

The Narrative Detector on Large-scale Facebook Posts:

A Technical Report

Honglin Carson Bao

A brief view of the data set

Let us begin by summarizing the data. We first create the labeled set, which is composed of messages that have been labeled by a human rater. The remainder is a collection of comments that wait to be labeled. Each message's URL is saved as the identical index for each post.

Multiple blank messages exist in both the labeled set and the set waiting to be labeled because that post contains no narrative/non-narrative message. It could be a simple hyperlink, a photograph, or a video. As a result, these posts contribute nothing to the training of the machine classifier. We obliterate these points and clean the data.

Finally, we have 860 messages for the labeled set, 49% of which are narrative (421 messages, label 1) and 51% of which are non-narrative (439 messages, label 0). We also have 4666 unlabeled messages, and we call it the “unlabeled set.” This is a fairly balanced data set for training a machine learning model. When you train a machine classifier on an imbalanced data set (say, 70% positive and 30% negative), your model is inherently biased toward one side. You must rely on more sophisticated techniques to resolve this issue. We are not required to do this, which is a good start.

The idea of the narrative classification model.

The goal of this classifier is to automatically identify the narrative-based posts. Let's first figure out what is a narrative:

A narrative refers to “a representation of connected events and characters that has an identifiable structure, is bounded in space and time, and contains implicit or explicit messages about the topic being addressed” (Kreuter et al., 2007, p. 222).

I design a two-stage model to identify narratives based on the definition.

- Stage 1: **Short Text Topic Modeling**: Can we identify *implicit or explicit messages about the topic being addressed* in such short messages? If not, they are unlikely narratives and will be labeled 0. However, we cannot assert a narrative even if a clear topic is identified. Because we have many messages like this:

“You have a chance to make an impact on the 1 in 8 women who will be diagnosed with breast cancer in her lifetime. Join us LIVE & DONATE during our Manduka #projectOM event in Bryant Park to hear from men and women affected by this disease.”

This message is being labeled by human rater as non-narrative because it looks like an advertisement (though many other lengthy, informative ads are regarded as narratives as well). However, we can still identify an explicit topic in it -- breast cancer. Thus, a second machine learning-based filter has been established. But the messages which have been labeled 0 will not be nor need not be put into the second filter.

about breast cancer (as shown in the two biggest words in the word cloud). The correct way to design a classifier should *NOT* start from **Semantics**. Instead, we should focus on the **Structure**.

I first set the topic-related words as stop words for removal, like 'life', 'women', 'treatment', 'help', 'diagnosed', 'support', 'breast', and 'cancer', because they will confuse the classifier from a semantic perspective. I then made the following assumptions based on the definition of narrative and some rhetoric/communication literature.

1. The narrative is more likely to describe an event that *already happened*. For example, "he has helped cancer patients for three years," "I was diagnosed with cancer three years ago."
[Cortazzi, Martin. "Narrative analysis in ethnography." *Handbook of Ethnography* 384 \(2001\): 394.](#)
2. The narrative is *personal storytelling*. It more likely begins with a persona. For example, "I experienced a big accident three years ago" -- this sentence is highly likely a start of personal storytelling, aka a narrative. "Dr. King developed a new drug. His team xxx," and "Susan was diagnosed with cancer when her child was 3 years old."
[Kajder, Sara B. "Enter here: Personal narrative and digital storytelling." *English Journal* \(2004\): 64-68.](#)
3. Narrating a thing is the most significant feature of a narrative. *Interrogative sentences* and *imperative sentences* are less likely narratives, "Do you want to try this new drug?" "Join us!" -- these two seem like ads. Many messages with paraphrases (quotation "") are also more likely narratives than others -- "I was diagnosed with cancer," Susan said.
[Longacre, Robert E. *The grammar of discourse*. Springer Science & Business Media, 2013.](#)

I study the high-frequency words (top 100) which exclusively appear in narratives but not in non-narratives.

```
['i', 'was', 'my', 'her', 'she', 'had', 'after', 'me', 'through', 'years',  
'just', 'am', 'no', 'were', 'family', 'time', 'during', 'going', 'felt',  
'go', 'there', 'tips', 'being', 'hope', 'told', 'other', 'first', 'i'm',  
'didn't', 'only', 'found', 'chemo']
```

Basically, this list verifies my presumption about personal storytelling and the past period. Also, interrogative sentences and imperative sentences are less likely narratives. We just focus on the "let's" or punctuation to capture narratives vs. interrogative/imperative sentences. It is a bit different from standard preprocessing but makes sense. We then vectorize text data and perform classification algorithms based on the three laws we discovered. You can also choose to add features to the data set manually. **These manually added features should be basically like positive correlations between the prior of persona/past period/particular punctuation and the posterior of narratives, like the work in [this](#) and [this](#).** Note that usual classification might reach high accuracy without manually added features if your data set is not too big. So, it depends.

