Phân tích mức độ tương tác của người dùng đối với bài dịch Reddit trên RedditVN

Leopard Bethemel

Ngày 30 tháng 1 năm 2019

Mục lục

Tóm tắt nội dung

RedditVN là phiên bản Việt Nam trên Facebook của Reddit. Các thành viên trong nhóm sẽ đóng góp các bài dịch từ những chủ đề trên Subreddit. Bản báo cáo này sẽ phân tích mức độ hưởng ứng của độc giả đối với bài dịch.

Chú ý: Tác giả không khuyến khích các bạn chạy dịch giả theo số tương tác mà bỏ bê chất lượng dịch thuật. Đồng thời cũng không coi thường việc dịch theo sở thích hay cảm hứng. Tác giả chỉ muốn phần nào xem xét hiện tượng "flop" và giúp các bạn có biện pháp để giảm thiểu khả năng bị chìm.

1 Thu thập dữ liệu

Dữ liệu được cung cấp ở đây:
https://mega.nz/#F!GMlQnKqZ!8FTQYeh8aprKPNE1qQT3ng

Trong thư mục `redditStats`, file `20170101T000000-20190101T000000.csv` chứa toàn bộ bài viết trong group RedditVN từ khi thành lập đến cuối năm 2018. Các bài đăng sau mốc trên vẫn còn tương tác ít nhiều, sẽ ảnh hưởng đến phân tích nói chung. Do đó, chúng ta chỉ tập trung vào khoảng thời gian trước năm 2019.

Ngoài ra, trước khi tool trợ giúp dịch thuật ra đời, bài dịch trong nhóm nói chung không theo một format thống nhất, dẫn đến không ít post không dẫn kèm link gốc. Hoặc hy hữu một số link gốc bị xóa hay khóa.

Sau khi loại bỏ các bài không thể truy ra ID Reddit hoặc có ID hỏng từ 28636 post, chúng ta còn 22869, chiếm 79,86% tổng số. Mẫu dữ liệu này có thể coi là hợp lệ để nghiên cứu.

2 Subreddit được dịch nhiều nhất

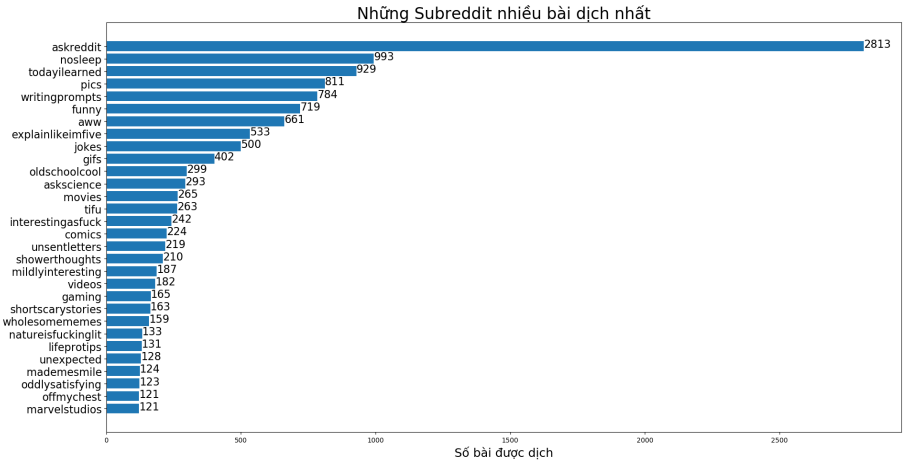
2.1 Trung bình

Trong vòng hai năm qua (2017-2018), đã có 1417 subreddit khác nhau được dịch. Trung bình mỗi sub có 16,13 bài.

- 783 sub (tức 55,26%) có trên 1 bài viết.
- 226 sub (tức 15,95%) có trên 10 bài viết.
- 38 sub (tức 2,46%) có trên 100 bài viết.

2.2 Tổng

Chúng ta xem xét 30 subreddit được dịch nhiều nhất.



