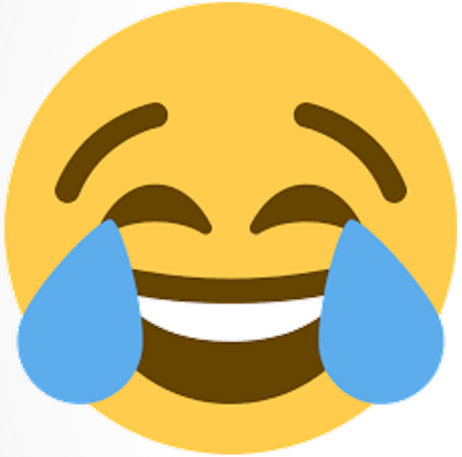
Sense Exploring of Emojis with Word2Vec Model and Using Emojis to Modify Short-text Sentiments Classification

Leyao Tan
120120910074
2022/05/31

01

Introduction



Motivation

Emojis are commonly used in social media platforms. but they are usually dropped from the dataset.

- Intuitively, emojis are strong sentiment signals
- They may be helpful for improving sentiment analysis models with social media text,
- and therefore be beneficial on research issues in marketing

Motivation

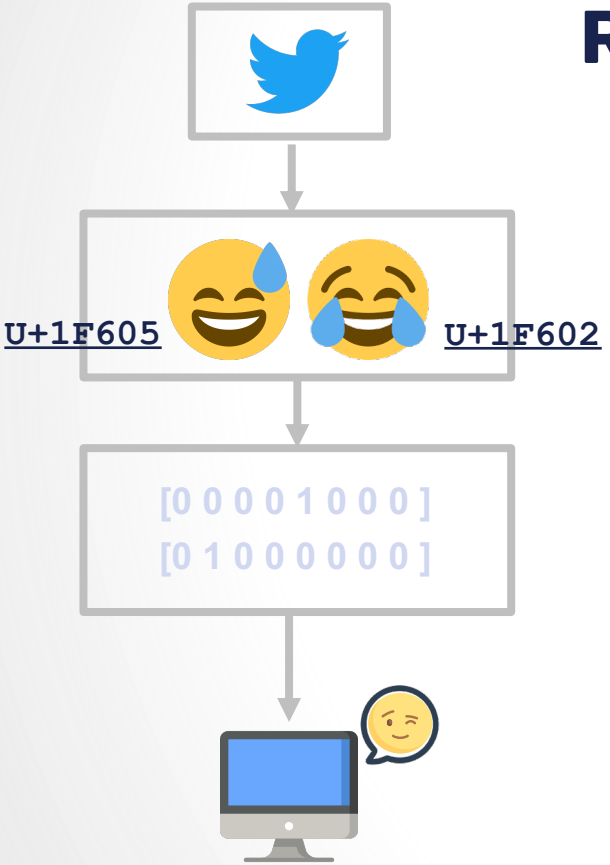
Is the description-based method effective on conveying affective factors?

№	Code	Browser	CLDR Short Name
1	U+1F600	😀	grinning face
2	U+1F603	😃	grinning face with big eyes
3	U+1F604	😄	grinning face with smiling eyes
4	U+1F601	😁	beaming face with smiling eyes
5	U+1F606	😆	grinning squinting face
6	U+1F605	😅	grinning face with sweat
7	U+1F923	🤣	rolling on the floor laughing
8	U+1F602	😂	face with tears of joy

[0 0 0 0 1 0 0 0]

Figure 4: Emoji analogy exmaples. Notice that the seemingly "correct" emoji often appears in the top three closest vectors, but not always in the top spot (furthest to the left).

Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., & Riedel, S. (2016



Research Question

RQ1: Representing emojis with Tweets

RQ2: Understand the semantic of emojis in different contexts, e.g., among languages

RQ3: Using emojis to improve sentiments classification

02

Dataset

Data Collecting

6

62

Emoji glyphs

joy	😀😂😃😄😆😇😉😊🙂😌😍😎😏🌞☺️😘😜😝😛😹😸😹😻😽❤️💖💕😁❤️
anger	😬😠😐😑😡😡😖😤💪
disgust	💩
fear	😅😨😧😱😰😰😿
sad	😔😕😞😫😩😢😥😪😓😭😿💔
surprise	😳😯😵😲

