sharing can have a snowball effect, increasing the visibility of posts beyond the immediate audience with whom users share those posts. Facebook's design is well-suited for stimulating within-country political discussion (at least in Europe), in contrast to Twitter, which better fosters international communication [4].

Researchers have applied diverse methods for exploring why users share content on social media. In a survey of university students in Singapore focusing on the sharing of misinformation [5], the top three responses highlighted characteristics of the information itself: the content was deemed new or eye-catching, a good topic of conversation, or interesting. A fourth reason highlighted information sharing: "sharing helps me get other people's opinions regarding the information/event." By contrast, survey responses for "the information seems accurate" and "the information comes from an authoritative source" were much lower than the top five reasons above, and also lower than reasons related to the information's usefulness, currency, and relevance to understanding a particular situation. Most of the high-ranking reasons focus on the post's content rather than the sharer's goals or the source's credibility.

Schreiner et al. [6] conducted a review of 45 studies focusing on which aspects of content influence engagement with commercially-oriented social media content. A few studies used physiological or behavioral measures, while others used content analysis. Most studies focused on topic (broadly defined), and a majority of these studies focused on one or more predefined topic categories. A smaller number of studies examined emotion, sentiment, emotional response, or humor.

Another class of studies use machine learning methods to examine sharing as a prediction task. In principle, such predictive models stand to complement explanatory models [7]. Predictive models can quickly explore a wide range of features over a large corpus, so long as the features can be easily obtained from the text itself or other readily available information without relying on labor-intensive annotation. On the other hand, the variables that are readily available without annotation may not be particularly relevant to a deeper understanding of sharing, as these variables may be peculiarities of the chosen dataset or task. In practice, moreover, the generality of insights from predictive models has been somewhat limited by an overly narrow focus on a few languages, typically English, and platforms, most notably Twitter [8].

2 Related Work

A multitude of studies have investigated the performance of a variety of metrics in predicting information spread online. These models vary in terms of features, machine learning models, and the definition of "success" used in the chosen prediction task, so it is difficult to do cross-comparisons of the effectiveness of various features in predicting popularity [9–12]. Moreover, while many studies have been conducted on Twitter, user sharing patterns may vary across social media platforms [13]. Here we discuss three studies that were particularly influential on our methodology.

Martin et al. [9] obtained a theoretical upper bound on predictive performance of an ex-ante prediction task on Twitter, where in an ex-ante task only information that is collected prior to the event of interest can be used for prediction (as opposed to situations where the first few hours/minutes of an event are used to predict future success). By attributing predictive error to two main sources, insufficient data and inherent unpredictability, they determined that predictive performance is limited by a theoretical upper bound even under perfect knowledge.

Jenders et al. [10] compared the performance of a Naive Bayes model against a generalized linear model in predicting whether a tweet would be viral, as defined by getting more than a threshold value T retweets, based on a combination of user and tweet features. The specific predictors included number of followers (user feature), as well as tweet length, hashtags and mentions, and sentiment (tweet features). They found that the generalized linear model generally provided more accurate predictions than the Naive Bayes model, with the generalized linear model additionally having the advantage of interpretability: analysis of the learned feature weights revealed the number of a user’s followers to have the most predictive power for the task.

Bakshy et al. [11] analyzed the sizes of Twitter cascades generated by users who “seed” a URL (i.e. they didn’t receive the URL through their follower graph) that subsequently gets diffused through the network of Twitter users. They defined user features in terms of both seed user attributes and past influence, finding that past local influence and number of followers are consistent predictors of future influence. Additionally, all defined content features (as determined by human annotators) were found to be less informative than the seed user features.

Our Contribution. We present two experiments that use predictive machine learning models as an initial exploration of content-based factors associated with sharing of Facebook posts of social and political relevance. The first experiment compares factors related to the textual content of the message itself with factors related to the user posting that message. The second experiment compares two methods for representing the content of a message computationally: one which represents presence or absence of individual words in the message (or, more precisely, lemmas) as separate features, and one which reduces the dimensionality of the lemma set with topic models. We use a very simple machine learning approach for both experiments, which allows for a more transparent examination of the role of individual features in the predictions. Unlike much of the work in this space, a non-English Facebook corpus is used for the investigation, thus probing the generality of earlier findings.