Hình 1: 30 Subreddit được dịch nhiều nhất

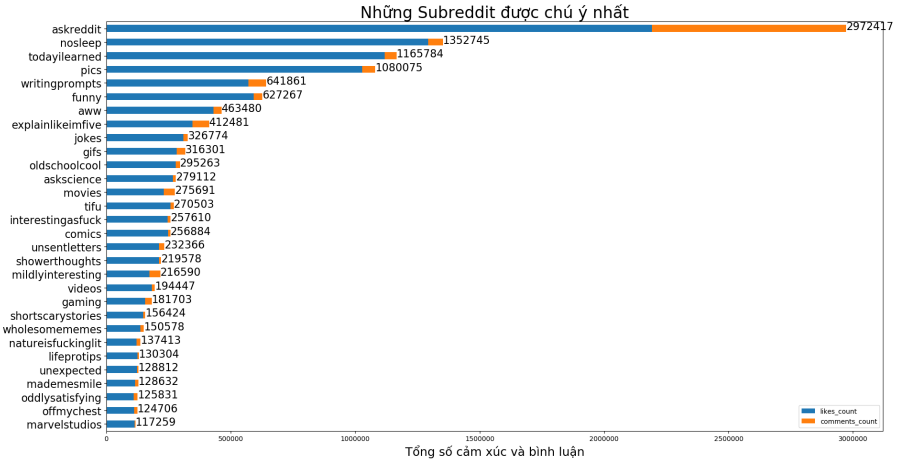
Dễ thấy chủ đề được các dịch giả RedditVN chọn nhiều nhất là **askreddit**, với xấp xỉ 2800 bài, nhiều gấp 2.5 lần so với **nosleep**, **todayilearned** đứng ở vị trí thứ thứ hai và ba. Các sub như **pics**, **writingprompts**, **funny** và **aww** cũng được dịch không ít, khoảng 660-810 bài cho mỗi sub. 23 subreddit còn lại có số bài viết phân bố trong khoảng [120, 480], nhưng một nửa là dưới 200 bài.

Nhận xét 1. *Chúng ta hầu hết tập trung dịch vào 7 subreddit kể trên.*

3 Subreddit được chú ý nhất

Để tiện cho so sánh, chúng ta coi như comment và like có giá trị ngang nhau và sẽ đếm tổng của số lượt like (kể cả cảm xúc) và số comment.

3.1 Tổng



Hình 2: 30 Subreddit được chú ý nhất

askreddit tiếp tục giữ vị trí độc tôn với tổng số lượt thích (cũng như cảm xúc khác) và bình luận cao chót vót gần 3 triệu, gấp đôi á quân **nosleep** quen thuộc. **todayilearned** và **pics** bám đuổi sát sao ở số 3 và số 4. Các vị trí tiếp theo, với tổng số like và comment dưới 650 nghìn, nhìn chung không khác mấy so với hình (1). Ta có thể kết luận rằng:

Nhận xét 2. ***"Sub được dịch nhiều"*** *tương đương với* ***"bài sẽ nhiều tương tác"****.*

Một điều đặc biệt là *tổng số comment* trong **askreddit** nhiều hơn hẳn so với các sub khác ít nhất 11 lần. Nói cách khác:

Nhận xét 3. ***askreddit*** *có tiềm năng thu hút bình luận cao hơn các subreddit khác rõ rệt.*

3.2 Trung bình

So về trung bình (bảng (1)), **askreddit** tuy thuộc top trong các sub "nhiều chữ" như **nosleep**, **todayilearned**, **tifu** nhưng thua xa các sub thiên về hình ảnh trực quan như **pics**, **aww**, **funny**, **gifs**. Do đó:

Nhận xét 4. *Độc giả nhìn chung ngại đọc nhiều chữ.*

Một giải pháp khắc phục là sử dụng hình minh họa bắt mắt để kéo thêm tương tác. (Cần phải có thêm một nghiên cứu khác về mức độ tương tác và hình ảnh để xác định tính hiệu quả của phương pháp này)

Subreddit	Số phản ứng trung bình	Số bình luận trung bình
tổng thể	**874**	**111**
askreddit	779	278
pics	1595	73
aww	1693	71
funny	1429	73
todayilearned	614	76
gifs	1473	88
jokes	861	65
nosleep	349	66
interestingasfuck	1275	75
writingprompts	360	44
comics	1245	73
wholesomememes	1679	77
explainlikeimfive	431	86
oldschoolcool	861	44
mildlyinteresting	1314	64
mademesmile	2005	67
videos	1158	118
oddlysatisfying	1712	73
movies	650	168
unexpected	1431	88
tifu	595	96
natureisfuckinglit	1113	63
blackpeopletwitter	1461	140
wtf	1020	135
beamazed	1710	75
memes	1777	90
natureismetal	979	111
showerthoughts	520	79
gaming	677	79
hmmm	988	41

