NLP and Social Media Analysis

Reshmi G Pillai

FNWI, University of Amsterdam

reshmi.g85@gmail.com

Roadmap

- Motivation
- NLP Pre-processing for Social Media
- Semantic Analysis of Social Media Data
- Data Collection – Sources, Format and Storage
- Challenges
- References

Roadmap

- Motivation
- NLP Pre-processing for Social Media
- Semantic Analysis of Social Media Data
- Data Collection – Sources, Format and Storage
- Challenges
- References

Motivation for Social Media Analysis

Popularity
megaphone
of the masses

General-purpose
Microblogs (Facebook, Twitter)

Goal- oriented services like LinkedIn, Instagram, Foursquare

Data

Heterogenous
Text, images, videos
Useful metadata
user profile, social network connections

JUL 2019

SOCIAL MEDIA OVERVIEW

BASED ON MONTHLY ACTIVE USERS OF THE MOST ACTIVE SOCIAL MEDIA PLATFORMS IN EACH COUNTRY / TERRITORY

TOTAL NUMBER OF ACTIVE SOCIAL MEDIA USERS	ACTIVE SOCIAL MEDIA USERS AS A PERCENTAGE OF TOTAL POPULATION	TOTAL NUMBER OF ACTIVE SOCIAL USERS ACCESSING VIA MOBILE DEVICES	ACTIVE MOBILE SOCIAL USERS AS A PERCENTAGE OF TOTAL POPULATION
3.534 BILLION	46%	3.463 BILLION	45%

we are social

Why Twitter ?

Popularity

Twitter has 326 million monthly active users. The number of messages sent per day is 500 million

Message Format

Users can post upto 140 characters; Brevity promotes several updates per day compared to traditional blogs

Mobile devices

80% of Twitter users access it from mobile devices; Real-time updates of daily events or opinions

Hashtags

To identify and collect messages with the same central topic

Roadmap

- Motivation
- NLP Pre-processing for Social Media
- Semantic Analysis of Social Media Data
- Data Collection – Sources, Format and Storage
- Challenges
- References

Text Normalization

- Identification and correction of orthographic errors in input text
- Reduces linguistic noise

Government confirms blast #nuclearplants #japan...don’t knw whts gona happen nw...

Why is there noise?

- Desire to save characters/keystrokes
- Social identity
- Conventions/limitations in this text sub-genre

Is it necessary?

“Lossy” translation task ?

Unique linguistic features give insights to demographic (age, gender etc.)

Relevant features might be lost in the conversion to standard English

NLP Pre-processing - Tokenizing

Tokenizer – Identify tokens(typically words) in the corpus

white space is usually a good indicator

Should punctuation be separated from words?

I love playing football, cricket etc.

football, - > football

etc. -> ?

Tokenizer for social media

- Treatment of punctuation even more complicated

  Emojis :-) :-(:-D :-x

- Usernames(starting with @), hashtags(starting with #) and URLs (links to webpages) should be treated as tokens and the symbols should be retained

Part-of-Speech Tagger

POS Tagger reads text in some language and assigns parts of speech to each word, such as noun, verb, adjective etc. The tags can be even more fine-grained like 'noun-plural'

Typically use Hidden Markov Models or Conditional Random Fields

PennTreeBank tagset (Marcus et al,1993)

Stanford log-linear POS Tagger (Toutanova et al, 2003)

POS Tagger for Twitter (1) Rittel et al, 2011

The POS-tagging accuracy drops from about 97% on newspaper text to 80% on the 800 tweets

The set of POS tags used must be extended in order to adapt to the needs of social media text

Words in these categories can be tagged with very high accuracy using simple regular expressions; they need to be considered as features in the re-training of the taggers (for example, the tags of the previous words to be tagged)

Rittel et al, 2011 used the PennTreeBank tagset to annotate 800 Twitter messages; They added a few new tags for Twitter - retweets,@usernames,#hashtags,and URLs.

Chunker-Parser

Chunker

- Detects noun phrases, verb phrases, adjectival phrases and adverbial phrases
- ''shallow'' parsers – they don't connect the detected phrases to the syntactic structure of a sentence.

Parser

- Performs syntactic analysis of sentences
- Produces parse tree
- Further used in semantic analysis/information extraction

Dependency Parser

- Extracts word pairs which are in syntactic dependency relation (e.g. verb-subject, verb-object, noun modifier)

Evaluation of Parser Performance (Rehbein and Genabith, 2007)

- ParsEval evaluation Compares the phrase structure bracketings produced by the parser(P) with bracketings of the annotated corpus (C)
- Precision M/P Number of bracketing matches with respect to number of bracketings P returned by parser
- Recall (M/C) Number of bracketing matches with respect to number of bracketings C in the corpus
- F-measure (Harmonic-mean)

NLP pre-processing Tool for Tweets – TweetNLP[1]

- Tokenization

RT @justinbieber : now Hailee get a twitter

Got #college admissions questions ? Ask them tonight during #CampusChat I'm looking forward to advice from @collegevisit http://bit.ly/cchOTk