3 Experiment #1: Author vs. Content Features in Predicting Popularity

3.1 Methodology

Dataset. We conducted all analyses on a corpus of posts from public Facebook accounts based in Poland. This corpus was collected as part of a larger effort investigating information propagation online.[1] Facebook was chosen as the primary platform for this collection because of its popularity in Poland: a recent survey of 427 Polish students found that 90% of respondents used Facebook daily, compared with only 8.9% who used Twitter regularly [14].[2]

To obtain the accounts of interest, a master list of influential Poles was compiled by conducting Google searches both in Polish and English, looking at trending topics via Google Trends and similar sources, and identifying the official social media pages for political parties or social organizations from a wide spectrum of political orientations. From this master list, potential influencers were excluded if they did not focus on social and/or political issues, were inactive, or did not have a social media presence on Facebook or YouTube. This process yielded 360 social/political influencers, including political parties, TV/Radio stations, non-profits, and think tanks, as well as individual politicians, journalists, activists, historians, and so on. We obtained the public posts for their Facebook accounts from January 1, 2015 to July 2, 2020. The posts, with their sharing statistics, were collected in 2020.

For this experiment, we selected as our testing period a 3-month time window around the Polish Women's March, which occurred on October 3, 2016 ("Black Monday") in protest to the proposal of a law that would criminalize abortion and make it illegal under any circumstance, including in cases of incest and rape, as well as when the life of the fetus was under threat. Accordingly, we built our models using posts from January 1, 2015 to November 2, 2016, where all training posts came from January 1, 2015 - August 3, 2016 (training period), and all testing posts came from August 4, 2016 - November 2, 2016 (testing period). Due to the temporal train-test split, we predicted future events strictly without any knowledge of the event itself or other events within the testing period.

Data Cleaning and Preprocessing. We removed posts that were written by an account that did not post at least once in the training period, resulting in 261 author accounts in the testing set. To prepare the text data for topic modeling, we removed URLs, excluded posts that were not detected as being in Polish,[3] replaced common emoticons with unique identifiers, removed numbers, Unicode punctuation, and stop words,[4] and lemmatized all word tokens using the *spaCy* Python package [15]. Furthermore, we excluded all posts that did not contain 10 tokens left after preprocessing, which left us with 533,203 posts. Splitting the posts temporally on August 3, 2016 resulted in a training set of 443,835 posts and a testing set of 89,368 posts (train-test split was thus 83.2%-16.8%). After preprocessing, the number of tokens in our vocabulary was 317,318.

[1] While the larger corpus collection includes both Facebook and YouTube posts from both Poland and Lithuania, only the Facebook portion of the Polish corpus is used for this study. See http://emotionsinsocialmedia.umd.edu/ for more information.

[2] Chmielewska-Szlajfer [3] cited surveys with similar figures from 2015: 80% use for Facebook versus only 13% use of Twitter.

[3] See https://pypi.org/project/langdetect/.

[4] We used the stop words list from the following Polish sentiment repository: https://github.com/Ermlab/pl-sentiment-analysis/blob/master/Data/PolishVocabulary/polishstopwords.txt

Defining "Unpopular" and "Popular" Posts. Following Jenders et al. [10], we tested several different thresholds T={50, 100, 500, 1000} for the number of shares that would differentiate "unpopular" from "popular" posts. Table 1 lists the percentage of all posts defined as "popular" given each T.

To run a balanced binary prediction task, we had to randomly down-sample the "unpopular" posts to match the number of "popular" posts, which balanced the class distributions and prevented the logistic regression from learning degenerate models. After down-sampling for each T, we fixed the training and testing sets in order to control for any effect of varying the random down-sampling on the performance of the logistic regression. Thus, we had four fixed train-test splits for each T, with Table 1 summarizing the sizes of each respective training and testing set.

Table 1. Percentage of "popular" posts (those with > T shares), sizes of training and testing sets depending on threshold T.