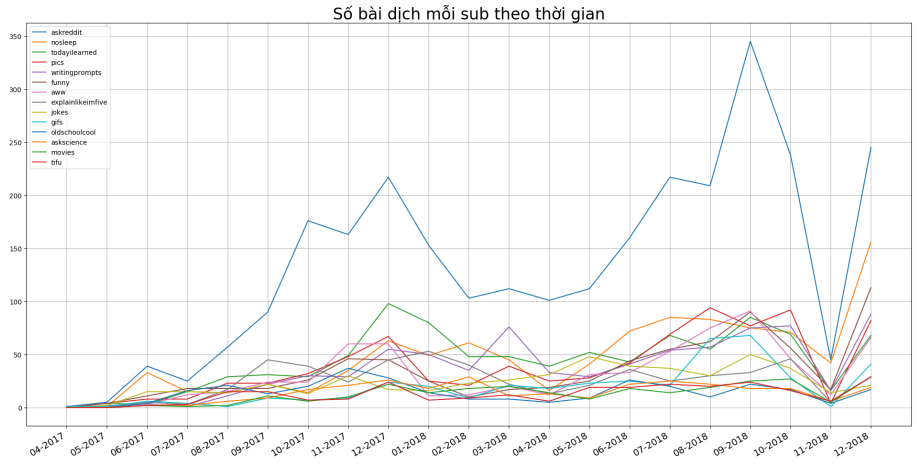
Bảng 1: Số phản ứng và bình luận trung bình của 30 sub nổi nhất và tổng thể

4 Xu thế đăng bài theo thời gian

Vì không có điều kiện gì khác, chúng ta sẽ giả sử rằng mọi bài viết trong dữ liệu đều được duyệt và các mod/admin không thiên vị khung giờ hay người đăng.

4.1 Mỗi tháng của từng sub

Bài dịch đầu tiên được đăng vào 18:59:17+0700 ngày 28/04/2017. Chúng ta sẽ quan sát số lượng post của 15 chủ đề được dịch nhiều nhất từ tháng 4/2017 đến tháng 12/2018.



Hình 3: Số lượng bài viết của 15 sub được dịch nhiều nhất, chia theo từng tháng

Tuy số lượng tăng giảm không đều nhưng có thể thấy **askreddit** luôn chiếm vị trí số 1 trong danh sách những sub được chọn để dịch. Nhờ vào nhận xét (2), ta có thể suy ra **askreddit** sẽ nhận được lượng tương tác cực kỳ lớn.

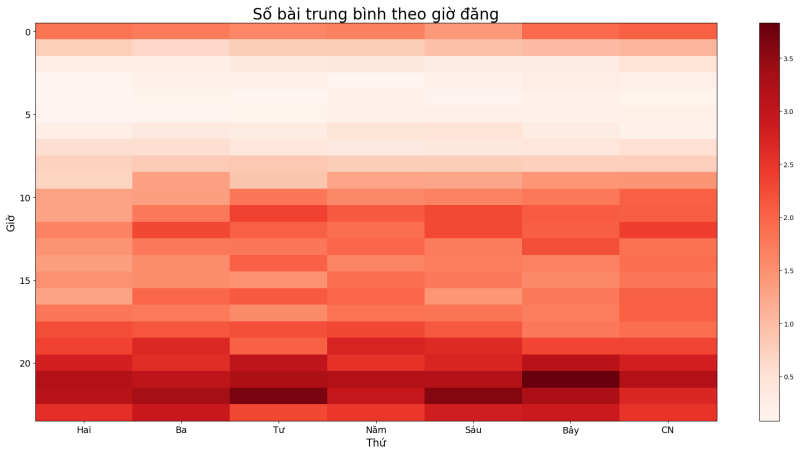
Mặc dầu vào cuối tháng 11, số lượng các bài giảm đột ngột xuống dưới 50 nhưng tháng tiếp theo đã kịp khôi phục. Nguyên nhân vẫn chưa được xác định, song có thể phỏng đoán là do cuộc "Đại Thanh Trừng" khiến số lượng thành viên sụt giảm nghiêm trọng. (Cần kiểm chứng)