Unicode code points

joy	U+1F600, U+1F602, U+1F603, U+1F604, U+1F606, U+1F607, U+1F609, U+1F60A U+1F60B, U+1F60C, U+1F60D, U+1F60E, U+1F60F, U+1F31E, U+263A, U+1F618 U+1F61C, U+1F61D, U+1F61B, U+1F63A, U+1F638, U+1F639, U+1F63B, U+1F63C U+2764, U+1F496, U+1F495, U+1F601, U+2665
anger	U+1F62C, U+1F620, U+1F610, U+1F611, U+1F620, U+1F621, U+1F616, U+1F624 U+1F63E
disgust	U+1F4A9
fear	U+1F605, U+1F626, U+1F627, U+1F631, U+1F628, U+1F630, U+1F640
sad	U+1F614, U+1F615, U+2639, U+1F62B, U+1F629, U+1F622, U+1F625, U+1F62A U+1F613, U+1F62D, U+1F63F, U+1F494
surprise	U+1F633, U+1F62F, U+1F635, U+1F632

Table 1: Selected emoji and their Unicode code points

Wood, I. D., & Ruder, S. (n.d.). *Emoji as Emotion Tags for Tweets*. 5.

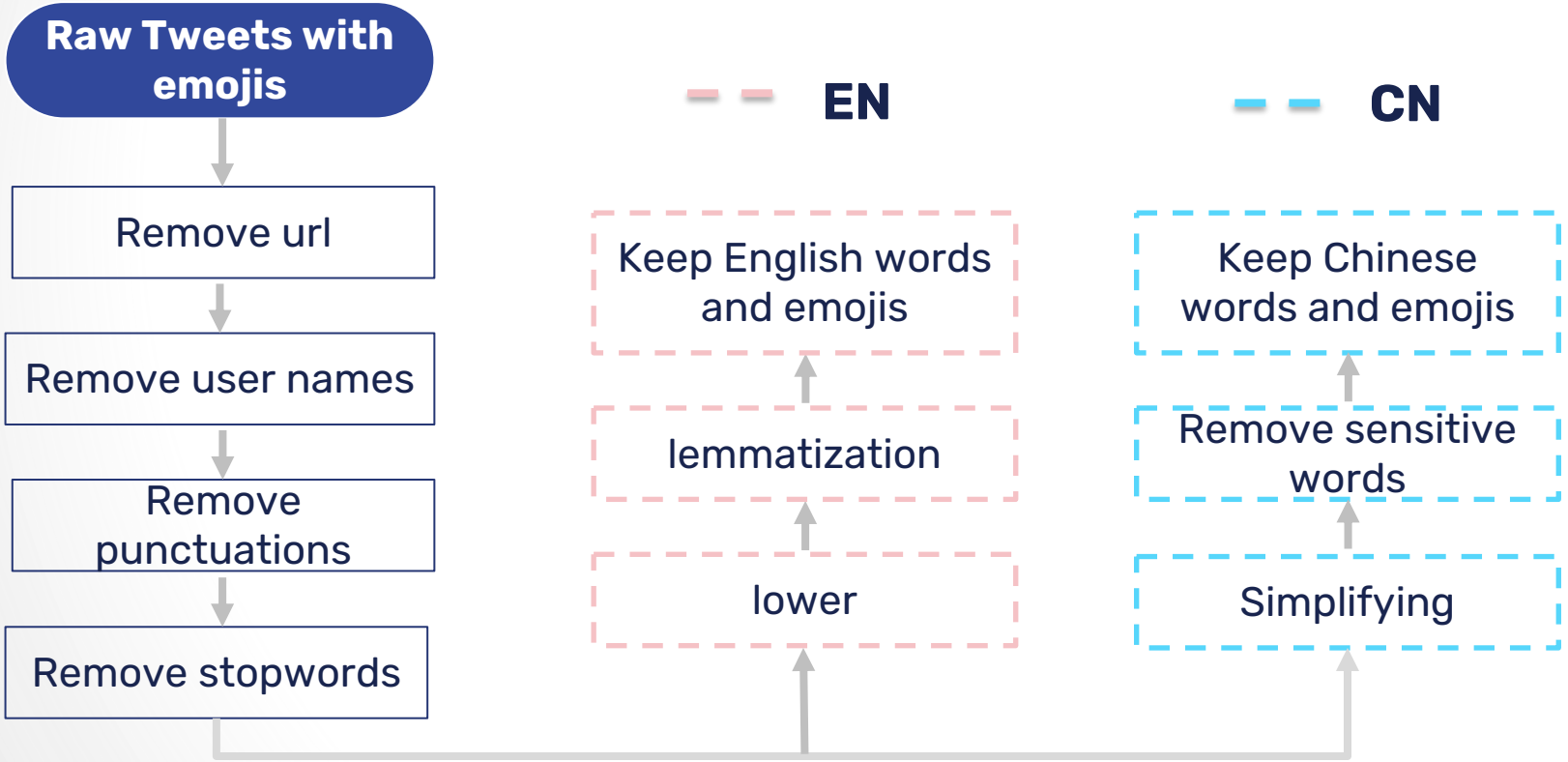
Query

- Period: 2022-0415-2022-0528
- Filter: retweet or media
- Query: tweets that contain at least one of the 62 emojis

EN 367,396 + CN 184,582

551,978

Data Cleaning



Data Cleaning

EN 367,396 → **160,041**

1. remove url
2. remove user names
3. remove punctuations
4. remove stopwords
5. lower the words
6. lemmatization
7. keep only english characters and emojis

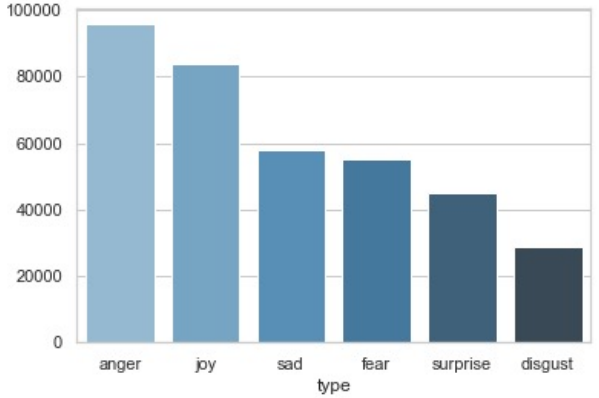
CN 184,582 → **179,925**

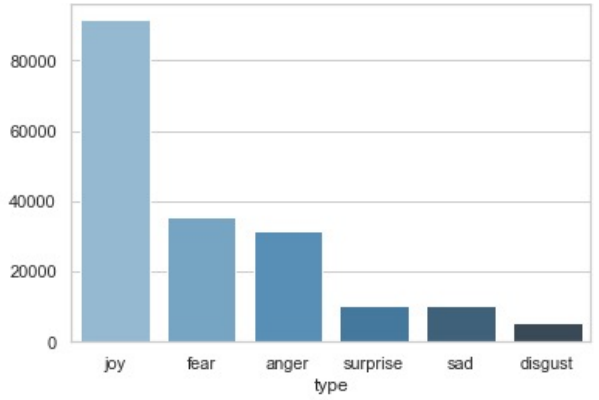
1. remove url
2. remove user names
3. cut the words
4. remove stopwords (en & zh)
5. transform to simplified Chinese
6. remove punctuations (en & zh)
7. remove sensitive Chinese words
8. remove non-chinese words