T	% "popular" posts	Size of training set	Size of testing set
50	25.12%	229,148 posts	38,768 posts
100	15.09%	137,484 posts	23,454 posts
500	2.20%	19,940 posts	3,496 posts
1000	0.67%	5,956 posts	1,142 posts

Content Features. We used topic modeling to obtain a quantitative representation of the content of a Facebook post. Topic modeling is a general term that encompasses a variety of statistical methods for discovering a prespecified number of hidden "topics" (groupings of words) in a corpus of text documents by assuming a generative approach of document production. Although there are many variants of topic modelling, we used the *gensim* Python package to run Latent Dirichlet Allocation (LDA) with K={25, 50} topics [16-17]. We did not feed posts in the testing set into our topic model, so there was a possibility of new topics emerging in the testing data that would remain hidden. Even though we randomly down-sampled the number of posts in the training set to ensure our binary prediction task was balanced, the topic model was fit on the *entire original training set* to make it immune to any random changes incurred from down-sampling. We obtained three topic models for each K={25, 50}, varying the random seed for each of the three trials to prevent our results from being affected by the quality of the topic model. Using these topic models we obtained the topic distribution (a vector summarizing what percentage of the post comes from each discovered topic) for posts in both the final training and testing sets. We refer to a post's topic distribution as its *topic features.*

Author Features. Following Bakshy et al.'s work [11] on predicting Twitter cascades, we identified five author account features: two seed user attributes (number of Facebook friends, number of posts in training period) and three "past influence" of user (Bakshy et al.'s term: i.e., *popularity of user's prior posts*) attributes (minimum, mean, and maximum number of times a post by the author from the training period had been shared by others at the time we downloaded the post). Note that although

Bakshy et al. [11] defined additional seed user and “past influence” of user attributes, we were not able to perfectly replicate their feature set due to differences in the information available for Twitter vs. Facebook. We refer to this set as *author features*.

Logistic Regression. We used the *scikit-learn* Python package [18] to fit a logistic regression model on the training dataset, where a logistic regression is a very simple machine learning model that predicts the categorical class of an item characterized by a set of features. We chose this model as analysis of the weights learned allows for the comparison of their relative importance (the higher the absolute value of the weight associated with a feature, the greater its predictive power). We specifically focused on three combinations of features to compare their predictive power: (1) author features, (2) topic features, and (3) author and topic features.

The author features predictor was as described above (# friends, # posts, min shares, mean shares, max shares), the topic features predictor consisted of a vector of length K={25, 50} (post’s topic distribution vector as obtained with our topic model), and the author and topic features predictor consisted of a vector of length K+5 (concatenation of topic distribution and author features vectors).

3.2 Results

Table 2 summarizes the accuracy of the logistic regression models for all tested T values and the three tested predictors. Since we had three different topic models for each K={25, 50} we report the average accuracy over these three models for predictors that include topic features (to control for the possibility of the quality of one topic model affecting the results).

Table 2. Accuracy of logistic regression models trained on three feature sets (author features only, topic features only, author + topic features) in predicting “popular” vs. “unpopular” posts.

		Topic features only		Author + topic features	
T	Author features only	K=25	K=50	K=25	K=50
50	0.6941	0.6019	0.5926	0.7043	0.7032
100	0.7106	0.6044	0.5950	0.7143	0.7159
500	0.7517	0.6174	0.6092	0.7418	0.7425
1000	0.7408	0.6565	0.6421	0.7498	0.7533

3.3 Discussion

As T increases, performance across all of the tested predictors also almost exclusively increases, with the total increase in performance from T=50 to T=1000 being approximately 5%. This suggests that differentiating between the lower 75% and upper 25% of posts (in terms of shares) is harder than differentiating between the lower 99% and upper 1% of posts. The effect of an increased T improving prediction performance was also observed by Jenders et al [10].

Additionally, we were able to replicate Bakshy et al.'s [11] findings for Twitter: content features were less important than author features in their predictive power. First, we found essentially no improvement in the accuracies obtained by the combined author and topic feature models as compared to the baseline author feature models for the same T. Second, analysis of the weights learned by the logistic regression showed that average over past shares was the single most powerful predictor of the "popular" posts class, followed by post count for T={50, 100} and number of friends of T={500, 1000}. The weights of even the most strongly weighted individual topics were an order of magnitude smaller than these features' weights. This suggests that there is at least some continuity between user sharing patterns between English-language Twitter and Polish-language Facebook.

4 Experiment #2: Topic Modeling vs. Bag-of-Lemmas in Representing Content for Predicting Popularity

Our second experiment investigated whether topic-based representations of content were more effective as predictors when compared to a baseline bag-of-words representation of a Facebook post, in which the frequency of each word in a post constitutes a separate input feature. Because Polish is morphologically complex, and we wished to treat all inflectional variants of a word as a single input feature, we lemmatized all word forms and used a bag-of-lemmas (BOL) model as the baseline.