4.2 Theo khung giờ

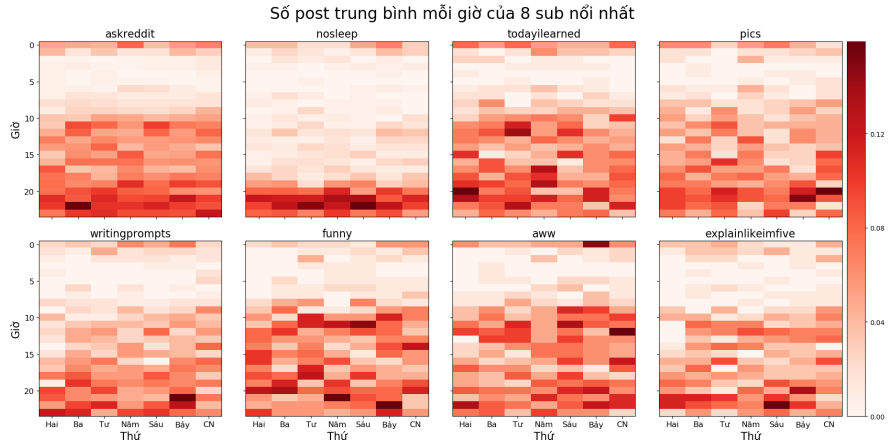
4.2.1 Tổng thể

Dựa vào dải màu đậm dưới cùng (hình (4)), chúng ta suy ra dịch giả có thói quen đăng bài nhiều vào chiều và đặt biệt là giờ tối sau 20 giờ cho tới nửa đêm. Cảnh tượng chen chúc giành sự chú ý của người đọc ắt hẳn không tránh khỏi.

Trái lại, khoảng 1 giờ tới 8 giờ lại vắng tanh. Bài post vào giai đoạn này hứa hẹn sẽ bớt tranh chấp hơn.



Hình 4: Tần số post bài theo giờ tổng thể



Hình 5: Tần số post bài theo giờ của 8 sub nổi nhất

4.2.2 Chi tiết

Các sub nổi nhất nói chung đều tuân theo quy luật của tổng thể, ngoại trừ **nosleep** hay xuất hiện vào giờ ngủ nghỉ 21 tới 24 giờ. Điều này khá dễ hiểu bởi tính chất ma quỷ hù dọa của **nosleep**, người dịch sẽ không muốn nhát ma người khác vào ban ngày ban mặt.

Ngoài ra, các sub khác đều tuân theo quy luật tổng thể. Nhưng chỉ có **askreddit** là đăng khá đồng đều giữa các khung giờ.

5 Mức độ tương tác

Một khi đã biết xu hướng đăng bài của dịch giả, hãy cùng xem độc giả hưởng ứng thế nào.

Từ phần (3), chúng ta thấy rằng số lượng comment so với số like là không đáng kể, trừ **askreddit**. Vì vậy, để xem xét xu hướng phản ứng, chúng ta chỉ nên tính số lượng like. Chú ý: Dữ liệu like sẽ bao gồm mọi cảm xúc khác như phẫn nộ, haha, v.v.

5.1 Thống kê tổng quát

Tổng số like: 19976881.
Số like trung bình mỗi post: 873.54
Số like cao nhất: 9481
Bài nhiều like nhất:
https://www.facebook.com/groups/redditvietnam/permalink/710564989340906/

- 20408 bài (89,24%) đạt 100 like trở lên.
- 15463 bài (67,62%) đạt 300 like trở lên.
- 12188 bài (53,29%) đạt 500 like trở lên.
- 6981 bài (30,53%) đạt 1000 like trở lên.
- 2555 bài (11,17%) đạt 2000 like trở lên.
- 305 bài (1,33%) đạt 4000 like trở lên.
- 29 bài (0,13%) đạt 6000 like trở lên.
- 2 bài (0,01%) đạt 8000 like trở lên.

Có ít hơn 5% số bài viết đạt 4000 like trở lên, nên để khảo sát mật độ tập trung của like, chúng ta chỉ cần chú ý vào những bài viết 4000 like trở xuống.