Data Cleaning

EN 367,396 → **160,041**

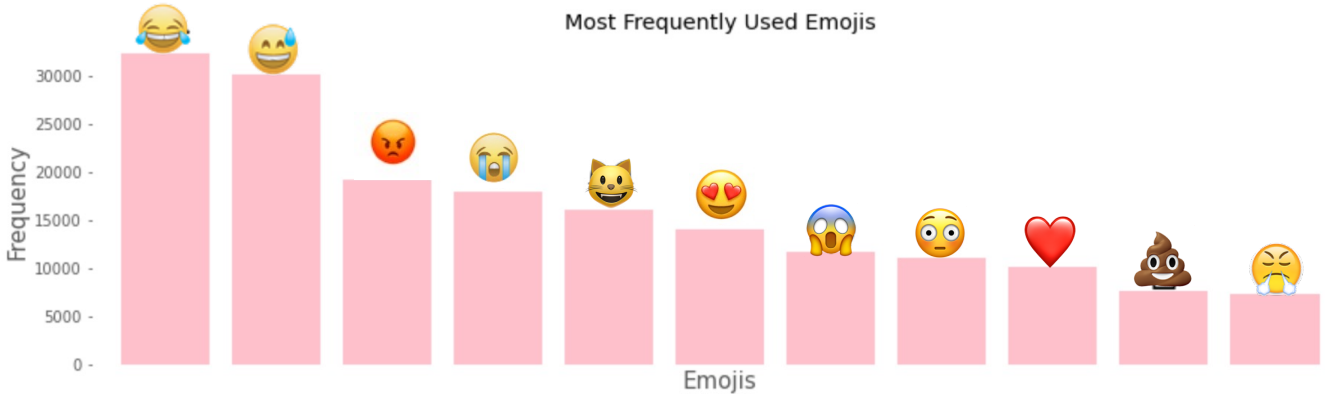
CN 184,582 → **179,925**



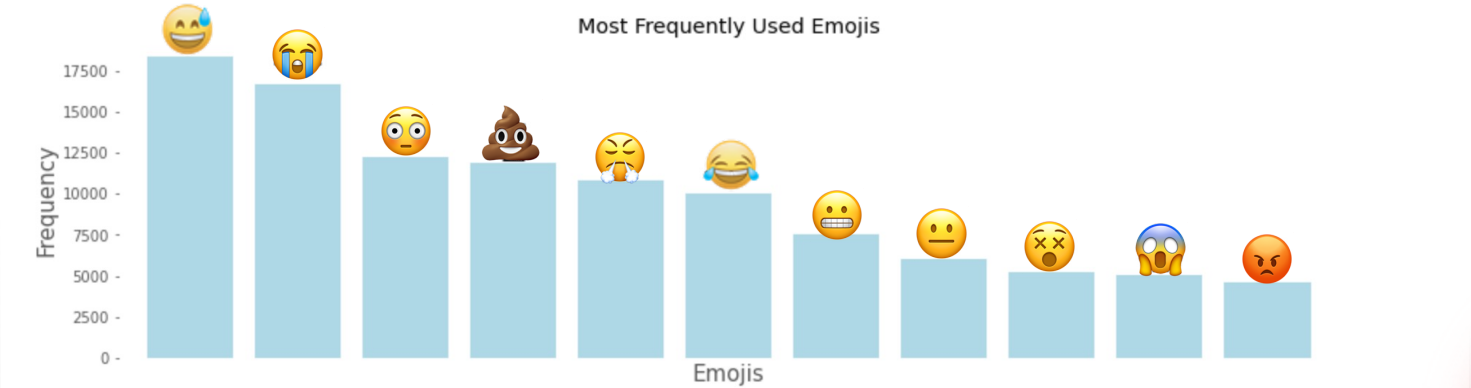


Exploring Data

CN



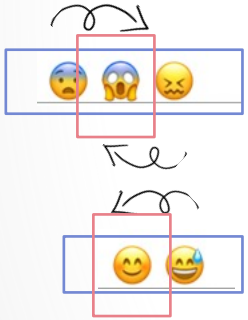
EN




03
Represent emojis

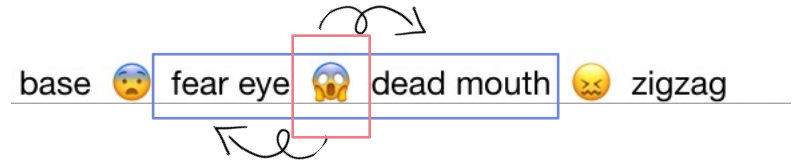
Word Embedding

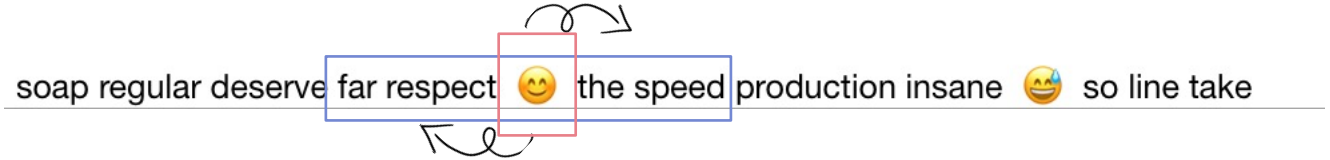
Emojis



num_features = 60
Window size = 5
Cbow model

Emojis+text





num_features = 300
Window size = 15
Skipgram model

Word Embedding: Emojis

EN

- The most similar emoji pairs?

