# Diachronic degradation of language models: Insights from social media

Kokil Jaidka, Niyati Chhaya, and Lyle Ungar | jaidka@sas.upenn.edu | nchhaya@adobe.com

## Background

Natural languages are constantly evolving and adapting to the needs of their users and the environment of their use. Diachronic differences measure semantic drift specifically for languages over time.

1. Because of diachronic differences, predictive models trained on language may go`stale'.
2. No existing work has investigated whether, and how, language models degrade over time.

## Research Questions

Supervised language models trained on user traits can degrade in performance over time. We explore the extent of the degradation by evaluating:

1. The predictive performance of language models trained at one point in time, in a subsequent time period
2. The implications of diachronic differences in language use on Twitter

## Data

1. 150,000 Twitter posts shared after informed consent by 554 adults in the United States who also answered a Qualtrics survey and reported their age and gender.
2. 130 and 179 million geolocated Twitter posts from the United States, collected from a 10% random sample of the real-time Twitter firehose courtesy the TrendMiner Project [3].

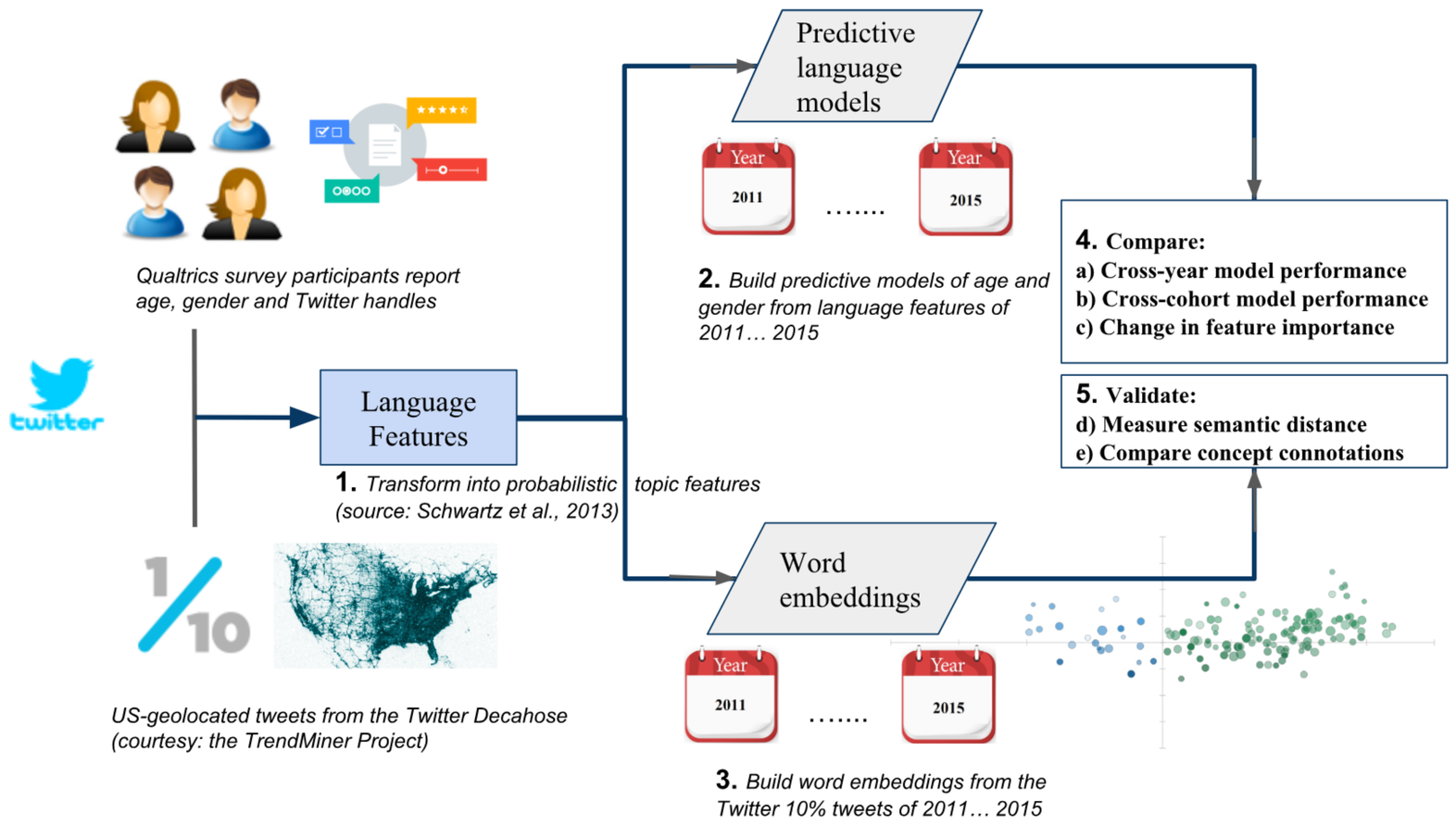
## Insights

1. In every subsequent year, the language of lateteens and early-twenties is more different from the language of their contemporaries from the year before
2. The language of each cohort of 35-year-olds changes little over the previous year.
3. Over time, young users from 2011 continued to use certain topics, while older users adopted newer trends.
4. A part of the language drift appeared because $1/5^{th}$ of the population was shifting along the temporal axis.

