Sentiment Analysis Symposium
-
Disambiguate Opinion Word Sense via Contextonymy

Guillaume Gadek, Josefin Betsholtz, Alexandre Pauchet,
Stéphan Brunessaux, Nicolas Malandain and Laurent Vercouter

Airbus, LITIS

28 June 2017

Tweets can be very difficult to analyze...

#hashtag #veryLongAndExplicitHashtag2 poorlywritten pseudo-ENglish by @user *http://spam.url* !ponctuation;signs #hashtag3

Tweets can be very difficult to analyze...

#hashtag #veryLongAndExplicitHashtag2 poorlywritten pseudo-ENglish by @user *http://spam.url* !ponctuation;signs #hashtag3

What is the sense of *!ponctuation;signs*?

Synonyms from the Oxford dictionary?

Tweets can be very difficult to analyze...

#hashtag #veryLongAndExplicitHashtag2 poorlywritten pseudo-ENglish by @user *http://spam.url* !ponctuation;signs #hashtag3

What is the sense of *!ponctuation;signs*?

Synonyms from the Oxford dictionary? No.

Outline

Introduction

State of the art

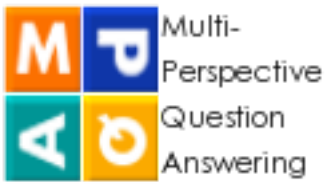
Experiment

Results

Conclusion

State of the art

Opinion as a sentiment analysis task



[Wilson et al., 2005]

[Pennebaker et al., 2001]

[Baccianella et al., 2010]

Opinion as a classification task
Stance Detection

[Andreevskaia and Bergler, 2008, Hasan and Ng, 2013]

Contextualization

Semantic relatedness of words

- use WordNet [Miller, 1995]

Contextualization

Semantic relatedness of words

- use WordNet [Miller, 1995]
- use Wikipedia [Zesch et al., 2008]

Contextualization

Semantic relatedness of words

- use WordNet [Miller, 1995]
- use Wikipedia [Zesch et al., 2008]
- $\Rightarrow$ poor results on Twitter content: different sentence structure, new vocabulary [Feng et al., 2015]

One lead amongst others: contextonyms

One lead amongst others: contextonyms

Two words are contextonyms if they occur in the same context.

One lead amongst others: contextonyms

Two words are contextonyms if they occur in the same context.

Contextonyms represent the minimal meanings of words.

One lead amongst others: contextonyms

Two words are contextonyms if they occur in the same context.

Contextonyms represent the minimal meanings of words.

Example tweet: *"What a great match!"*

One lead amongst others: contextonyms

Two words are contextonyms if they occur in the same context.

Contextonyms represent the minimal meanings of words.

Example tweet: *"What a great match!"*

Some minimal meanings of the word **match**

- match, light, fire, wood
- match, couple, date, love
- match, football, sports, game

If we know something about the context(s) of the word(s) in a tweet, we might be able to disambiguate the meaning of that tweet.

- introduced by [Hyungsuk et al., 2003]

If we know something about the context(s) of the word(s) in a tweet, we might be able to disambiguate the meaning of that tweet.

- introduced by [Hyungsuk et al., 2003]
- used for sentiment: [Serban et al., 2012]

If we know something about the context(s) of the word(s) in a tweet, we might be able to disambiguate the meaning of that tweet.

- introduced by [Hyungsuk et al., 2003]
- used for sentiment: [Serban et al., 2012]
- used for machine translation:
  [Ploux and Ji, 2003, Wang et al., 2016]

Contextonyms

Example: Contextonyms for the word **support**

Contextosets
(support, continued, foolery),
(climate, support, advocacy, preventing, change),
(support, bae, naten, kanta),
(support, tennessee, thank, trump2016)

Contextonyms

Example: Contextonyms for the word **support**

Contextosets
(support, continued, foolery),
(climate, support, advocacy, preventing, change),
(support, bae, naten, kanta),
(support, tennessee, thank, trump2016)

Contextoset: a set of words representing the context for the *target word*.

Contextonyms

Example: Contextonyms for the word **support**

Contextosets
(support, continued, foolery),
(climate, support, advocacy, preventing, change),
(support, bae, naten, kanta),
(support, tennessee, thank, trump2016)

Contextoset: a set of words representing the context for the *target word*.