Emoji	Most similar emojis
😂	('😀', 0.9360405206680298), ('💀', 0.9042941927909851), ('😕', 0.8956789374351501), ('🙃', 0.8887747526168823), ('🤣', 0.8830084800720215)
😭	('🥺', 0.7425079941749573), ('😐', 0.7375286817550659), ('🤨', 0.7003530263900757), ('😄', 0.6826454401016235), ('☹️', 0.6647094488143921),
😡	('🌈', 0.8902453184127808), ('📸', 0.8781907558441162), ('😃', 0.8713417053222656), ('🤪', 0.8629465103149414), ('😫', 0.8613309264183044)
😰	('💘', 0.984454333782196), ('😰', 0.9711648225784302), ('🆘', 0.9670186638832092), ('☑️', 0.9496175646781921), ('🧯', 0.9322992563247681)
💕	('😘', 0.9101722836494446), ('💗', 0.9008756279945374), ('📌', 0.8778780698776245), ('🌸', 0.8776981234550476), ('💕', 0.8737573027610779)
😜	('😝', 0.9149106740951538), ('😆', 0.8982559442520142), ('🌎', 0.8856846690177917), ('🤤', 0.8847092390060425), ('😫', 0.8774237036705017)
💩	('🤡', 0.8676602840423584), ('🤝', 0.8398205041885376), ('🤎', 0.8028566241264343), ('🍆', 0.7870866060256958), ('🤮', 0.7760229706764221)

Table 5: Ten Most Similar Emoji Pairs Based on Jaccard Similarity

Emoji Pair	😃😀	💕💙	💕💕	💚💕	💕💚	😭😭	🎉🎈	😄😃	😆😜	😆😃
Similarity	0.60	0.57	0.56	0.52	0.52	0.50	0.50	0.50	0.48	0.47

Wijeratne, S., Balasuriya, L., Sh[illegible]A., & Doran, D. (2017).

Word Embedding: Emojis

CN

- The most similar emoji pairs?

Emoji	Most similar emojis
😂	('🤣', 0.8511608839035034), ('🤝', 0.7738974094390869), ('😅', 0.7089278101921082), ('👏', 0.6739214062690735), ('👍', 0.582727313041687)
😭	('💤', 0.7611076831817627), ('🐍', 0.5911054015159607), ('😦', 0.5841449499130249), ('🥺', 0.5437110066413879), ('😢', 0.5362480878829956)
😡	('🐂', 0.8700661659240723), ('❓', 0.7024880647659302), ('👎', 0.6896219849586487), ('🤮', 0.6372858881950378), ('🤬', 0.6215999722480774),
😰	('🐗', 0.7079207897186279), ('🌑', 0.6555851697921753), ('😰', 0.631428599357605), ('🤫', 0.6080030202865601), ('😩', 0.5364194512367249)
💕	('🍼', 0.807707667350769), ('👶', 0.6825952529907227), ('💐', 0.6243771910667419), ('🍓', 0.6016101241111755), ('🔊', 0.5767290592)
😜	('👴', 0.6747037768363953), ('🌊', 0.6414020657539368), ('😔', 0.5879266858100891), ('➕', 0.5531941652297974), ('👻', 0.547359049320221)
💩	('🚘', 0.8539419770240784), ('👎', 0.7785320281982422), ('🤜', 0.6942105293273926), ('👆', 0.634078323841095), ('🤮', 0.6230592727661133)