- Features

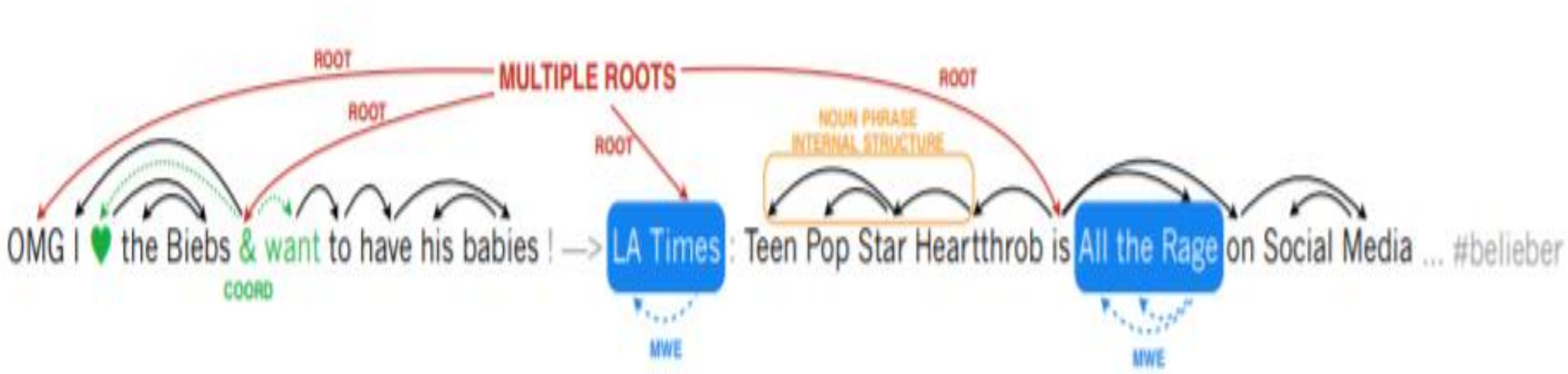
POS; shape features that recognize the retweet marker, hashtags, usernames, and hyperlinks; capitalization; and a binary feature for tokens that include punctuation

- Annotators decide MWEs

Including: proper names (Justin Bieber, World Series), non-compositional or entrenched nominal compounds (grilled cheese), connectives (as well as), prepositions (out of), adverbials (so far), and idioms (giving up, make sure)

[1] http://www.cs.cmu.edu/~ark/TweetNLP/

Parsing by TweetNLP - Example



Named Entity Recognition

- Detects names, dates, currency amounts and other entities in the text
- Identifies Person, Organization and Location and the boundaries of these phrases
- Coreference resolution- detect the noun to which a pronoun is referring to OR the different noun phrases that refers to the same entity

More on NER

- Subtasks- detecting entities
  Determining/ classifying the type of entity
  - Methods
    Statistical methods based on linguistic grammars
    Semi-supervised
    Supervised based on CRFs
      Handcrafted grammar based features
      Annotated training data
  - Evaluation measures – precision, recall, f-measure
    Sequence level or token level

Stanford NER on a Tweet

Yess! Yess! Its Official ! It is announced today that they will release the Nintendo 3DS in north America march 27 for $250

[Yess]ORG! [Yess]ORG! Its official! It is announced today that they will release the [Nintendo]ORG 3DS in north [America]LOC march 27 for $250

Yess wrongly identified as an NE (organization)

Roadmap

- Motivation
- NLP Pre-processing for Social Media
- Semantic Analysis of Social Media Data
- Data Collection – Sources, Format and Storage
- Challenges
- References

Semantic Analysis of Social Media Data

- Geo-location Identification
- Sentiment Analysis
- Event Detection
- Automatic Summarization
- Machine Translation
- Psycho-Social Analysis