Contextonyms: two words are contextonyms if they appear in the same context.

Word Embeddings, contextosets and WordNet synsets for the nearest words of **support**

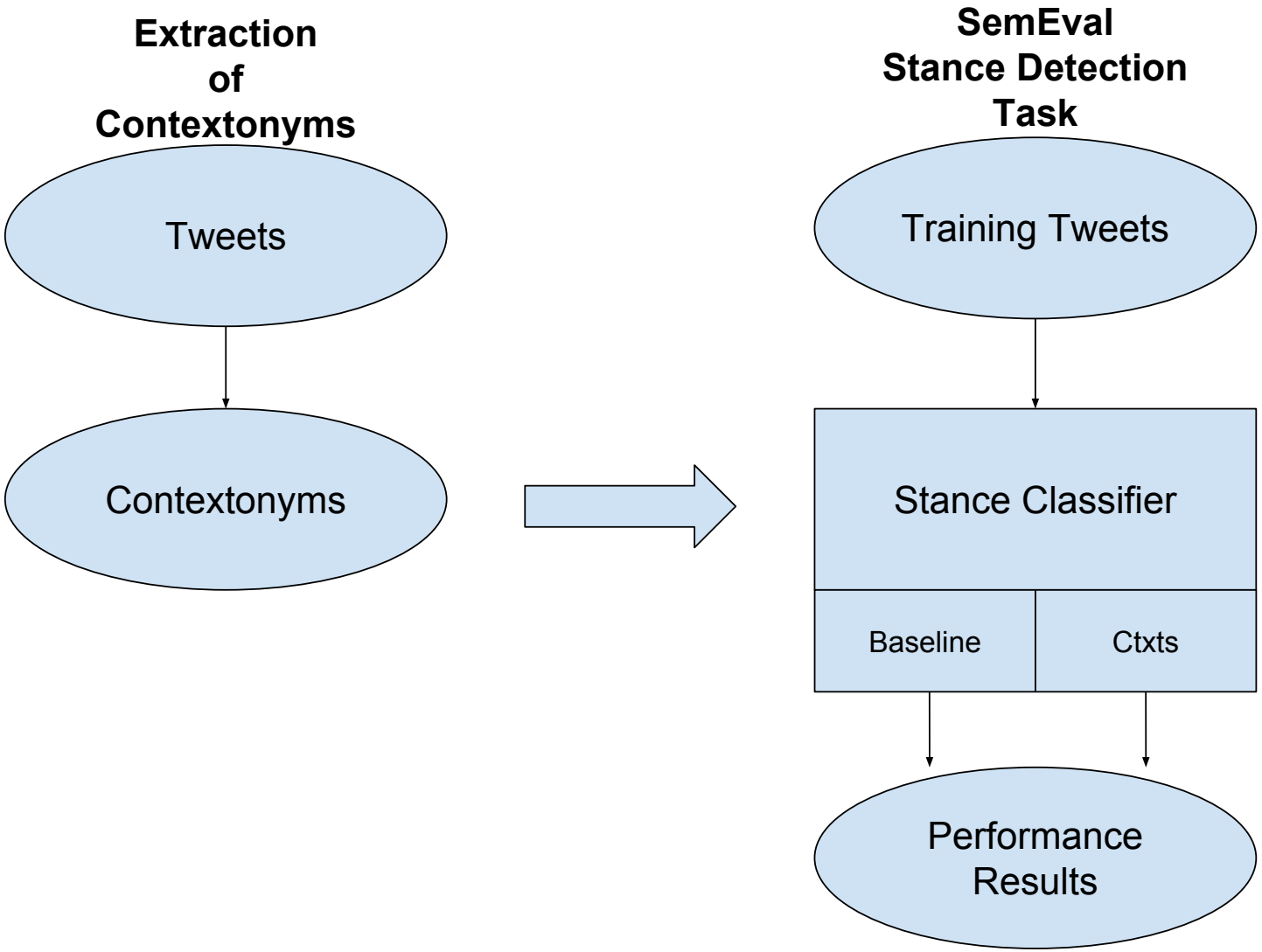
Method: Word Embeddings
supporting, supported, supports, respect, vote, encourage, voting, voted, organize, helping
Method: Synsets (extract, total:14)
(documentation, support) (support, keep, livelihood, living, bread and butter, sustenance) (support, supporting) (accompaniment, musical accompaniment, backup, support)
Method: Contextosets
(support, continued, foolery), (climate, support, advocacy, preventing, change), (support, bae, naten, kanta), (support, tennessee, thank, trump2016)

Word2Vec: [Mikolov et al., 2013]

Experiment

Aim: to improve stance detection on tweets using contextonyms

Experiment overview



Extraction-Theory

What do we need to extract contextonyms?

