LocAndFeel

Location-Based Tweet Retrieval and Sentimental Analysis for Turkish

Mert Emin Kalender, Havva Gulay Gurbuz, Fatma Balci, Elif Dal

May 2, 2013

Outline

- Introduction
- Tweets as Data Source
- Methodology
- Classification
- Evaluation
- Tool
- Conclusion
- References

Introduction

- Social services
  - rapid growth
  - mass adoption
- Twitter
  - microblogging service
  - messages up to 140 characters
- As of June 2012 [1]
  - more than 517 million users
  - generating billions of tweets per month

Tweets as Data Source

- Various use cases [2]
  - daily updates
  - conversations
  - information sharing
  - news reporting
  - commentary on news and events
- Valuable source of data
  - public opinion on products, services etc.
  - details about recent events, emergencies with geographical information [3]

Methodology

Gathering Data

- 1,392,556 tweets gathered
  - 696,278 positive and 696,278 negative
- Emoticons as (noisy) labels
  - :), ;), :D, :)) etc. for positivity
  - :(, :’(, :((etc. for negativity
- Data characteristics
  - average length 61 characters, 8 words
  - tweets from publicly accessible profiles
  - between April 13, 2013 and April 20, 2013

Methodology

Feature Reduction

- tweet as bag-of-words
  - a set of steps applied in terms of feature reduction

Reduction Step	Feature Size	Percent of Original
None	1,619,858	100.00%
Emoticons	1,498,479	92.50%
Usernames	1,117,533	68.98%
Links	1,027,994	63.46%
Hashes	1,016,681	62.76%
Other Reductions	598,782	36.96%
Stemming	150,696	9.30%

Classification

- Naive Bayes (NB) [4]
  - simple and reasonable in terms of performance
  - assumes features are independent
- Maximum Entropy (ME) [5]
  - similar to Naive Bayes
  - no assumption on feature independence
- Support Vector Machines (SVM) [6]
  - universal learners
  - ability to learn independent of feature space size

Evaluation

Classification Accuracies

- Test Data
  - manually marked 143 negative, 125 positive tweet
- Accuracy around 80%
  - similar to [7]
  - less than [6] (reported 90%)

Min. Frequency	NB	ME	SVM
1	79.1%	81.7%	79.4%
2	78.7%	81.3%	79.4%
3	78.7%	80.6%	79.4%
4	78.4%	80.6%	
5	78.4%	79.9%	77.6%

Evaluation

Limitations

- Use of language
  - casual
  - contains grammatical mistakes
- Labels
  - emoticons
- Stemming
  - specific stemming approach

Tool

- Web application
  - Turkish and English (Python NLTK API)
  - analysis on different result types
    - recent tweets
    - popular tweets
    - mixed
  - specifying location via Google Maps API
- Demo (link)

Conclusion

- Vast usage of Twitter
  - tweets as valuable data sources
- Emoticons as label for training data
  - more than one million tweet collected
  - cleaned and refined via feature reduction steps
  - classified with three different Machine Learning algorithms
  - achieved 80% accuracy

References

- [1]: I. Lunden. Analyst: Twitter passed 500m users in june 2012, 140m of them in us; jakarta 'biggest tweeting' city @ONLINE, July 2012.
- [2]: A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.
- [3]: D. Maxwell, S. Raue, L. Azzopardi, C. Johnson, and S. Oates. Crisees: Real-time monitoring of social media streams to support crisis management. In Advances in Information Retrieval, pages 573–575. Springer, 2012.
- [4]: B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 79–86. Association for Computational Linguistics, 2002.
- [5]: K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In IJCAI-99 workshop on machine learning for information filtering, volume 1, pages 61–67, 1999.
- [6]: T. Joachims. Text categorization with support vector machines: Learning with many relevant features. Springer, 1998.
- [7]: A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, pages 1–12, 2009.

Thanks

Questions?
Comments?