Geo-Location Identification

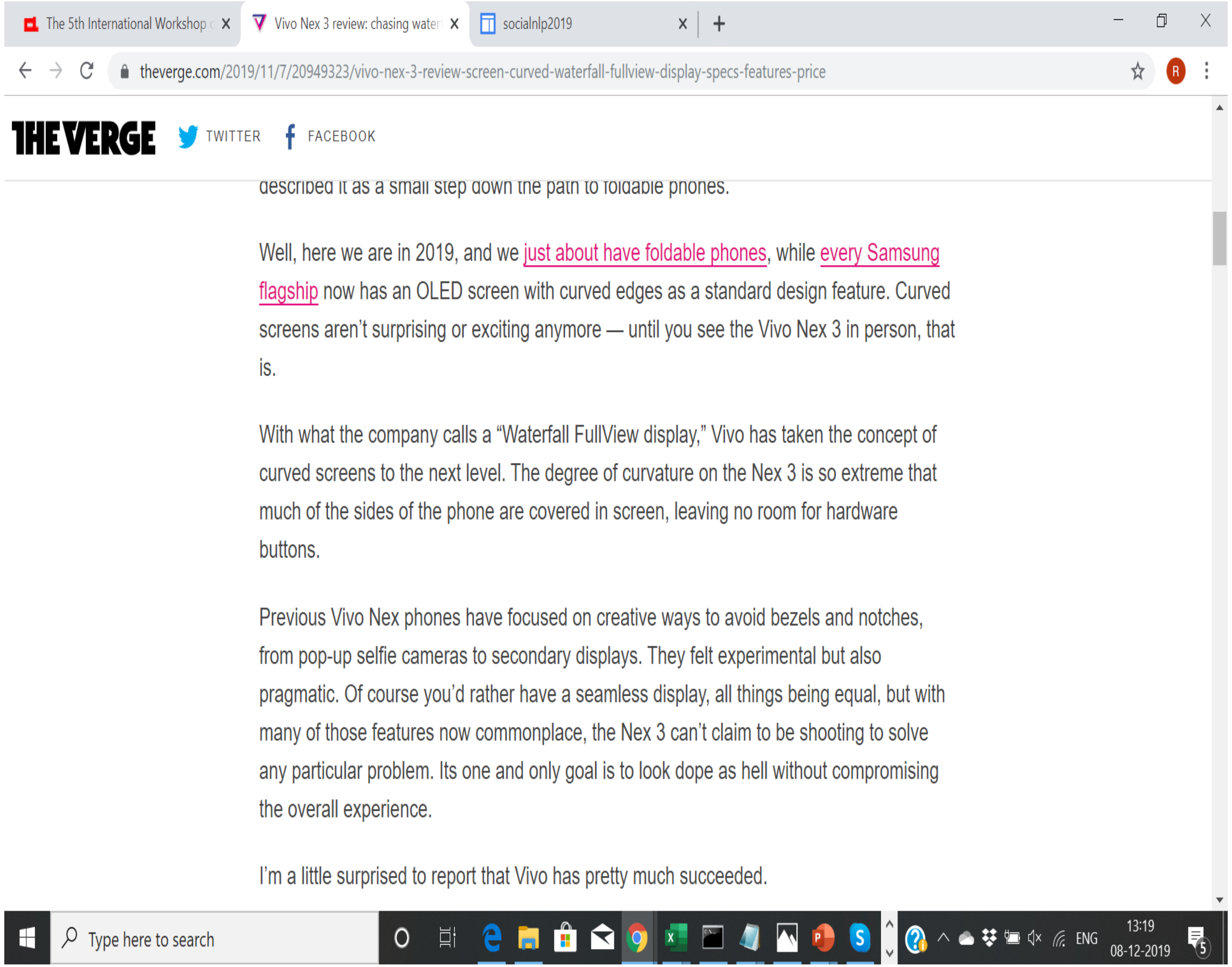
Only about 1% of Tweets are geo-tagged

Inferring a user’s location

- Statistical distribution of words in tweets to find words which have strong geo-scope
- Geographical topic models to model the language across a certain region

Sentiment Analysis

Searching for Sentiments in a Review!



described it as a small step down the path to foldable phones.

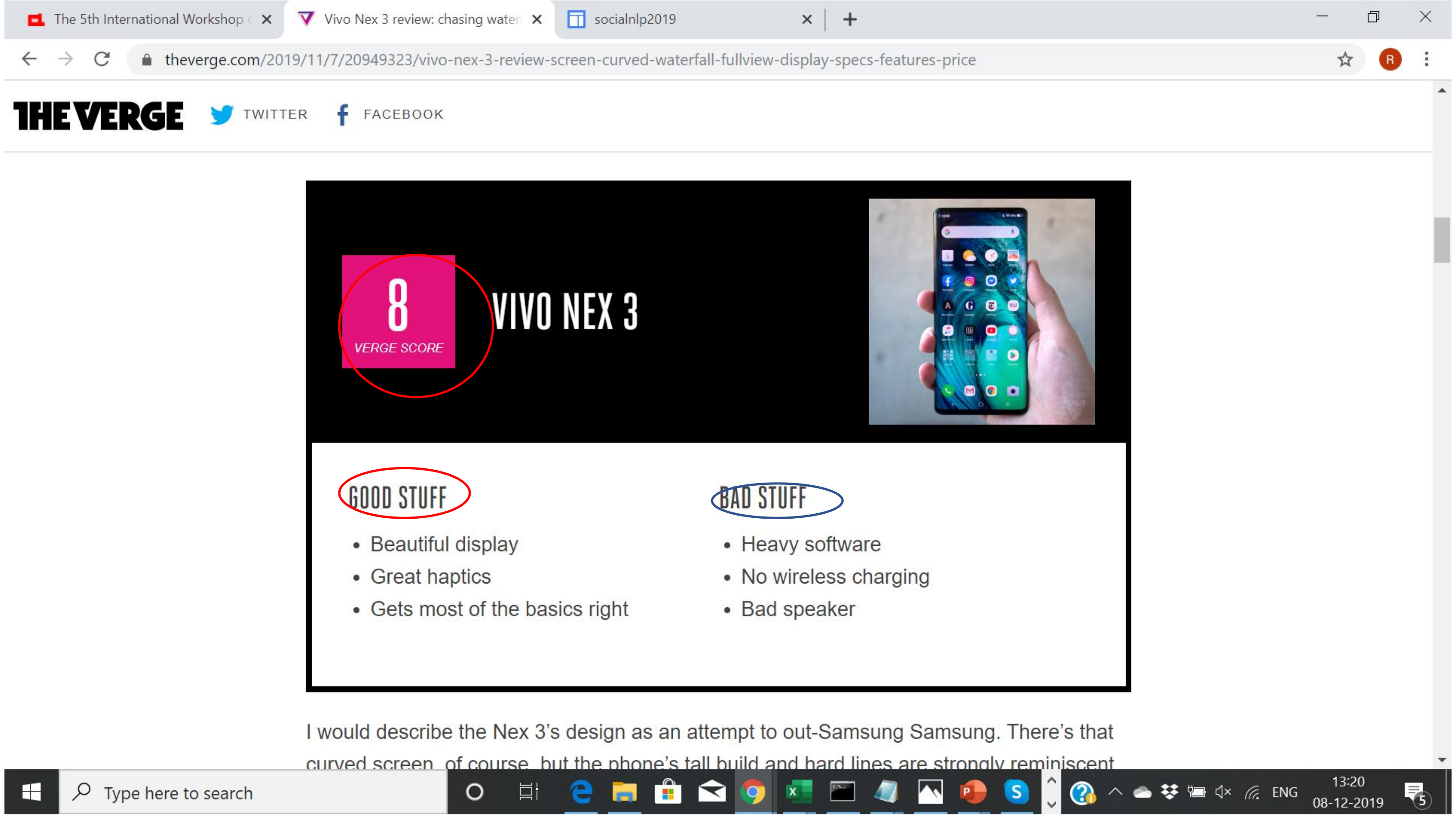
Well, here we are in 2019, and we just about have foldable phones, while every Samsung flagship now has an OLED screen with curved edges as a standard design feature. Curved screens aren't surprising or exciting anymore — until you see the Vivo Nex 3 in person, that is.

With what the company calls a "Waterfall FullView display," Vivo has taken the concept of curved screens to the next level. The degree of curvature on the Nex 3 is so extreme that much of the sides of the phone are covered in screen, leaving no room for hardware buttons.