Source corpus

- huge (millions of tweets? more?)
- not annotated
- same language
- same topic, if possible

Contextonyms Extraction Algorithm

1. preprocess tweets
2. words co-occurrence graph
3. filter:
   - 3.1 α (filter nodes)
   - 3.2 β (filter edges)
4. k-cliques are our contextonyms.

Contextonyms Extraction Algorithm

1. preprocess tweets
2. words co-occurrence graph
3. filter:
   - 3.1 α (filter nodes)
   - 3.2 β (filter edges)
4. k-cliques are our contextonyms.

Tweet:

Hillary is the best candidate #hillary2016

Contextonyms Extraction Algorithm

1. preprocess tweets
2. words co-occurrence graph
3. filter:
   - 3.1 α (filter nodes)
   - 3.2 β (filter edges)
4. k-cliques are our contextonyms.

Tweet:

Hillary is the best candidate #hillary2016

Processed Tweet:

hillary best candidate #hillary2016

Contextonyms Extraction Algorithm

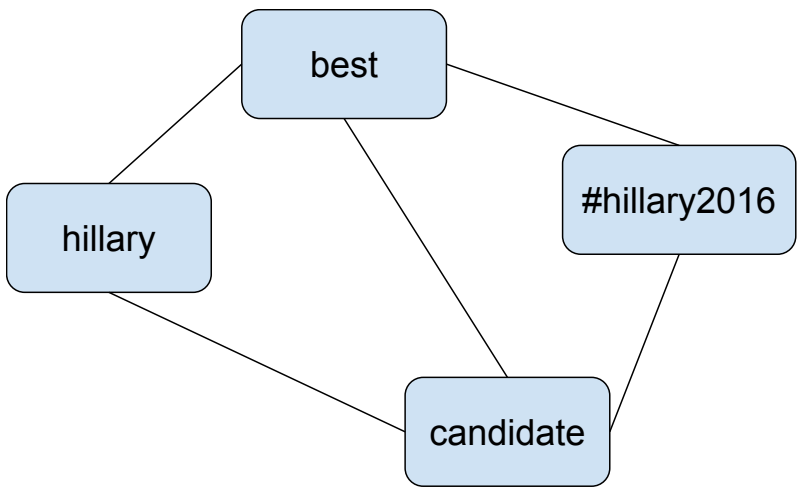
1. preprocess tweets
2. words co-occurrence graph
3. filter:
   - 3.1 α (filter nodes)
   - 3.2 β (filter edges)
4. k-cliques are our contextonyms.

Tweet:

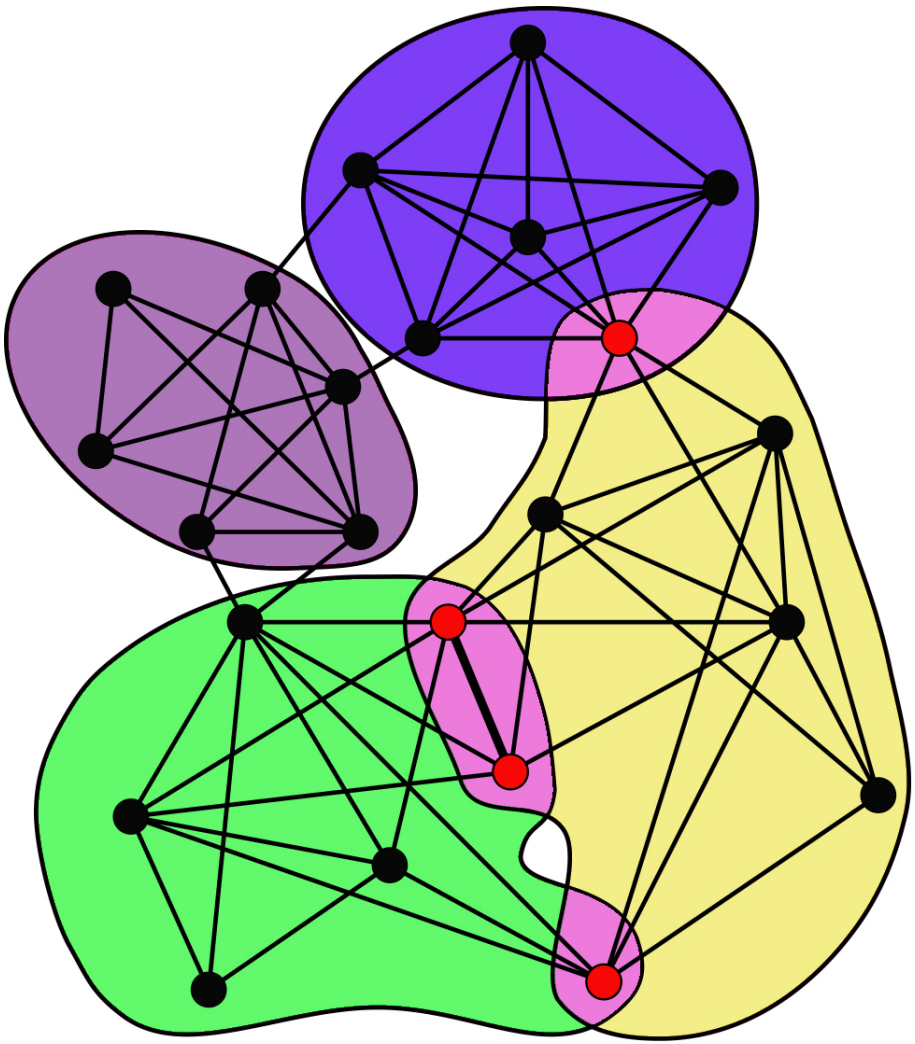
Hillary is the best candidate #hillary2016

Processed Tweet:

hillary best candidate #hillary2016



Communities of words: k-cliques



Stance in tweets

Stance in tweets

Target: Hillary Clinton.

Stance in tweets

Target: Hillary Clinton.

Sample tweet FAVOR

I'm proud to announce I support #HillaryClinton!!!!

Stance in tweets

Target: Hillary Clinton.

Sample tweet FAVOR

I'm proud to announce I support #HillaryClinton!!!!

Sample tweet AGAINST

#WhyImNotVotingForHillary <<<<<<<SHE IS A CRIMINAL

Stance in tweets

Target: Hillary Clinton.

Sample tweet FAVOR

I'm proud to announce I support #HillaryClinton!!!!

Sample tweet AGAINST

#WhyImNotVotingForHillary <<<<<<<SHE IS A CRIMINAL

Sample tweet NONE

Ding Dong ____ _____ __ ____


SemEval-2016 Task 6

Details about the SemEval task

- **Title:** SemEval2016-task6 subtask-A
- 5 independent **targets**:
  - Atheism
  - Climate Change is a Real Concern
  - Feminist Movement
  - Hillary Clinton
  - Legalization of Abortion

Details about the SemEval task

- **Title:** SemEval2016-task6 subtask-A
- 5 independent **targets**:
  - Atheism
  - Climate Change is a Real Concern
  - Feminist Movement
  - Hillary Clinton
  - Legalization of Abortion
- 2 **datasets**:
  - training: 2914 texts of tweets, unbalanced
  - test: 1250 texts of tweets, quite well balanced

Details about the SemEval task

- **Title:** SemEval2016-task6 subtask-A
- 5 independent **targets**:
  - Atheism
  - Climate Change is a Real Concern
  - Feminist Movement
  - Hillary Clinton
  - Legalization of Abortion
- 2 **datasets**:
  - training: 2914 texts of tweets, unbalanced
  - test: 1250 texts of tweets, quite well balanced
- and some **critics**:
  - training set not large enough to train a classifier
  - man-made annotation: need to perfectly know the topic to understand the stances

Baselines

To benchmark our classifier, we created two baselines using two standard stance detection approaches:

1. Sentiment based: SENT-BASE
2. Learning based: SVM-UNIG

Baselines

1. Sentiment based: SENT-BASE

Idea

Each word is associated with a "positivity" score, a "negativity" score, and an "objectivity" score.
Use SentiWordNet 3.0 [Baccianella et al., 2010].