4.1 Methodology

We used the same preprocessed dataset, the same topic models obtained for K={25, 50}, and the same fixed under-sampled training sets for each threshold T of popularity as described in the methodology for Experiment 1.

Logistic Regression. We fit a logistic regression model on the under-sampled training sets and compared the predictive power of two sets of features: (1) topic features, and (2) BOL features.

The topic features predictor consisted of a vector of length K={25, 50} (topic distribution for a post obtained via our topic model), and the BOL features predictor consisted of a sparse vector representation of all the lemmas in our processed post (using the same word dictionary of 317,318 types as in Experiment 1).

4.2 Results

Table 3 summarizes the accuracy of the logistic regression models for all tested T values and the two tested predictors. Again, we report the average accuracy over the three topic models used to determine the topic features.

Table 3. Accuracy of logistic regression models trained on BOL and topic features.

		Topic features only	
T	Bag-of-Lemmas	K=25	K=50
50	0.6879	0.6019	0.5926
100	0.6964	0.6044	0.5950
500	0.7017	0.6174	0.6092
1000	0.7242	0.6565	0.6421

4.3 Discussion

The topic features performed substantially worse than the baseline performance of BOL features, which suggests that the dimensionality reduction offered by topic modeling loses valuable information for predicting popularity. However, the advantage of using topic features is that the weights learned by the logistic regression allow for an easier qualitative interpretation of the socio-political topics of interest in Poland. Specifically, we found that topics associated with family and health, as well as money and the workplace, were some of the most predictive of "popular" posts at T=1000. Additionally, since our training dataset was taken over a time period leading up to the Women's March, several topics that were strong predictors of the "popular" posts included the lemma for woman (*kobieta*). (Interestingly, however, the top 100 lemmas most predictive of "popular" posts from the BOL model did not include this lemma.) We found different patterns for lower thresholds of T: topics about President Andrzej Duda and immigration were strongly predictive of post popularity. One possible explanation is that although some semantic spaces are somewhat more popular than others (e.g. politics), those differ from the topics that are hugely popular.

It is not immediately clear how this pattern of topics at different values of T can become transparent through the analysis of the weights learned by the BOL model, although we did see changes in the ordering of the most positively weighted lemmas in the BOL model as we compared different T values. While many of the BOL model's most positively weighted lemmas were also strongly associated with several of the more positively correlated topics, it might be difficult to interpret the relevance of these lemmas in isolation without having seen them grouped into topics. On the other hand, given that the topic modeling process can generate different topics for the same set, it may require multiple runs to find all the relevant topics. For example, for all T values some of the most positively correlated lemmas included explicit calls to share (*udostępniać, udostępnić*), which only some of the generated topic models captured. Additionally, comparing the weights learned by the BOL model may be useful for ranking socio-political influencers, politicians, and/or political parties/organizations in terms of popularity, which is harder to do with the topic-based model.

These findings suggests that BOL and topic models may be picking up on different but complementary interpretations of semantic content. When we examined the intersection between the predictions for the BOL model and the topic model, we found only 66% overlap, suggesting that a model combining both methods could outperform either alone. It is possible that topic models are able to find the broader semantic spaces that are more/less popular, and BOL models are able to pick up on words or short phrases that have particularly strong predictive power.

5 Limitations

Our Facebook posts were categorized into one of four types: status, link, photo, or YouTube. Of these four, only status posts consisted solely of author-generated text, as the other three included multimodal content: photo(s), link to website, or link to a YouTube video. Analyzing status posts only would have heavily reduced the size of our dataset. Instead, we tried to mitigate the effect of non-linguistic data through the exclusion of posts that had less than 10 tokens left after preprocessing. However, the presence of non-linguistic content that could not possibly be captured by our BOL or topic models is an inherent limitation of the current methods.

Second, it is possible we did not choose the optimal topic models for our prediction task. The dictionary size used for topic modeling was quite large, consisting of well over 300,000 types. Trimming the vocabulary to only the most informative lemmas could result in better models, which in turn could lead to better predictions.