## Resources

2000 LDA topics modeled from Facebook posts are available at
`http://www.wwbp.org/data.html`

## Method




1. We establish the diachronic validity of language-based models through predictive evaluations, and also compare against [1].
2. We use topic models and word embeddings to study the diachronic differences in the language of social media users, using methods described in [2].
3. We use linear methods to interpret diachronic differences in user trait prediction as the differences between standardized coefficients in the language models.

## Results

- Models have the lowest mean error in the year that they were trained on, but error increases or decreases in previous or subsequent years respectively.
- This is observed even for in-sample prediction and irrespective of the size of the training set.
- The direction and magnitude of prediction errors are different for different cohorts

| Test set year | 1970 | 1971 | 1972 | 1973 | 1974 | 1975 | 1976 | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 | 1984 | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2011 | **4.4** | **4.6** | **4.7** | **5.1** | **3.4** | **1.7** | **1.8** | **1.4** | **0.7** | **0.9** | **-0.1** | **-1.0** | **-0.9** | **0.0** | **-3.5** | **-2.7** | **-2.0** | **-4.3** | **-4.9** | **-2.5** | **-4.0** | **-4.3** | **-3.3** | **-2.7** | **-4.8** | **-5.6** | **-3.1** | **-2.7** |
| 2012 | 6.5 | 5.0 | 6.8 | 6.2 | 4.4 | 3.1 | 3.2 | 3.1 | 1.2 | 2.2 | -1.1 | -0.8 | -1.2 | -2.1 | -3.7 | -2.9 | -2.9 | -5.2 | -6.0 | -2.9 | -5.6 | -2.9 | -4.3 | -4.7 | -4.0 | -7.2 | -3.9 | -6.1 |
| 2013 | 8.4 | 6.0 | 8.2 | 5.6 | 6.5 | 4.2 | 4.1 | 3.0 | 0.5 | 1.3 | -0.4 | 0.2 | -0.8 | -1.4 | -3.7 | -0.8 | -2.9 | -4.0 | -4.5 | -3.7 | -4.9 | -3.5 | -4.4 | -4.8 | -4.3 | -8.2 | -5.0 | -7.0 |
| 2014 | 9.1 | 8.7 | 8.1 | 7.0 | 7.2 | 6.1 | 5.7 | 3.4 | 1.1 | 2.7 | 0.5 | 1.9 | -0.1 | -3.4 | -2.4 | -1.8 | -2.3 | -3.2 | -4.1 | -2.9 | -3.8 | -4.2 | -4.4 | -4.1 | -4.7 | -6.6 | -4.8 | -7.1 |
| 2015 | 10.9 | 8.7 | 10.8 | 6.7 | 8.1 | 6.6 | 6.1 | 4.3 | 3.4 | 2.7 | 2.7 | 2.0 | 0.1 | -2.9 | -2.9 | -1.2 | -3.2 | -4.0 | -4.5 | -2.7 | -4.3 | -3.3 | -3.8 | -6.7 | -5.9 | -6.2 | -6.2 | -4.9 |
| N | **10** | **9** | **10** | **9** | **17** | **7** | **23** | **19** | **16** | **13** | **28** | **26** | **22** | **21** | **23** | **24** | **19** | **27** | **20** | **25** | **34** | **15** | **27** | **23** | **16** | **14** | **11** | **15** |

Figure 2: Cross-year performance for predicting (a) age (reported as Mean Error = $actualage - predictedage$). The rows reflect the test sets: language samples posted in the same or different year. The columns reflect users stratified according to their year of birth. Deeper shades of blue reflect higher underestimation errors; deeper shades of red reflect higher overestimation errors.