Dễ thấy thông thường like tập trung ở 100-200 rồi giảm dần. Mô hình hóa phân bố này cho ta nhiều lựa chọn ban đầu: exponential, gamma, lognormal. Kết hợp với test Kolmogorov-Smirnov, chúng ta thấy exponential và lognormal là hợp nhất. Mặt khác, exponential không tạo ra một "đỉnh" núi ở giữa phân bố như lognormal. Cho nên chúng ta sẽ chọn mô hình lognormal với $p-value = 8.85521235143507e-24 < 1\%$, bất chấp $p-value$ của exponential nhỏ hơn.

5.2 Theo giờ đăng

5.2.1 Tổng thể

Ô màu trên sơ đồ (8) càng đậm thì càng được nhiều like. Nhìn tổng quan, hầu hết sơ đồ có màu đỏ vừa ở mức 600 like trở lên, tức là group RedditVN có lượng tương tác lớn.

Trái ngược với xu hướng đăng bài, các bài đăng vào khoảng 4 đến 10 giờ sáng nhìn chung được độc giả tương tác đáng kể. Những post buổi chiều và tối thì ít dần và thấp nhất là khuya tới sáng sớm 3 giờ.

Nhận xét 5. *Nên dàn xếp đăng bài vào giờ sáng để tránh chen chân nhau.*

5.2.2 Đối với các sub nổi nhất

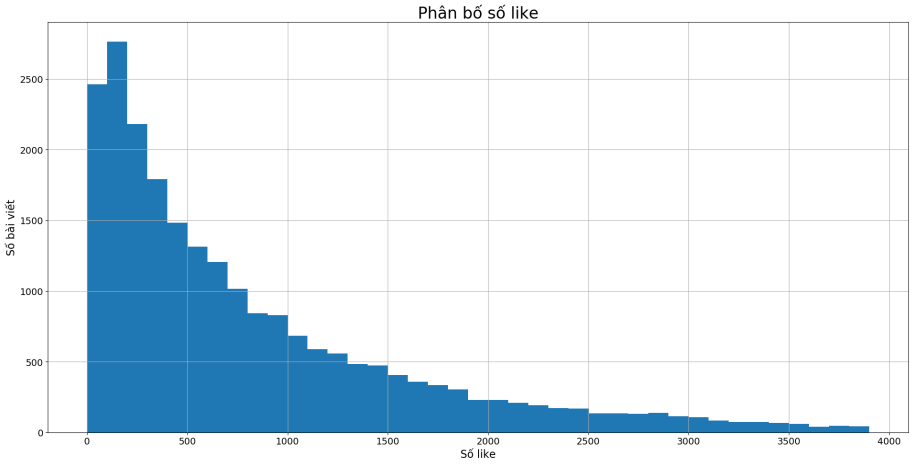
Chúng ta sẽ tiếp tục mổ xẻ mức tương tác đối với từng sub nổi nhất. **askreddit** tuy có số like trung bình theo giờ thấp (dưới 800), song mức tương tác lại trải đồng đều ban ngày và ban đêm cả tuần, trừ hai trường hợp cá biệt 3 giờ sáng thứ hai và 6 giờ sáng thứ ba. Tức **askreddit** không hoàn toàn tuân theo xu hướng tổng thể (hình (8)). Nhìn theo hướng khác, tương tác của **askreddit** là ổn định. Ngoài ra, **writingprompts** và **explainlikeimfive** cũng theo xu hướng như thế.

Mặc dù **nosleep** thường xuyên được post vào ban đêm và có số like khá khẩm hơn **askreddit**, số phản ứng lại đến nhiều hơn với những bài đăng ban ngày vào ban ngày. Quy tắc không thể hiện rõ nhưng có vài điểm cá biệt vào hai, tư và sáu. Phải chăng các độc giả sẽ ít sợ nút like vào ban ngày hơn?