Previous Vivo Nex phones have focused on creative ways to avoid bezels and notches, from pop-up selfie cameras to secondary displays. They felt experimental but also pragmatic. Of course you'd rather have a seamless display, all things being equal, but with many of those features now commonplace, the Nex 3 can't claim to be shooting to solve any particular problem. Its one and only goal is to look dope as hell without compromising the overall experience.

I'm a little surprised to report that Vivo has pretty much succeeded.

Much easier!!


The 5th International Workshop
Vivo Nex 3 review: chasing water
socialnlp2019
theverge.com/2019/11/7/20949323/vivo-nex-3-review-screen-curved-waterfall-fullview-display-specs-features-price
THE VERGE
TWITTER
FACEBOOK
8
VERGE SCORE
VIVO NEX 3
GOOD STUFF
Beautiful display
Great haptics
Gets most of the basics right
BAD STUFF
Heavy software
No wireless charging
Bad speaker
I would describe the Nex 3's design as an attempt to out-Samsung Samsung. There's that
Type here to search
ENG
13:20
08-12-2019

Sentiment Analysis of Social Media

- community
- another person
- user / author
- document
- sentence or phrase
- aspect (e.g. product feature)

Sentiment Analysis - Approaches

Lexicons – SentiWordNet, SentiStrength

Classification Problem

- Machine Learning Algorithms

Neural Networks

Features

- n-grams
- Stylistic features
- Social media specific features

Event detection

Event detection can be classified based on

- Event type - Unspecified or Specified

  Unspecified - Driven by emerging events, breaking news general topics that attract the attention of a large number of Twitter users; Identified using temporal patterns, sudden increased usage of specific keywords

  Specified – Known or planned social events with metadata information like location, time and performers

- Detection task – retrospective or new event detection

- Detection method - Supervised or Unsupervised

Automatic Summarization

- Less focus on individual documents
- More on how they contribute to a summary of some real-world phenomenon
- Four types
  - Update summarization
  - Network activity summarizati
  - Event summarization
  - Opinion summarization

An ideal summary

Coverage

The extracted summary can conclude every aspect of all documents

Sparsity

One sentence in the document set should be precisely represented by only a small number of summary sentences

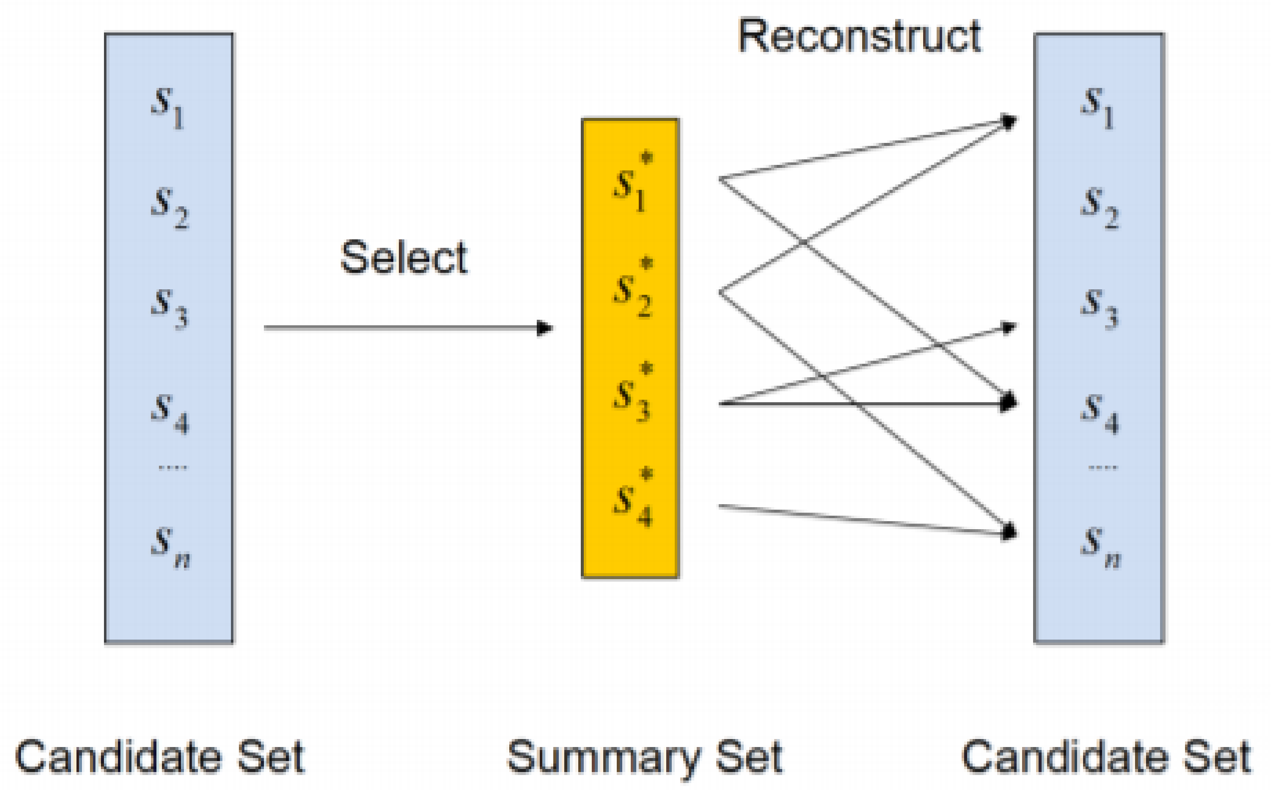
Structure

Multi-document set have one central topic and some sub-topics, indicating the summary sentences should be categorized into groups too

Diversity

To eliminate redundancy.. A good summarization finds the most obvious topic and other sub-topics that help us understand the whole document set

Summarization



Twitter summarization

Traditional summarization only considers text information

Twitter summarization techniques

Extending PageRank algorithm incorporating social properties (Duan et al, 2012; Liu et al, 2012)

Temporal and retweet information (Alsaedi, Burnap and Rana, 2016)

As a supervised classification task through mining rich social features such as temporal signal and user influence (Chang et al. 2013;2016)

Criticism

Social information used is mostly static or limited to user-level

Tweet-level network relations are unexplored

Twitter summarization based on social network and sparse reconstruction(He and Duan, 2018)

Expression consistency: Whether the tweets posted by the same user are more consistent than two randomly selected tweets?