Word Embedding: Emojis+Text

EN

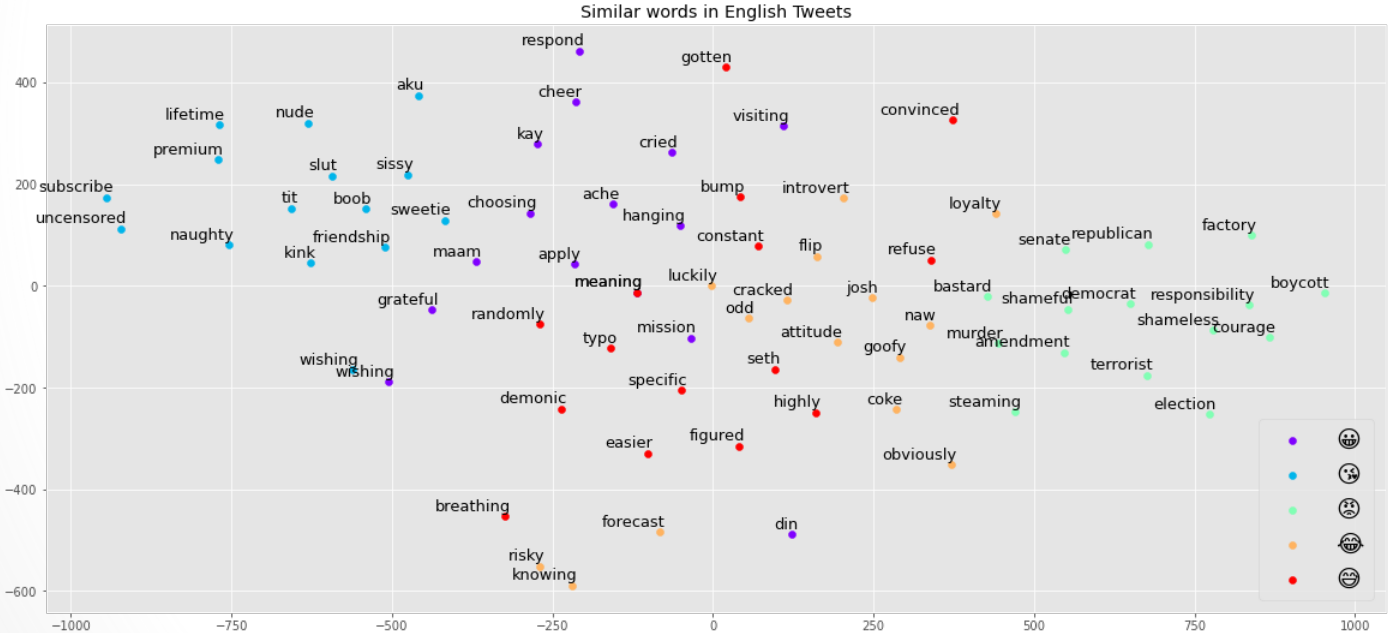
- The most similar text tokens of emojis?

Emoji	Most similar text tokens
😂	('knowing', 0.3973119258880615), ('naw', 0.3947875201702118), ('goofy', 0.3873693346977234), ('josh', 0.3872291147708893), ('attitude', 0.3867221772670746)
😭	('instantly', 0.4060744345188141), ('dry', 0.3985782265663147), ('cringing', 0.3970121145248413), ('had', 0.3934396505355835), ('outfit', 0.3757479190826416)
😡	('amendment', 0.47141000628471375), ('shameful', 0.463161826133728), ('responsibility', 0.45262280106544495), ('democrat', 0.44781386852264404), ('shameless', 0.43762820959091187)
😰	('plea', 0.6059387922286987), ('compact', 0.5996894836425781), ('bug', 0.5943024158477783), ('past', 0.5849696397781372), ('ca', 0.5556657314300537)
🌞	('sunny', 0.6493528485298157), ('wishing', 0.6011807918548584), ('bright', 0.5642414093017578), ('sunshine', 0.5567154288291931), ('filled', 0.5429767370223999)
😜	('classic', 0.487289160490036), ('slut', 0.4687976539134979, ('awhile', 0.45238691568374634), ('sandwich', 0.4442844092845917), ('naughty', 0.4438404440879822)
💩	('worthless', 0.43902328610420227), ('steaming', 0.4276275038719177), ('crap', 0.41803112626075745), ('garbage', 0.4112972021102905), ('poop', 0.3946281969547272)