Baselines

1. Sentiment based: SENT-BASE

Idea

Each word is associated with a "positivity" score, a "negativity" score, and an "objectivity" score.
Use SentiWordNet 3.0 [Baccianella et al., 2010].

Valence

A weighted sum of the scores of all the words in a tweet.

Baselines

1. Sentiment based: SENT-BASE

Idea

Each word is associated with a "positivity" score, a "negativity" score, and an "objectivity" score.
Use SentiWordNet 3.0 [Baccianella et al., 2010].

Valence

A weighted sum of the scores of all the words in a tweet.

The valence indicates the overall sentiment of the tweet: a positive valence means that the tweet is favorable, etc.

Baselines

2. Learning based: SVM-UNIG

Idea

1. Select the 10,000 unigrams (words) that are **most indicative** of stance from training corpus.

Baselines

2. Learning based: SVM-UNIG

Idea

1. Select the 10,000 unigrams (words) that are **most indicative** of stance from training corpus.
2. Construct a feature vector of boolean indicators of unigram presence in each tweet.

Baselines

2. Learning based: SVM-UNIG

Idea

1. Select the 10,000 unigrams (words) that are **most indicative** of stance from training corpus.
2. Construct a feature vector of boolean indicators of unigram presence in each tweet.
3. Train SVM classifier on annotated training corpus.

Improving baselines with contextonyms

Sentiment based approach: SENT-CTXT

Idea

We can improve the sentiment analysis of a tweet by looking at the contextonyms associated with that tweet.

Improving baselines with contextonyms

Sentiment based approach: SENT-CTXT

Idea

We can improve the sentiment analysis of a tweet by looking at the contextonyms associated with that tweet.

1. Associate each tweet with contextonyms.

Improving baselines with contextonyms

Sentiment based approach: SENT-CTXT

Idea

We can improve the sentiment analysis of a tweet by looking at the contextonyms associated with that tweet.

1. Associate each tweet with contextonyms.
2. Compute the valence of those contextonyms. This indicates the sentiment of the tweet.

Improving baselines with contextonyms

Learning based approach #1: SVM-CTXT

Idea

Contextonyms can improve the SVM classifier because we get more information about the context of the tweet.

Improving baselines with contextonyms

Learning based approach #1: SVM-CTXT

Idea

Contextonyms can improve the SVM classifier because we get more information about the context of the tweet.

Same classifier as SVM-UNIG but feature vector is a boolean indicator of contextonym presence.

Improving baselines with contextonyms

Learning based approach #2: SVM-EXP

Idea

The fact that tweets are short make them difficult to analyze and contextonyms are adding information about context.

Improving baselines with contextonyms

Learning based approach #2: SVM-EXP

Idea

The fact that tweets are short make them difficult to analyze and contextonyms are adding information about context.

Expand the tweets with the associated contextonyms and then train a SVM on the best unigrams.

Results

Extraction-Implementation

Source corpus used to extract contextonyms

- huge: 7,773,089 tweets
- not annotated: this part is easy
- same language: English-written tweets
- same topics: Clinton, Trump, the abortion debate, religion, and miscellaneous US politics
- gathered between November 20th and December 1st, 2015 using the free Twitter Stream API.

Tools used

- stopwords list
- regexp to remove URLs
- NetworkX library
- **not used:** POS taggers and lemmatisers

Parameters

1. $\alpha_{threshold} = 10$; consequence: vocabulary size at 50,000
2. $\beta_{threshold} = 0.06$; number of edges at 300,000
3. $k_c = 4$ the size of smallest clique; 6278 contextonyms ("contexto-sets")