I tested four models with splitting training/testing 70% and 30% in the original labeled set.

- A naive Bayes classifier with additive (Laplace/Lidstone) smoothing parameter 0.1
- A logistic regression classifier with the inverse of regularization strength 10
- A random forest classifier

- A dense-dropout-dense deep learning model

To evaluate the result, I focus on measurements called **“ROC”** and **“AUC”**. All of them achieve high performance ([Check out this for the report](#)). Furthermore, I average the four generated predictive probabilities of narratives to smooth the result and reduce random errors (It can be regarded as a very naive form of ensemble learning). I also consign a human rater to evaluate the generated labels manually within small generated sets multiple times (because the quality of the original labeled set is not perfect. I will discuss this more later). These examined data and labels will be used as feedback information to further improve the machine classifier model and be re-added into the labeled set to expand it and achieve higher performance. The overall result and generated labels should be attached to the email.

Integrated Smoothing											
URL	Result_NB_1	Result_NB_0	Result_LR_1	Result_LR_0	Result_RF_1/Result_RF_0	Result_DL_1	Result_DL_0	Smoothing_1	Smoothing_0	Message	Label
https://www.facebook.com/	9.4382253393115E-01	5.6177446069838E-01	9.96216710781639E-01	3.78328921639438E-01	7.8E-01	2.2E-01	9.95338191067679E-01	4.9818039332254E-01	9.28769363810607E-01	Our hearts go out to Janet St. James as she works through her treatment plan for her metastatic breast cancer diagnosis. This	1
https://www.facebook.com/	1.92191866304521E-01	8.07808133995479E-01	2.24423070934294E-01	7.75576929065706E-01	5E-02	9.8E-01	2.13783088891717E-01	7.86218911110829E-01	1.70099506531997E-01	Think about all of the strong, powerful women you know. Who	0
https://www.facebook.com/	9.93861680795011E-01	6.13831920499045E-01	9.98990861934863E-01	1.00913806513728E-01	9E-01	6E-02	9.99151857639329E-01	8.48142365059538E-01	9.7300110009967E-01	Jennifer was a healthy, fit and active 44-year-old with no family	1
https://www.facebook.com/	9.9386207833947E-01	6.31792164805332E-01	9.99315100814543E-01	6.84899185457249E-01	9E-01	5E-02	9.9966208363129E-01	3.37916388710364E-01	9.73164815699945E-01	The harsh reality that breast cancer does not discriminate is	1
https://www.facebook.com/	9.97607230514967E-01	2.38276948503209E-01	9.82716751353663E-01	1.72832486463399E-01	7.5E-01	2.8E-01	9.75024462458424E-01	2.4875537541579E-01	9.26337111081794E-01	excellent and caring. Rebecca Thomas is a 72-year-old in Redding,	1
https://www.facebook.com/	8.01068192077785E-02	9.19893180792222E-01	1.824780850954E-01	9.8175219149046E-01	1E-02	9.9E-01	2.2489246687944E-01	9.77510735331209E-01	3.27109730965282E-02	In 2000, Kelly Moncey's mom passed away from breast cancer.	1
https://www.facebook.com/	9.9896286914937E-01	1.05713385006448E-01	9.98871099223507E-01	1.12390077649329E-01	9.8E-01	1E-02	9.98848466205239E-01	1.31503376749099E-01	9.89130982894671E-01	Ever since later while she's been a widow, she found a home.	0
https://www.facebook.com/	9.29428866347527E-01	7.05711336524723E-01	9.96951191849115E-01	3.04880815088489E-01	8.8E-01	1E-01	9.97074265567189E-01	3.2573433811135E-01	9.46013580940958E-01	This is our new dream: to be the "strong" woman. This is	1
https://www.facebook.com/	9.98819754488086E-01	1.18024551191351E-01	9.99968870105731E-01	3.11296942691889E-01	8.1E-01	1.4E-01	9.9960766358081E-01	3.82336419187204E-01	9.52187347737975E-01	At age 30, a breast cancer diagnosis was the last thing on	1
https://www.facebook.com/	9.95695204951739E-01	4.30479504826195E-01	9.97560494126591E-01	2.43950587340333E-01	7.9E-01	2.7E-01	9.9795106235927E-01	2.04893767407309E-01	9.45301690351064E-01	It's a miracle and Andrew were already preparing to make life's	1
https://www.facebook.com/	4.61638884031213E-02	9.5383611159688E-01	6.53550691941713E-01	9.34644930805629E-01	7E-02	9.6E-01	4.3144473553480E-01	9.5685529446514E-01	5.61658577876947E-02	renewment of ourselves through thick and thin, she says.	1
https://www.facebook.com/	9.98153983552459E-01	1.8460164475441E-01	9.99416033532409E-01	5.8396467593808E-01	9.7E-01	4E-02	9.99405415064841E-01	5.94584953158897E-01	9.91743858037427E-01	"I never wanted any of my peers to see me anything other than	1
https://www.facebook.com/	9.98913317780034E-01	1.08668221905476E-01	9.9895656003622E-01	1.43439963780378E-01	9.8E-01	2E-02	9.99892614665625E-01	1.073853437484E-01	9.946656223120695E-01	Thanks for joining our Facebook Live! Our panelists Monisha	0
https://www.facebook.com/	8.00565139168527E-01	1.99343860831471E-01	4.10385745430243E-01	5.89614245469757E-01	2.8E-01	7.2E-01	4.32782221160125E-01	5.5721777839875E-01	4.60956028714724E-01	Byron, Nicholas, Monisha, and Tim Monisha Moore are excited in.	1