Word Embedding: Emojis+Text

EN

- Visualizing of similar pairs of emojis and text tokens



Word Embedding: Emojis+Text

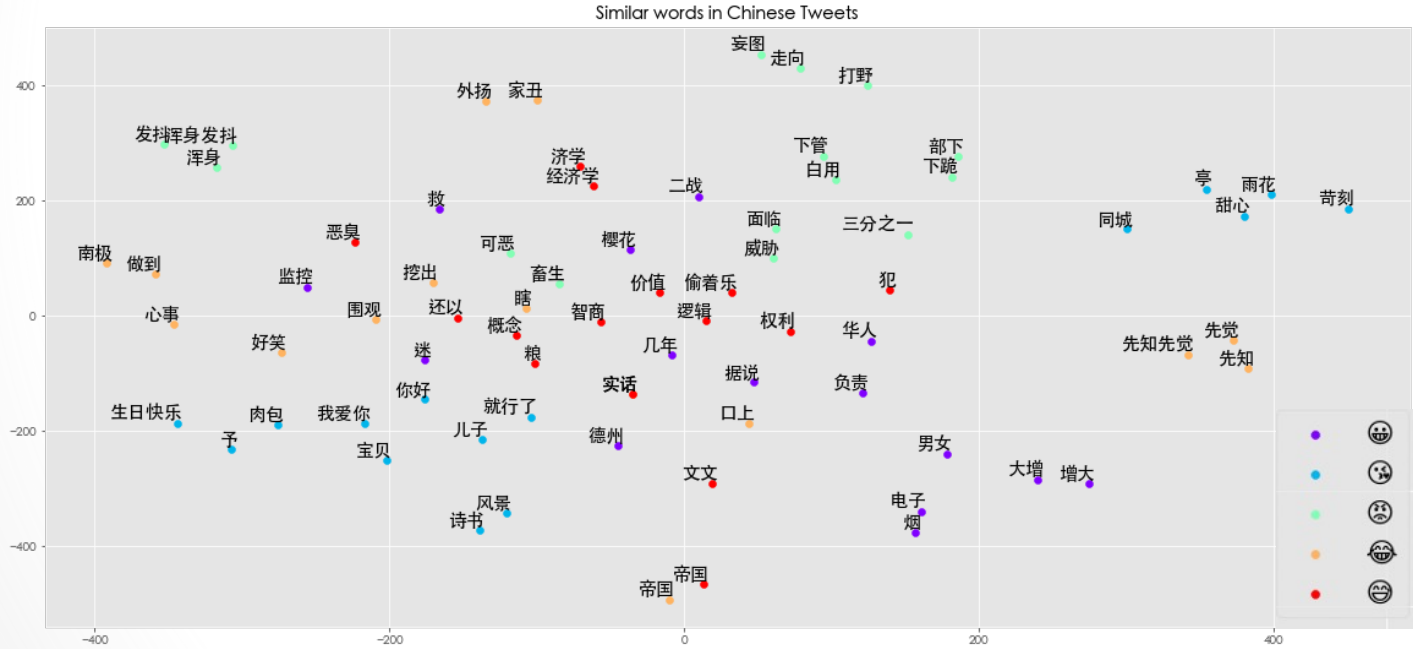
CN • The most similar text tokens of emojis?

Emoji	Most similar text tokens
😂	('南极', 0.4719521105289459), ('做到', 0.4268992245197296), ('先觉', 0.363065242767334), ('外扬', 0.35706448554992676), ('好笑', 0.354017436504364)
😭	('肉包', 0.4347709119319916), ('呜呜', 0.39678218960762024), ('委屈', 0.389453649520874), ('扭来扭去', 0.3862420916557312), ('面如', 0.3726605176925659)
😡	('浑身发抖', 0.368216335773468), ('白用', 0.33962294459342957), ('可恶', 0.33646467328071594), ('发抖', 0.332061231136322), ('部下', 0.3207872807979584)
😨	('宪政', 0.49934372305870056), ('优越', 0.48226654529571533), ('博爱', 0.4657065272331238), ('发烧', 0.4393883943557739), ('病根', 0.4294809103012085)
🌞	('早上好', 0.4115558862686157), ('太阳', 0.4089129865169525), ('晒', 0.39826592803001404), ('阳', 0.3389401137828827), ('真好', 0.3219277560710907)
😜	('满脸', 0.43915051221847534 ('复刻', 0.39226189255714417), ('不准', 0.3786528706550598), ('高技', 0.37621140480041504), ('老头', 0.33869895339012146)
💩	('屎', 0.488468736410141), ('大便', 0.4037770628929138), ('吃屎', 0.3799259662628174), ('百丑图', 0.3389178216457367), ('拉屎', 0.3383464217185974),