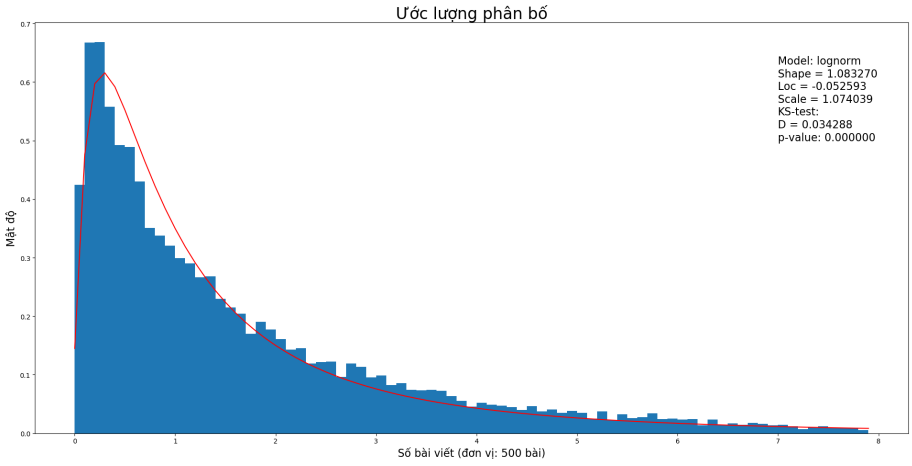
Mỗi sub còn lại ít nhiều gì cũng có số like tương đồng bất kể giờ đăng. Riêng **pics**, **funny** và **aww** tiếp tục được nhận tương tác nhiều hơn các sub "tường chữ" một cách rõ rệt.

6 Dịch giả RedditVN

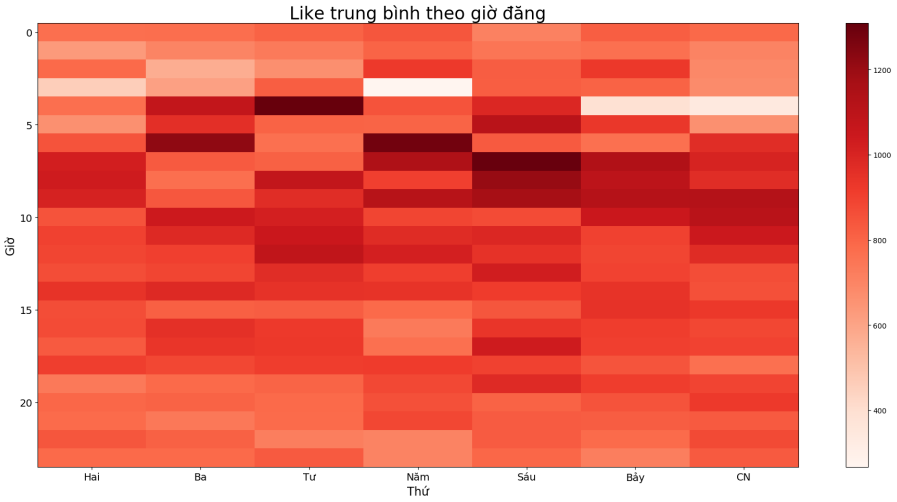
Cùng điểm một số số liệu thống kê về những chú kiến thợ của RedditVN.



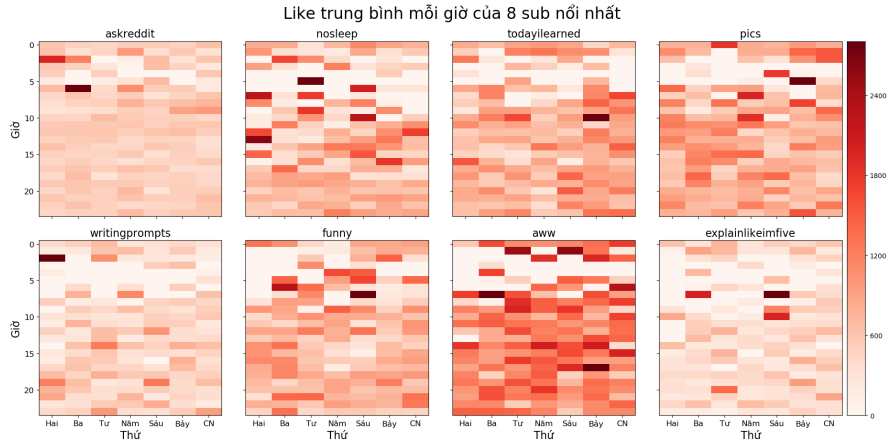
Hình 6: Phân bố like



Hình 7: Mô hình phân bố like



Hình 8: Like trung bình của post tính theo giờ đăng



Hình 9: Like trung bình tính theo giờ đăng của 8 sub nổi nhất