Expression contagion: Whether the two tweets posted by friends are more similar than the two randomly selected tweets?

Machine Translation for Social Media

Informal to informal translation

- Preserve informal features such as stylistic effects, short forms (GOAAAAL -> TOOOOR)

Informal to formal translation

- Twitter users are encouraged to use short forms at word or phrase level in order to fit all the contents within the 140 characters limitation
- Most of the time they intentionally make acronyms for a group of words or a phrase
- u (informal for ‘you’ in English) -> dir (formal in German)

Sentiment preservation

- The tweet “YEEEEEESSSS!!!” contains a higher level of positive sentiment than the tweet “YES!!” The correct translation will be “JAAAAAA!!!” in German

TweetMT

TweetMT, a parallel corpus of tweets in four language pairs that combine five languages (Spanish from/to Basque, Catalan, Galician and Portuguese)

Participating teams used Statistical Machine Translation and Rule Based Machine Translation

A Case Study of Machine Translation in Financial Sentiment Analysis

Native approach Create a gold standard corpus for German from the ground up, manually annotate and cross review it, and then train the new classifier on it

Derived approach Take the English sentiment gold standard corpus, translate it (either manually or automatically) to German, and train the German classifier on it

Direct Translation Approach Use machine translation to convert the German input to English, and feed the English translations to the English classifier

Understanding the user – Modelling user personality profiles

Topic	Authors	Focus	Social Media	Factors	Analysis Tool
Personality And Traits	Kulkarni, 2018	Latent Traits	Facebook	Linguistic features	Differential analysis & correlation coefficients
	Buffone et al.,et al., 2018	Empathy and distress	Online news articles	Linguistic features	CNN-based predictive model
	Liu and Preotiuc-Pietro, 2016	Big-five personality traits	Twitter	Profile picture (colour, composition, type, demographics, expressions)	Pearson correlation between facial features and traits
	Souri, Hosseinpour and Rahmani, 2018	Big-five Personality traits	Facebook	Likes, profile, networking and posts information	Classification algorithms
	Yilun Wang, 2018	MBTI personality traits	Twitter	Bag of ngrams, POS tags and word vectors	Logistic Regression classification model
	Zamani, Buffone and Schwartz, 2018	Human trustfulness	Facebook	Ngrams(1 to 3) and LDA topics of status updates	Ridge regression

Understanding the user – Modelling user personality profiles

Psychological Disorders	Kotikalapudi, 2012	Depression	Email/chatting	Internet usage	Statistical analysis using correlation
	Moreno, 2011	Depression	Facebook	Status updates	Negative binomial regression analysis
	De Choudhury et al.,et al., 2013a	Depression	Twitter	Emotional and linguistic features	SVM classifiers
	CopperSmith, Harman, 2014a	PTSD	Twitter	Ngrams and LIWC data of Tweets	Logistic regression classifier
	Eichstaedt, Smith et al.,et al., 2018	Depression	Facebook	Linguistic, emotional, interpersonal and cognitive	Logistic regression
	De Choudhury et al.,et al., 2013c	Post-partum depression	Twitter	Social engagement, emotional and linguistic styles	Predictive model
	Coppersmith and Harman, 2014b	PTSD, Depression Bipolar and Seasonal Affective disorders	Twitter	LIWC, language models	Correlation coefficients and classification models
	Lin, Jia, Huang, 2016a	Stress	Sina Weibo	Tweet level and user level attributes	CNN based mobile App (Moodee)

Example- World Wellness Project[1] – University of Pennsylvania

Measuring psychological well-being and physical health based on the analysis of language in social media

https://www.wwbp.org/

Wordcloud from Facebook statuses of more/less extroverted people ?