https://www.facebook.com/	9.97803472510066E-01	2.19652748993247E-01	9.98143289527034E-01	1.85671437296635E-01	9.3E-01	7E-02	9.98747046636914E-01	1.252953306386E-01	9.81173451207679E-01	"I was 14 years old when my father, who had a BRCA-2	1
https://www.facebook.com/	6.928252022711E-01	3.83439597356837E-01	6.16560402643163E-01		1.3E-01	8.9E-01	5.00484056991982E-01	4.9853543028038E-01	3.30319612944425E-01	mutation, passed away, a time when women broke and	0
https://www.facebook.com/	1.9943434368871E-01	8.005565563113E-01	2.5708965783449E-01	7.4291044216551E-01	4E-01	6.5E-01	2.37415927238184E-01	7.6258490276181E-01	2.73487024635389E-01	You've got to watch this 13-year-old, Jordan Phillips, whose	1
https://www.facebook.com/	8.3770621520796E-01	1.62293784792041E-01	8.5975987359386E-01	1.40240102640614E-01	7.8E-01	1.8E-01	8.81283468328511E-01	1.18718531671489E-01	1.8826548792321E-02	through all the ups and downs, Jordan Myers has been there	0
https://www.facebook.com/	9.94348575081771E-01	5.65324291822866E-01	9.99984018238029E-01	1.59817619721099E-01	8.1E-01	1.4E-01	9.9984807955104E-01	1.53924348955463E-01	9.5107884571226E-01	for his wife, Thomas. It's a reminder, like him, who can make a	1
https://www.facebook.com/	9.9978577830211E-01	2.14223867090007E-01	9.99956305146728E-01	4.36948532720323E-01	9.9E-01	2E-02	9.9966454171423E-01	9.34528587697288E-01	9.97426657155217E-01	Anna was shocked when she was diagnosed with breast cancer.	1
https://www.facebook.com/	3.47351579657756E-01	6.52648423342244E-01	4.28018041618501E-01	9.5719819583815E-01	6E+00	1E+00	1.8524305574659E-01	9.81473639425344E-01	1.0216943534855E-01	Dr. Yilin Kang is fighting metastatic breast cancer. And so are	1
https://www.facebook.com/	9.967180396006348E-02	9.03891960390966E-01	2.70741148900854E-01	9.72925885109915E-01	1.3E-01	9.5E-01	2.29905624722557E-01	9.7700943752674E-01	6.9020679263494E-02	While the news about the COVID-19 pandemic is changing daily,	0
https://www.facebook.com/	9.3473796898021E-01	6.52620311019796E-01	9.88130931316416E-01	3.18698988635843E-01	8.2E-01	2.7E-01	9.81581428384037E-01	1.58142578159329E-01	9.20256910649626E-01	on how much research has been done since we first heard of	1
https://www.facebook.com/	2.40192109105394E-01	7.5980789084607E-01	1.38140050139434E-01	8.61859949860566E-01	3.2E-01	5.9E-01	7.5440403376576E-01	9.24595991662342E-01	1.93443141895621E-01	Dr. Yilin Kang is fighting metastatic breast cancer. And so are	1
https://www.facebook.com/	2.44579651509285E-01	7.55420348490716E-01	9.19481435631628E-01	9.08151856436837E-01	3.3E-01	6.4E-01	1.0982334794331E-01	8.9017865205489E-01	1.940627857944339E-01	We're so moved by the way this mother took her friendship with a late	0
https://www.facebook.com/	1.7423649635626E-01	8.2578303643674E-01	5.52278214375765E-01	9.44772178652424E-01	1E-01	8.1E-01	3.77916877923909E-01	9.82208312207800E-01	9.18140013965734E-02	breast cancer survivor to a whole other level <http://on.today.com/	1
https://www.facebook.com/	9.9075672987886E-01	9.2432732017141E-01	9.9996651703911E-01	3.03348293089898E-01	7.4E-01	2E-01	9.96761705060802E-01	2.3829493919787E-01	8.0655858104379E-01	"When the cancer cell learns to outsmart the treatment drug,	0
https://www.facebook.com/	9.91684407600097E-01	8.31559239990288E-01	9.92644399247075E-01	7.35560075292518E-01	9E-01	1.2E-01	9.9374585026255E-01	6.2534149737446E-01	8.69518847968357E-01	the treatment then falls on you and you're another treatment. Then,	1
https://www.facebook.com/	8.22168675484882E-01	1.77831324515319E-01	7.300265508062638E-01	2.06073449137363E-01	7.3E-01	2.3E-01	7.73046565602662E-01	2.26053044037344E-01	6.79785545509994E-01	Starting to spread the Mother's Day love by surprising these	0
https://www.facebook.com/	1.64429435966197E-01	8.35571564033604E-01	2.80866610035299E-01	9.719133399547E-01	5E-02	8.7E-01	2.98916500529158E-01	9.70118349847394E-01	6.80991857556607E-02	inspiring breast cancer patients, survivors and mothers with.	1
										BREAKING: New therapeutic approach could stop breast cancer	1
										metastasis from forming and spreading. Read about this revolutionary	0
										"Never in a million years, did I think I was going to be diagnosed	1
										with breast cancer. I was 26 years old. I was married to a healthy	1
										As sisters, you dream about sharing friends or clothes, but	0
										never do you dream of about this same disease. That's exactly	1
										He walked his wife into the hospital for her double-mastectomy.	1
										The moment showed emotion, to be understood how his	1
										A little vent session is healthy once in a while. So c'mon, let it	0
										out and tell breast cancer how you really feel in the comments	0