| | Age (Mean Error) | | | | | | Gender (Accuracy %) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Test set | Sap et. al | 2011 | 2012 | 2013 | 2014 | 2015 | Sap et. al | 2011 | 2012 | 2013 | 2014 | 2015 |
| **2011** | 2.2 | **0.0** | 0.2 | 1.1 | 1.6 | 1.8 | 83 | **.86** | .79 | .75 | .75 | .75 |
| **2012** | 3.1 | 0.2 | **-0.1** | 1.1 | 1.8 | 2.1 | 82 | .78 | **.87** | .78 | .77 | .74 |
| **2013** | 3.9 | -0.2 | -0.3 | **0.4** | 0.9 | 1.2 | 80 | .78 | .78 | **.87** | .77 | .77 |
| **2014** | 4.4 | -1.2 | -1.2 | -1.1 | **0.0** | 0.7 | 77 | .78 | .78 | .84 | **.84** | .72 |
| **2015** | 5.0 | -1.3 | -1.5 | -1.3 | -0.4 | **0.0** | 75 | .77 | .78 | .77 | .77 | **.87** |

Figure 1: Cross-year performance for predicting (a) age (reported as $MeanError = Age_{actual} - Age_{predicted}$) and (b) gender (reported as Accuracy). The columns depict the training set for regression models: language samples posted in a particular year. The rows depict the test sets. Deeper shades of blue reflect higher underestimation errors; deeper shades of red reflect higher overestimation errors. Deeper shades of green depict higher accuracy.

| Age | $\beta_{2011}$ | $\beta_{2015}$ |
|---|---|---|
| **Email communication** *(send, email, message, contact)* | 168.5 | -53.7 |
| **Accommodation** *(place, stay, found, move)* | 162.2 | -101.8 |
| **Sleep** *(bed, lay, sleep, head, tired)* | 59.6 | -88.5 |
| **Swear** *(wtf, damn, sh**, wth, wrong, pissed)* | 38.1 | -46.0 |
| **Tiredness** *(i'm, sick, tired, feeling, hearing)* | 33.6 | -98.3 |
| **Hacking** *(virus, called, open, steal, worm, system)* | -99.6 | 253.5 |
| **Software** *(computer, error, photoshop, server, website)* | -87.2 | 80.7 |
| **Feeling** *(feeling, weird, awkward, strange, dunno)* | -70.5.0 | 23.2 |
| **Meetings** *(meeting, conference, student, council, board )* | -44.0 | 38.6 |
| **Skills** *(management, business, learning, research)* | -26.4 | 158.0 |
| **Gender** | $\beta_{2011}$* | $\beta_{2015}$* |
| **Apple products** *(iphone, apple, ipad, mac, download)* | 4.1 | (0) |
| **Sports** *(win, lose, game, betting, streak, change)* | 3.2 | (0) |
| **Bills** *(pay, money, paid, job, rent)* | 2.8 | (0) |
| **Government** *(government, freedom, country, democracy)* | 2.8 | (0) |
| **Prom** *(dress, prom, shopping, formal, homecoming)* | 1.8 | (0) |
| **Hairstyles** *(hair, blonde, dye, color, highlights)* | 1.7 | (0) |
| **Relationships** *(amazing, boyfriend, wonderful, absolutely)* | 1.7 | (0) |
| **Negative emotions** *(inside, deep, feel, heart, pain, empty)* | 1.6 | (0) |

Table 1: The features whose coefficients had the biggest change and flipped sign when comparing the age and gender prediction models trained on 2011 language against those trained on 2015 language. (0) depicts that the feature was no longer significant in the 2015 model. *:(X10$^-$4)

- A number of the predictors in a 2011 language model changed the sign of their coefficients in a 2015 language model for predicting age, or were no longer relevant for predicting gender.
- The kind of adjectives associated with positive emotions and LGBTQ issues changed in 2014 vs 2011.

| Concept | Year | Context words |
|---|---|---|
| LGBTQ issues | 2011 | strippers, conservative, pedophile, subjective, shocking |
| | 2014 | coping, passed, balance, yayy, finally, harmony |
| Positive emotion | 2011 | fagazy, bomb, totally, awesomeness, tight, fly |
| | 2014 | kickback, swag, winning, dontgiveafuck, bi*ch, thicka** |

Table 2: Context words for concepts in the language of Twitter 2011 vs. 2014, selected among the words with the highest relative norm difference in distances from the concepts in the first column, between the two sets of Twitter embeddings.

## Conclusion

1. Language models degrade over time!
2. The language of social media posts can be used to study semantic drift over short periods of time, even from small datasets.
3. There is a need to disentangle which differences are due to the changing use of language from the ones due to changes in topics and trends.
4. Domain adaptation techniques can potentially resolve diachronic performance differences

## References

[1] Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP). pages 1146–1151.

[2] Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635-E3644.

[3] Daniel Preotiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. 2012. Trendminer: An architecture for real time analysis of social media text