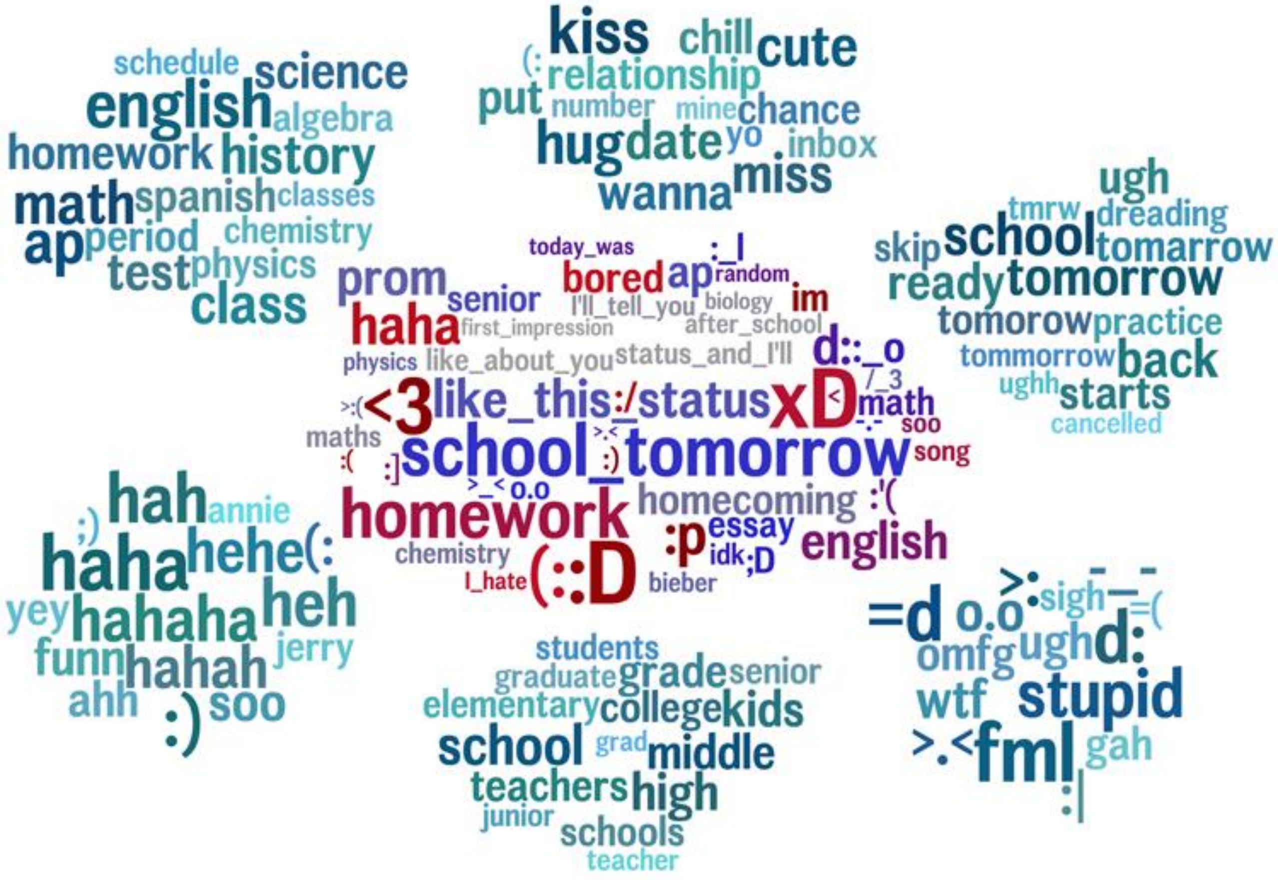
Wordcloud from Facebook statuses of more/less extroverted people?

Wordcloud from Facebook statuses of more extroverted people

Wordcloud from Facebook statuses of less extroverted people

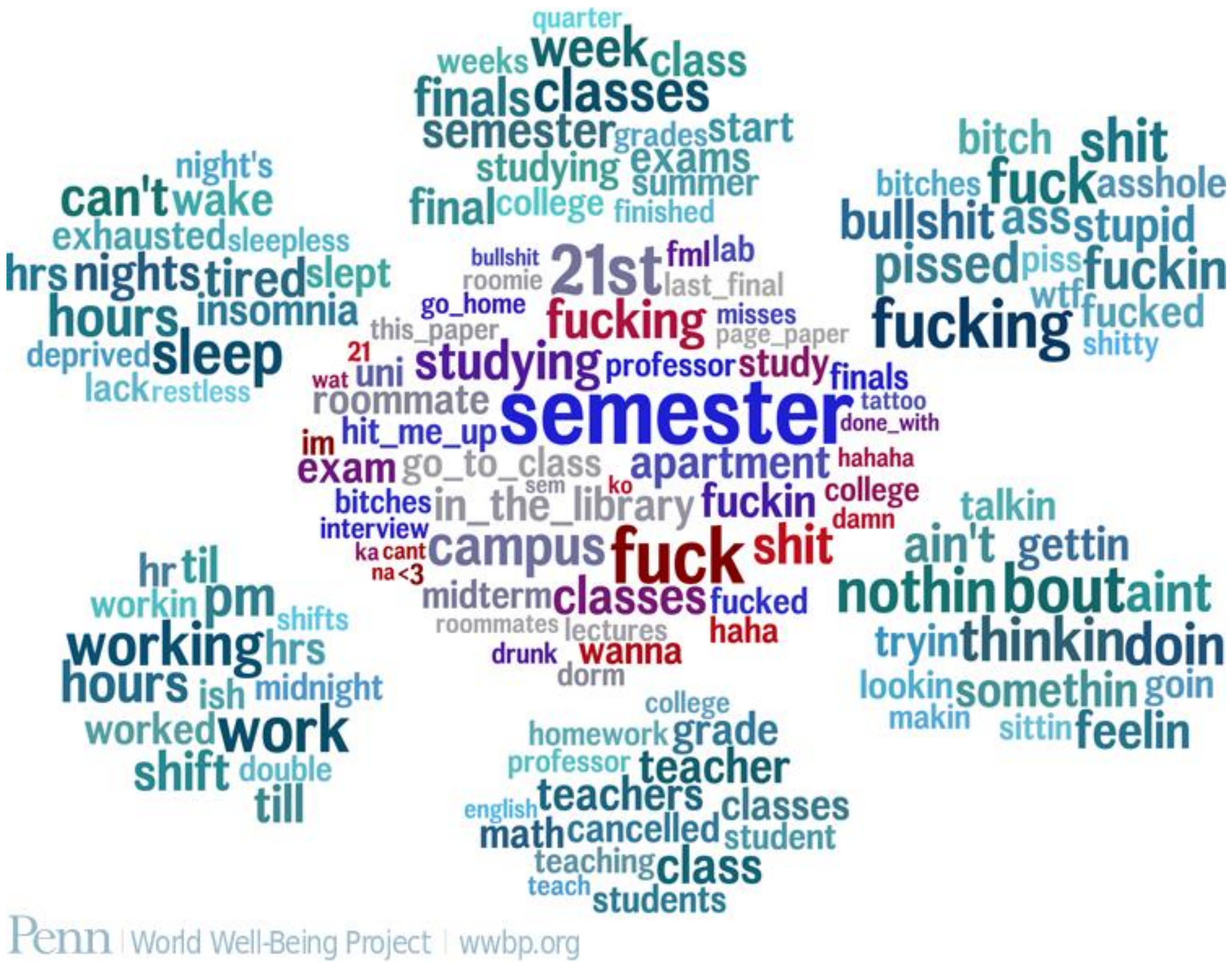
Penn | World Well-Being Project | wwbp.org