Measure the performance of Results

$$P_s = Precision \tag{1}$$

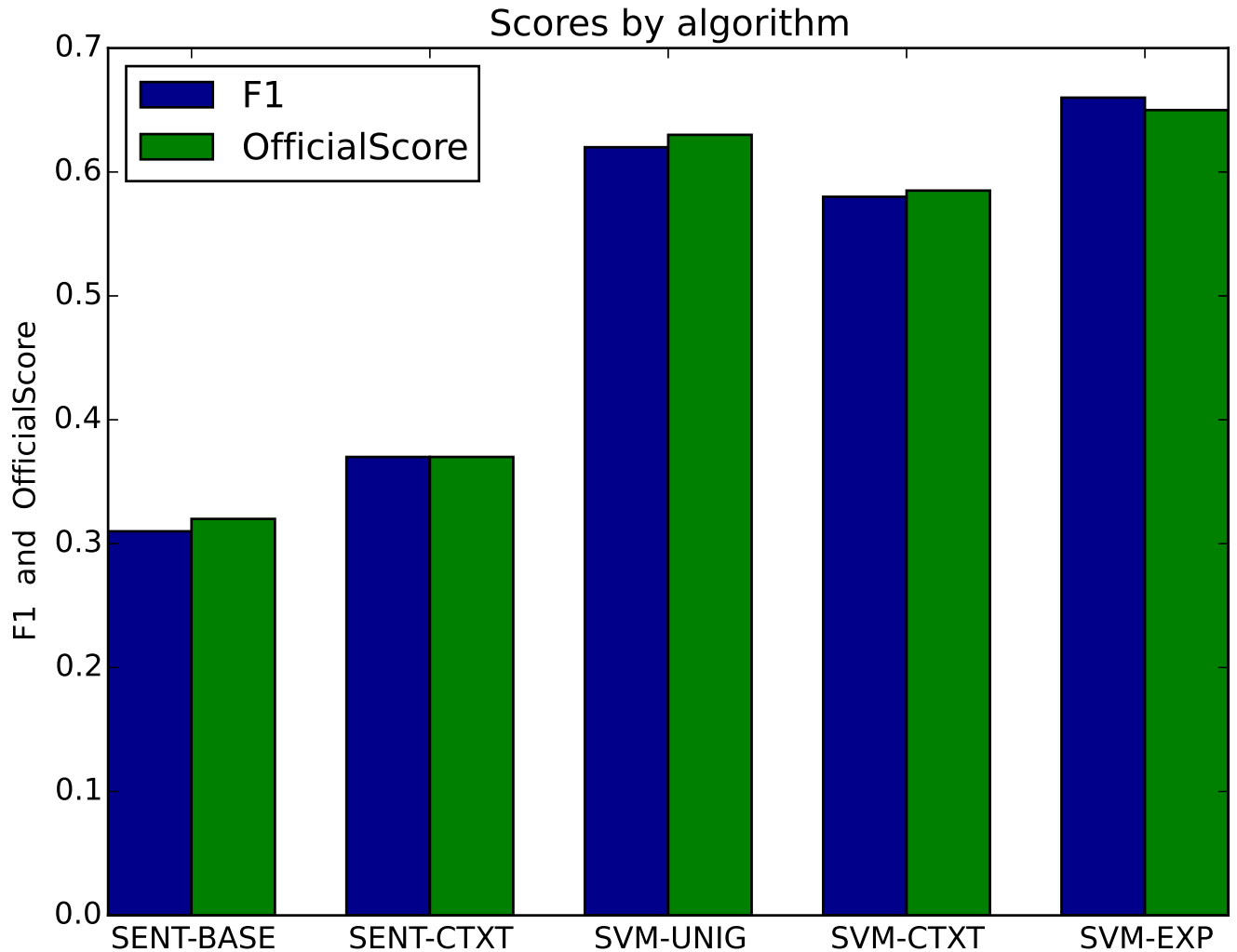
$$R_s = Recall \tag{2}$$

$$F_1(s) = 2\frac{P_s R_s}{P_s + R_s} \tag{3}$$

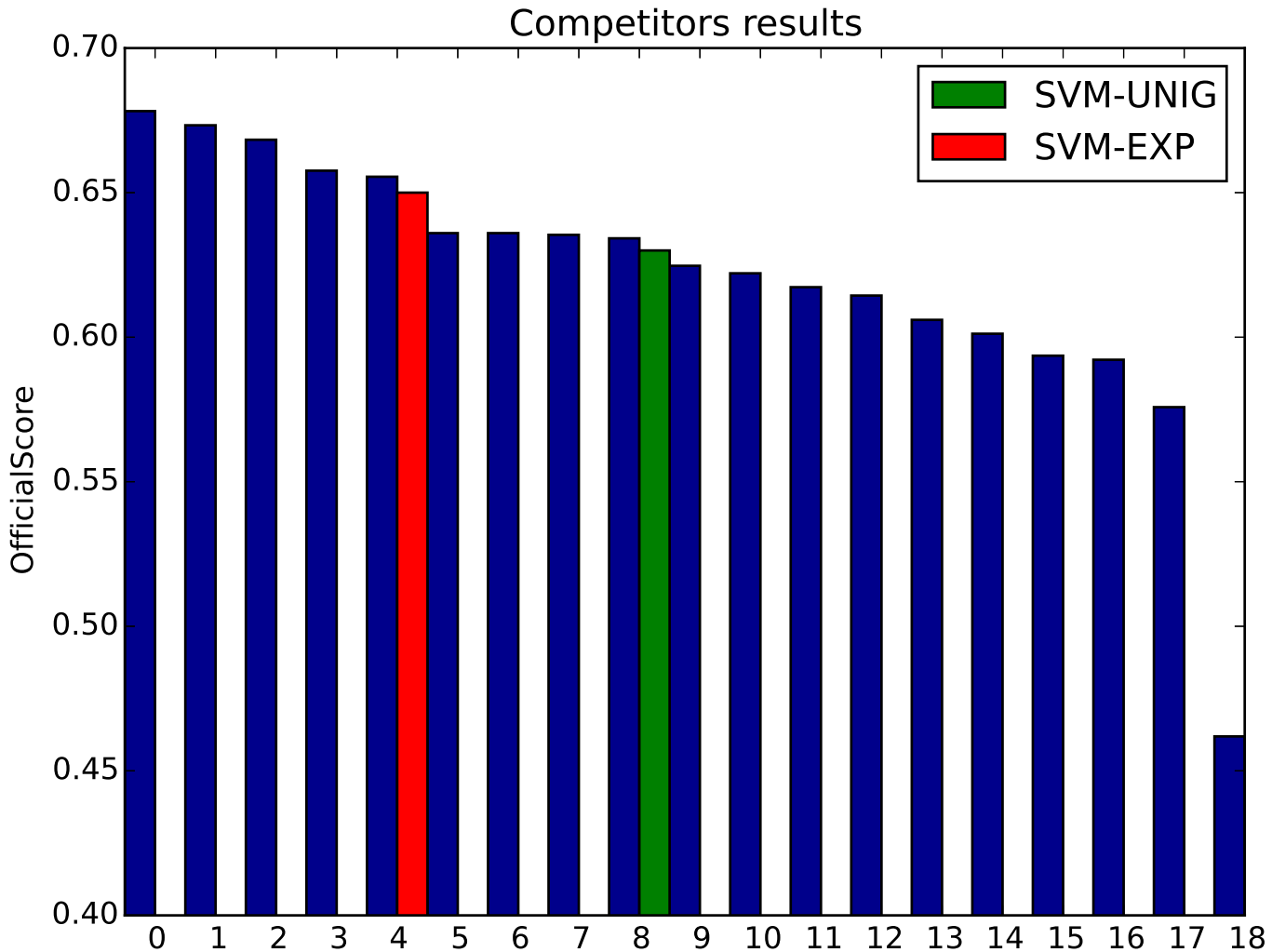
Official Score for benchmarking purposes:

$$Score = \frac{1}{2}\ (F_1(F) + F_1(A)) \tag{4}$$

Comparison of classifiers



Comparison of competitors



Conclusion

Summary

- Challenging task: shortness of tweets, innovative spelling and specific usage of words.
- Lexical relatedness: more information to understand the tweets
- Contextonyms: a tool to be adapted to one's needs and resources

Conclusion

Summary

- Challenging task: shortness of tweets, innovative spelling and specific usage of words.
- Lexical relatedness: more information to understand the tweets
- Contextonyms: a tool to be adapted to one's needs and resources

Future Works

- Disambiguate ambiguous tweets only
- Focus on user's opinion
- Look for groups of users

Questions

Thank you for your attention!

Remarks, questions?