Additionally, we faced computational constraints. Since fitting a logistic regression on unbalanced classes can result in a degenerate model that always predicts the most frequent class, we had to randomly down-sample the “unpopular” posts class, which may have altered which features (whether topics or lemmas) are discriminatively associated with popular posts. Jenders et al. [10] similarly used equal numbers of viral and non-viral tweets for training and testing in their experiment. Still, it would be preferable to solve the problem of unbalanced classes through weighting or choosing a model that is not sensitive to class size differences. Also, specifically for the BOL model, the number of predictors used as features consisted of over 300,000 lemmas (more than the size of our training sets for all T), which caused some overfitting of our BOL model.

6 Conclusion

Using a logistic regression for a binary prediction task that separates “popular” from “unpopular” posts around the Women’s March event in Poland, we: (1) replicated Twitter content sharing findings comparing the roles of content vs. an author’s prior shares in predicting sharing on our Facebook dataset (in Experiment 1), and (2) compared a topic-based approach with a naive BOL approach in modeling a post’s semantic content (in Experiment 2). Although the BOL model had a higher accuracy than the topic-based model, topic distributions were easier to interpret. Ultimately, BOL and topic-based models were capturing complementary aspects of the semantic content of a post: while a topic model revealed what general semantic spaces were more popular, BOL features captured terms that were particularly viral independent of their participation in specific topics. Future work in this space will investigate newer variants of topic models geared towards short texts and textual features associated with emotion to elicit further insights into users’ sharing patterns.

References

1. Allcott, H., Gentzkow, M.: Social Media and Fake News in the 2016 Election. J. Econ. Perspect. 31, 211–236 (2017). https://doi.org/10.1257/jep.31.2.211.
2. Shifman, L.: Memes in Digital Culture. MIT Press, Cambridge, MA (2014).
3. Chmielewska-Szlajfer, H.: Opinion dailies versus Facebook fan pages: the case of Poland's surprising 2015 presidential elections. Media Cult. Soc. 40, 938–950 (2018). https://doi.org/10.1177/0163443718756065.
4. Bossetta, M., Dutceac Segesten, A., Trenz, H.-J.: Engaging with European Politics Through Twitter and Facebook: Participation Beyond the National? In: Barisione, M. and Michailidou, A. (eds.) Social Media and European Politics: Rethinking Power and Legitimacy in the Digital Era. pp. 53–76. Palgrave Macmillan UK, London (2017).
5. Chen, X., Sin, S.-C.J., Theng, Y.-L., Lee, C.S.: Why Do Social Media Users Share Misinformation? In: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 111–114. Association for Computing Machinery, New York, NY, USA (2015).
6. Schreiner, M., Fischer, T., Riedl, R.: Impact of content characteristics and emotion on behavioral engagement in social media: literature review and research agenda. Electron. Commer. Res. (2019). https://doi.org/10.1007/s10660-019-09353-8.
7. Hofman, J.M., Sharma, A., Watts, D.J.: Prediction and explanation in social systems. Science. 355, 486–488 (2017).
8. Tufekci, Z.: Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. Proc. Int. AAAI Conf. Web Soc. Media. 8, (2014).
9. Martin, T., Hofman, J.M., Sharma, A., Anderson, A., Watts, D.J.: Exploring limits to prediction in complex social systems. In: Proceedings of the 25th International Conference on World Wide Web. pp. 683–694 (2016).
10. Jenders, M., Kasneci, G., Naumann, F.: Analyzing and predicting viral tweets. In: Proceedings of the 22nd International Conference on World Wide Web. pp. 657–664. Association for Computing Machinery, New York, NY, USA (2013).
11. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11. p. 65. ACM Press, Hong Kong, China (2011).
12. Petrovic, S., Osborne, M., Lavrenko, V.: Rt to win! predicting message propagation in twitter. In: Proceedings of the International AAAI Conference on Web and Social Media (2011).
13. Smith, A., Fischer, E., Yongjian, C.: How Does Brand-related User-generated Content Differ across YouTube, Facebook, and Twitter? J. Interact. Mark. 26, 102–113 (2012).
14. Król, K., Zdonek, D.: Social media use and its impact on intrinsic motivation in Generation Z: a case study from Poland. Glob. Knowl. Mem. Commun. ahead-of-print, (2020).
15. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spacy: Industrial-strength natural language processing in python. spaCy. https://spacy.io/ (2016).
16. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003).
17. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks. ELDA, Valletta, Malta (2010).
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.: Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).