Wordcloud from Facebook statuses of age group 13-18



Penn | World Well-Being Project | wwbp.org

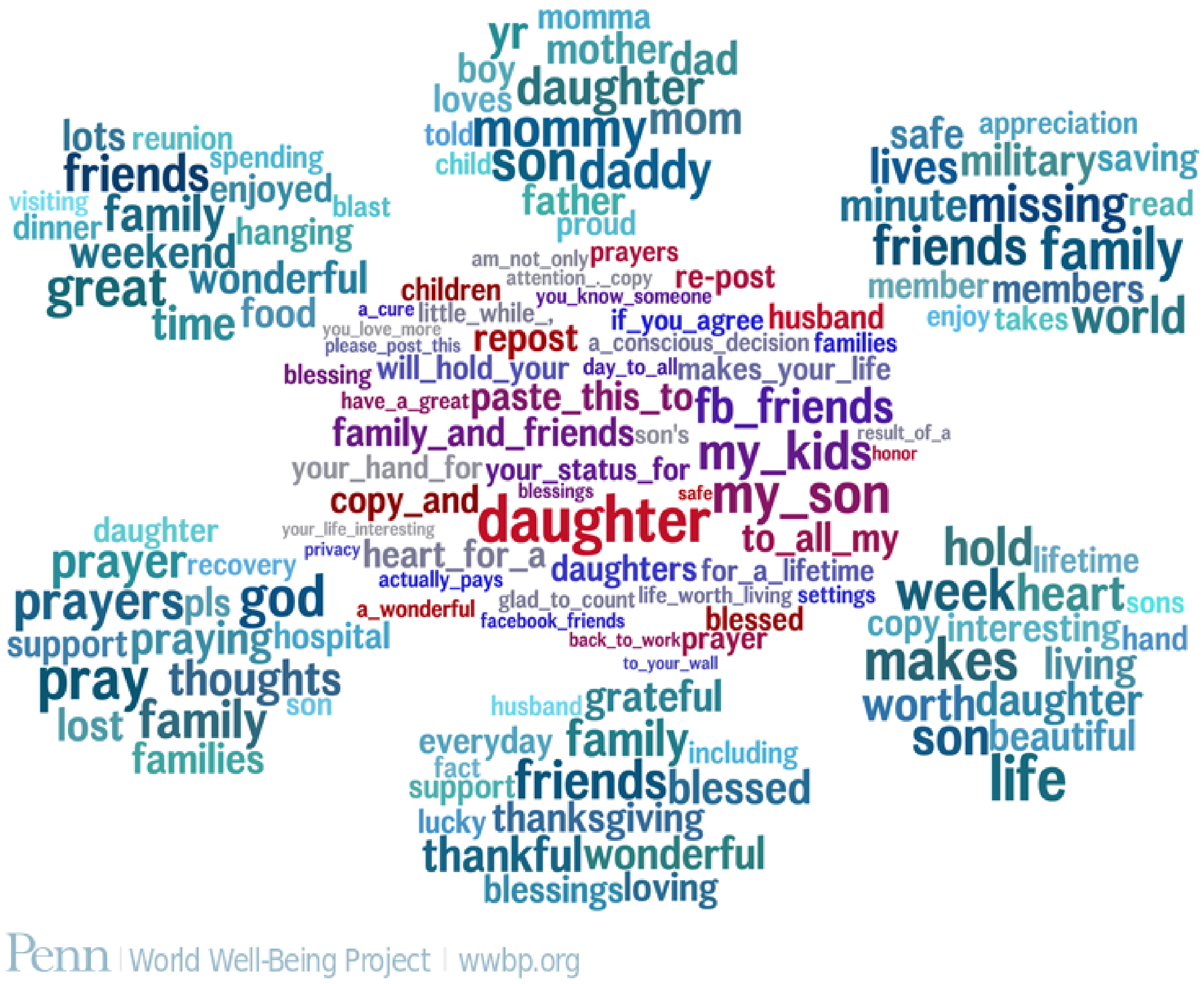
Wordcloud from Facebook statuses of age group 19-22



Wordcloud from Facebook statuses of age group 23-29

Penn | World Well-Being Project | wwbp.org

Wordcloud from Facebook statuses of age group above 30



Roadmap

- Motivation
- NLP Pre-processing for Social Media
- Semantic Analysis of Social Media Data
- Data Collection – Sources, Format and Storage
- Challenges
- References

Data collection – Sources, Format and Storage

Data scraping— scrape any type of social media (social networking media, RSS feeds, blogs, wikis, news, etc.) through easily programmable APIs

Data streaming—to access and combine real-time feeds and archived data for analytic

APIs

Search API

Query Twitter for recent Tweets containing specific keywords

It is part of the Twitter REST API v1.1 (it attempts to comply with the design principles of the REST architectural style, which stands for Representational State Transfer)

Requires an authorized application (using oAuth, the open standard for authorization) before retrieving any results from the API.

Streaming API

A real-time stream of Tweets, filtered by user ID, keyword, geographic location or random sampling.

Data collection – Sources, Format and Storage

Missing data when a piece of information existed but was not included for whatever reason in the raw data supplied.

numeric data when ‘blank’ or a missing value is erroneously substituted by ‘zero’ which is then taken (for example) as the current price

textual data when a missing word (like ‘not’) may change the whole meaning of a sentence.