As shown above, the URL column is just the identical link to the post. Result_NB (Naive Bayes), Result_LR (logistic regression), Result_RF (random forest), and Random_DL (deep learning) all have two columns, narrative probability (1) and non-narrative probability (0), respectively. Smoothing_1 and Smoothing_0 are averaged probabilities of four heterogeneous classifiers. Label 1 and label 0 are complementary, and their sum is 1. When the smoothing_1 probability is greater than 0.5 (the smoothing_0 probability is fewer than 0.5 simultaneously), the message will be labeled “1”, i.e., narrative.

Generally, the result is accurate. The rater examined 1-1800 generated labels in total and found 141 misclassified labels (that's why you will see smooth probability doesn't match labels in very sporadic cases -- because these unmatched labels have to be human corrected ones). The overall accuracy is **92.22%**. I think it's acceptable. These misclassifications are mainly because of the ambiguous posts and labels in the labeled set -- We do have some inconsistent labels and narrative criteria, but not too much. The quality of human labeling is fine.

Limitation

One issue might be that I set the narrative standard a bit high, leading to fewer identified narratives than non-narratives.

Another issue is around “narrative” vs. “advertisement.” The classifier performs poorly when it encounters messages like this:

Susan G. Komen and Scopely invite you to Compete for the Cure! From now through December 7, Scopely will make a donation for every Yahtzee With Buddies app installation, for each Komen stick pack collected and for each Komen sticker set completed while playing the game! Scopely is guaranteeing a minimum donation of \$50,000 in support of our mission to end breast cancer.

Download and play:

<https://apple.co/32sh3C3>:=<https://apps.apple.com/app/id1206967173?mt=8>

The classifier, in most cases, labels it as “non-narrative.” It indeed is an advertisement but still contains a useful information structure: “*Scopely will make a donation for every Yahtzee With Buddies app installation, for each Komen stick pack collected and for each Komen sticker set completed while playing the game!*”

Also, whether the messages like this are narratives or non-narrative -- it’s not very precise even in the labeled set. Some labels are 1, but some are 0 though messages look very similar.