Word Embedding: Emojis+Text

CN

- Visualizing of similar pairs of emojis and text tokens



Word Embedding: Analogy Task

EN

Task	Most similar emojis results	Most similar text token results
😂+😅=	('🙃', 0.9031633138656616), ('😁', 0.893641471862793), ('😥', 0.8903817534446716)	('revenge', 0.47016441822052), ('bell', 0.4490945637226105), ('dislike', 0.4458730220794678)
💩+😡=	('🤡', 0.9046067595481873), ('🤮', 0.8951444029808044), ('🍆', 0.8924004435539246)	('democrat', 0.5384275913238525),('shameful', 0.5239399075508118),('amendment', 0.5211712121963501)
😀+👋=	('🙌', 0.9314199090003967), ('🤜', 0.9312358498573303), ('👌', 0.9258456826210022)	('achievement', 0.563830494881), ('congratulation', 0.558100879), ('proven', 0.556374669075)
🎉+😆=	('💗', 0.9507796764373779), ('📌', 0.9384811520576477), ('🌸', 0.9177654385566711)	('holder', 0.5954164266586304), ('illogic', 0.5848940014839172), ('profit', 0.5381441712379456)
😘-♥=	('😒', -0.09562597423791885), ('💰', -0.12754710018634796), ('🏆', -0.15446999669075012)	('law', 0.06858835369348526), ('force', 0.05814095586538315),('dive', 0.04655774310231209)

Word Embedding: Analogy Task

CN

Task	Most similar emojis results	Most similar text token results
😂+😅=	('🤣', 0.8153877258300781), ('🤝', 0.8039407134056091), ('👏', 0.6794406175613403)	('对对', 0.3655748963356018), ('智商', 0.35177725553512573), ('丑', 0.34466344118118286)
💩+😡=	('👎', 0.8597309589385986), ('🐎', 0.8537186980247498), ('🚘', 0.8050117492675781)	('屎', 0.4502533972263336), ('命', 0.4128126800060272), ('大便', 0.39148688316345215)
😀+👏=	('🤝', 0.7928808927536011), ('👍', 0.7326189875602722), ('🙏', 0.675881564617157)	('小伙', 0.5226402878761292),('平安', 0.4999120831489563),('就行了', 0.452870637178421)
🎉+😆=	('🎂', 0.550155758857727), ('🍀', 0.5165693759918213), ('💝', 0.48833051323890686)	('祝贺', 0.4266037940979004)('假期', 0.35669994354248047)('你好', 0.34463343024253845)
😘-♥=	('👹', 0.1686173677444458), ('🤡', 0.1681564897298813), ('🔪', 0.16678594052791595)	('粮食', 0.39395880699157715), ('战争', 0.3772900402545929), ('政府', 0.3613773584365845)

04

Emojis & Sentiment Analysis

Sentiment Analysis

Dataset

- 13200 observations with 0-1 sentiment labels;

	Unnamed: 0	tweets	labels
0	0	lmfaoo 😭😭😭😭😭	0
1	1	i hate this feeling 😢	0
2	2	ca n't believe i just went out in this cold to...	0
3	3	i need a new trap house so if you really fuck ...	0
4	4	so very sorry for your loss 💔	0
...	...	...	...
13195	13195	i love waking up skinny ahaha wish it lasted a...	1
13196	13196	magnificent pair of tits 😍 my cock is hard 🍆 😀	1
13197	13197	soon mamsh 😘 god will give you the best among ...	1
13198	13198	i trust u 😎	1
13199	13199	aww thanks 😁	1

13200 rows × 3 columns

Sentiment Analysis

Method 1

Only text

Accuracy:
- LSTM: 0.57
- DNN: 0.56

Method 2

Text + Emojis' descriptive names

Accuracy:
- LSTM: 0.77
- DNN: 0.78

Method 3

Text + Emojis' top 5 similar text tokens

Accuracy:
- LSTM: 0.76
- DNN: 0.75

Method 4

With pre-trained Word2Vec embeddings

Accuracy:
- LSTM: 0.74
- DNN: 0.74

References

1. Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., & Riedel, S. (2016). emoji2vec: Learning Emoji Representations from their Description. *ArXiv:1609.08359 [Cs]*.
2. Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1615–1625.
3. Feng, Y., Lu, Z., Zhou, W., Wang, Z., & Cao, Q. (2020). New Emoji Requests from Twitter Users: When, Where, Why, and What We Can Do About Them. *ACM Transactions on Social Computing*, *3*(2), 1–25.
4. Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of Emojis. *PLOS ONE*, *10*(12), e0144296.
5. Wijeratne, S., Balasuriya, L., Sheth, A., & Doran, D. (2017). EmojiNet: An Open Service and API for Emoji Sense Discovery. *ArXiv:1707.04652 [Cs]*.
6. Wood, I. D., & Ruder, S. (n.d.). *Emoji as Emotion Tags for Tweets*. 5.

Thanks for Listening

Current version of code & dataset:
https://github.com/Liagogo/Sentiments-in-Tweets-with-Emojis

😀 + 💕 = 💐