Incorrect data when a piece of information is

incorrectly specified (such as decimal errors in numeric data or wrong word in textual data)

incorrectly interpreted (such as a system assuming a currency value is in $ when in fact it is in £ or assuming text is in US English rather than UK English).

Inconsistent data when a piece of information is inconsistently specified

For example, with numeric data, this might be using a mixture of formats for dates: 2012/10/14, 14/10/2012 or 10/14/2012.

Data collection – Sources, Format and Storage

Flat file - a flat file is a two-dimensional database (somewhat like a spreadsheet) containing records that have no structured interrelationship, that can be searched sequentially.

Relational database—a database organized as a set of formally described tables to recognize relations between stored items of information, allowing more complex relationships among the data items. Examples are row-based SQL databases

noSQL databases—a class of database management system (DBMS) identified by its non-adherence to the widely used relational database management system (RDBMS) model

Tweepy and MongoDB

Tweepy to download Tweets

PyMongo to

- Open a connection to MongoDB server
- Store the JSON tweets
- Query MongoDB to retrieve the tweets

Roadmap

- Motivation
- NLP Pre-processing for Social Media
- Semantic Analysis of Social Media Data
- Data Collection – Sources, Format and Storage
- Challenges
- References

Privacy and ethics

Why is there so much negativity in the social media ?

- Dissociative anonymity (“You don’t know me”)
- Invisibility (“You can’t see me”)
- Asynchronicity (“See you later”)
- Solipsistic Introjection (“It’s all in my head”)
- Dissociative Imagination (“It’s just a game”)
- Minimization of Status and Authority (“Your rules don’t apply here”)


DOGBERT CONSULTS
COMPANY RULES FORBID YOU FROM INSULTING YOUR CO—WORKERS.
DILBERT.COM SCOTTADAMSSAYS
I'LL TEACH YOU HOW TO INSULT EACH OTHER WHILE STAYING WITHIN COMPANY GUIDELINES.
8-18 2018 Scott Adams, Inc./Dist. by Andrews McMeel
THAT DOESN'T SEEM POSSIBLE.
YOU SHOULD LOOK INTO GETTING A STANDING DESK.

Focus areas to ensure a healthy social media environment

- Security and privacy of information shared on social media
- Avoiding offensive language
- Hate-speech detection

General Directions for NLP in Social Media Analysis

aclweb.org/anthology/events/socialnlp-2017/

↑up

pdf (full) bib (full) **Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media**

pdf bib **Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media**
Lun-Wei Ku | Cheng-Te Li

pdf bib abs **A Survey on Hate Speech Detection using Natural Language Processing**
Anna Schmidt | Michael Wiegand

pdf bib abs **Facebook sentiment: Reactions and Emojis**
Ye Tian | Thiago Galery | Giulio Dulcinati | Emilia Molimpakis | Chao Sun

pdf bib abs **Potential and Limitations of Cross-Domain Sentiment Classification**
Jan Milan Deriu | Martin Weilenmann | Dirk Von Gruenigen | Mark Cieliebak

pdf bib abs **Aligning Entity Names with Online Aliases on Twitter**
Kevin McKelvey | Peter Goutzounis | Stephen da Cruz | Nathanael Chambers

pdf bib abs **Character-based Neural Embeddings for Tweet Clustering**
Svitlana Vakulenko | Lyndon Nixon | Mihai Lupu

Roadmap in retrospection

- Motivation
- NLP Pre-processing for Social Media
- Semantic Analysis of Social Media Data
- Data Collection – Sources, Format and Storage
- Challenges
- References

References

[1] Z. Chong, C. Matteo, P. Alexandros, S. Thorben, and N. Jane, 2017. A Case Study of Machine Translation in Financial Sentiment Analysis

[2] A. Farzindar and D. Inkpen. 2015. Natural Language Processing for Social Media. Morgan & Claypool Publishers.

[3] Giachanou, Anastasia & Crestani, Fabio. (2016). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. ACM Computing Surveys. 49. 1-41. 10.1145/2938640.

[4] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. Smith. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, page 42--47. Stroudsburg, PA, USA, Association for Computational Linguistics, (2011)

[5] Han, B., Cook, P., & Baldwin, T. (2012). Geolocation Prediction in Social Media Data by Finding Location Indicative Words. COLING.

[6] He, R., & Duan, X. (2018). Twitter Summarization Based on Social Network and Sparse Reconstruction. AAAI.

[7] Liu, H., Yu, H., & Deng, Z. (2015). Multi-Document Summarization Based on Two-Level Sparse Representation Model. AAAI.

References

[8] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider and N.A. Smith. (2013). Improved part-of-speech tagging for online conversational text with word clusters. Proceedings of NAACL-HLT. 2013. 380-390.

[9] I. Rehbein, J. van Genabith, (2007) Evaluating Evaluation Measures, Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)

[10] Sluyter-Gäthje, H., Lohar, P., Afli, H., & Way, A. (2018). FooTweets: A Bilingual Parallel Corpus of World Cup Tweets. LREC.

[11] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (NAACL '03), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 173-180. DOI: https://doi.org/10.3115/1073445.1073478

[12] I.S. Vicente, I. Alegria, C. España-Bonet, P. Gamallo, H.G. Oliveira, E.M. Garcia, A. Toral, A. Zubiaga, and N. Aranberri, N. (2016). TweetMT: A Parallel Microblog Corpus. *LREC*.

[13] Han, B., Cook, P., & Baldwin, T. (2012). Geolocation Prediction in Social Media Data by Finding Location Indicative Words. COLING.

Software and tools

The World Wellness Project https://www.wwbp.org/

TweetNLP http://www.cs.cmu.edu/~ark/TweetNLP/

Thank You !!

Questions ??