References I


Andreevskaia, A. and Bergler, S. (2008).
When specialists and generalists work together: Overcoming domain dependence in sentiment tagging.
In *ACL*, pages 290–298.


Baccianella, S., Esuli, A., and Sebastiani, F. (2010).
Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.
In *LREC*, volume 10, pages 2200–2204.


Feng, Y., Fani, H., Bagheri, E., and Jovanovic, J. (2015).
Lexical semantic relatedness for twitter analytics.
In *Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on*, pages 202–209. IEEE.

References II


Hasan, K. S. and Ng, V. (2013).
Extra-linguistic constraints on stance recognition in ideological debates.
In *ACL (2)*, pages 816–821.


Hyungsuk, J., Ploux, S., and Wehrli, E. (2003).
Lexical knowledge representation with contexonyms.
In *9th MT summit Machine Translation*, pages 194–201.


Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013).
Distributed representations of words and phrases and their compositionality.
In *Advances in neural information processing systems*, pages 3111–3119.

References III


Miller, G. A. (1995).
Wordnet: a lexical database for english.
Communications of the ACM, 38(11):39–41.


Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001).
Linguistic inquiry and word count: Liwc 2001.
Mahway: Lawrence Erlbaum Associates, 71:2001.


Ploux, S. and Ji, H. (2003).
A model for matching semantic maps between languages (french/english, english/french).
Computational linguistics, 29(2):155–178.

References IV

Serban, O., Pauchet, A., Rogozan, A., Pécuchet, J.-P., and LITIS, I. (2012).
Semantic propagation on contextonyms using sentiwordnet.
In *WACAI 2012 Workshop Affect, Compagnon Artificiel, Interaction*, page 86.

Wang, R., Zhao, H., Ploux, S., Lu, B.-L., and Utiyama, M. (2016).
A bilingual graph-based semantic model for statistical machine translation.
In *International Joint Conference on Artificial Intelligence*.

References V

Wilson, T., Wiebe, J., and Hoffmann, P. (2005).
Recognizing contextual polarity in phrase-level sentiment analysis.
In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Zesch, T., Müller, C., and Gurevych, I. (2008).
Using wiktionary for computing semantic relatedness.
In *AAAI*, volume 8, pages 861–866.

Sentiment Analyser

Compute a valence

Let $S(n)$ be the set of i synsets s_i containing the word n. Each synset has a positive and a negative valence s_i^+, s_i^-. .
Let S_t be the set of all the N synsets taken into account for the whole tweet. We therefore define the valence $v(t)$:

$$v(t) = \frac{1}{N} \sum_{s_i \in S_t} s_i^+ + s_i^- \tag{5}$$

If $v(t)$ is positive (negative), we assume the tweet is supportive (opposed), thus having a stance *FAVOR* (*AGAINST*).

Stance classifier

SVM-UNIG, using a SVM on word unigrams:
Comparison:

- different algorithms: SVM-linear, **SVM-RBF**, NN, Bayes, ...
- parameter settings:
  - $C = 100.0$
  - $\gamma = 0.01$

The feature vector is composed of the boolean indicators of the unigrams presence.
Vocabulary size is fixed at 10,000, which limits the feature vector length.

Graph filtering parameters - α

$g_t = (V_t, E_t)$ the co-occurrence graph for a single tweet t. Average degree ϕ of a word n, due to its position:

$$\phi(n) = \frac{1}{K} \sum_{j=1}^{K} d(n)_{g_j} \tag{6}$$

$\alpha(n)$, the ratio of degree in G to average degree position for word n:

$$\alpha(n) = \frac{d(n)_G}{\phi(n)} \tag{7}$$

A large score implies that word n occurs in a great variety of contexts. A word n would then be removed if $\alpha(n) < \alpha_{threshold}$.

Graph filtering parameters - β

β consists of two weight-node count ratios.

$$\beta(e) = \frac{w_e}{c_{n_{1,e}}} + \frac{w_e}{c_{n_{2,e}}} \tag{8}$$

- w_e is the weight of edge $e = (n_1, n_2)$,
- $c_{n_{1,e}}$ and $c_{n_{2,e}}$ are the word counts for the two words n_1 and n_2 connected by e.

A value approaching 2 implies the association is very important for both words.
Filter away the edges whenever $\beta_e < \beta_{threshold}$, to get rid of unimportant associations.