

# Northumbria Research Link

Citation: Chalothorn, Tawunrat and Ellman, Jeremy (2013) TJP: using Twitter to analyze the polarity of contexts. In: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 14-15 June 2013, Atlanta, Georgia.

URL:

This version was downloaded from Northumbria Research Link:  
<http://nrl.northumbria.ac.uk/13077/>

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: <http://nrl.northumbria.ac.uk/policies.html>

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher's website (a subscription may be required.)

[www.northumbria.ac.uk/nrl](http://www.northumbria.ac.uk/nrl)



# Using SentiWordNet for Analyzing Radical Contents on Web Forums

Tawunrat Chalothorn  
(1<sup>st</sup> year Ph.D. student)

Dr. Jeremy Ellman and Dr. Paul Vickers  
(Principle Supervisor and Second Supervisor)

University of Northumbria at Newcastle

# Contents

---

- Background
- Research Question
- Data
- Method
- Results
- Conclusion
- Future Work

## Who is interested in this topic?

UK government interests in this topic according to, having launched a 'Prevent Strategy' to prevent radicalization of youth in Great Britain and blocked networks that support terrorists (Morgan, 2011).

# Background (2)

---

- Internet and web forums have become important place for social communication.
- Some radical groups also use them for communication and disseminating their ideologies to the public.
- So I intended to develop techniques to classify and detect radical contents. The final result obtained could be built into a web browser based tool.

How effective is SentiWordNet for detecting opinions and emotions on web forums?

- Selected by using research from 21 people (Arabic speaker).
- Two Arabic forums have been selected: Montada and Qawem.

# Method (1)

---

1. Collected data from web forums.
2. Translated sentences to English Language by Arabic speaker.\*
3. Built system by using Python programming, SentiWordNet, WordNet and NLTK.
4. Used POS tagging each word in sentences.
5. Removed stopwords from sentences.
6. Separated sentences into words for calculating scored.

\*Khaled Nakkachi (Translated sentences from Arabic to English)



# Method (2)

7. Calculated scored of each word by adapting formulas from Neivarouskaya, et al. (2007).

$$Pos\_weight = \left[ \frac{pos}{senses} \right]$$

$$Neg\_weight = \left[ \frac{neg}{senses} \right]$$

*pos* is the number of lemma that have  $Pos(s)(i) \geq Neg(s)(i)$  and  $Pos(s)(i) \neq 0$   
*neg* is the number of lemma that have  $Neg(s)(i) \geq Pos(s)(i)$  and  $Neg(s)(i) \neq 0$   
*senses* is the total number of lemma in synsets.

# Method (3)

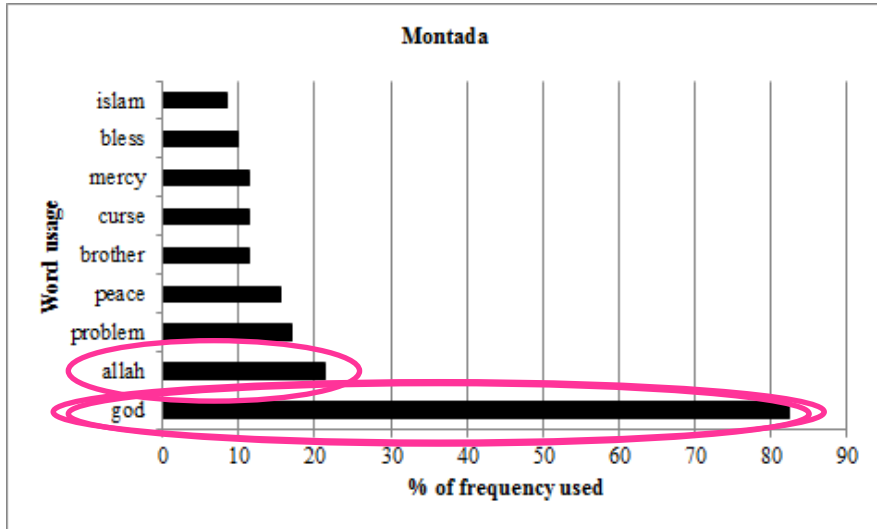
8. Calculate sentences' scores from words' scores by using formula from Khan and Baharudin (2011)

$$Sentence\_score = \left[ \frac{\sum_{i=1}^n Score(i)}{n} \right]$$

*Score(i)* is the positive/negative scores of the word in sentences.

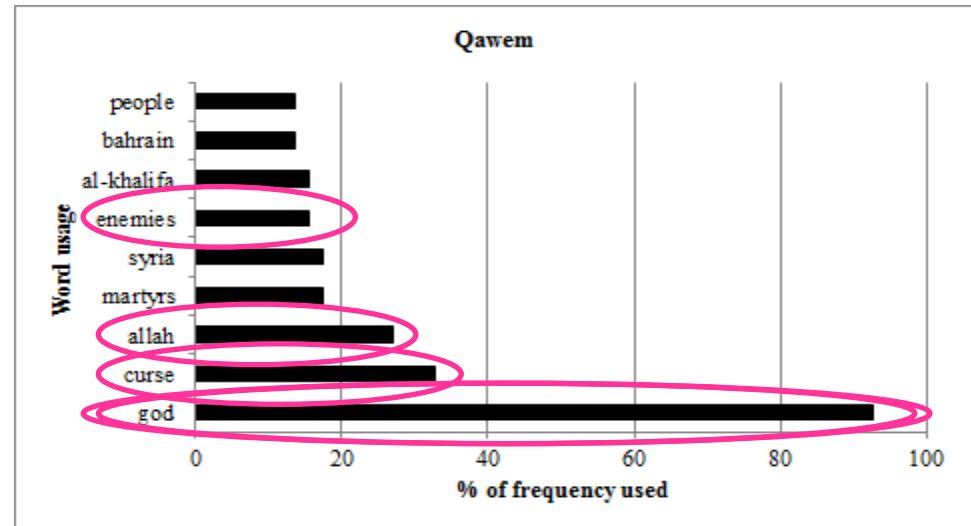
*n* is the number of words in sentences.

# Results (1)

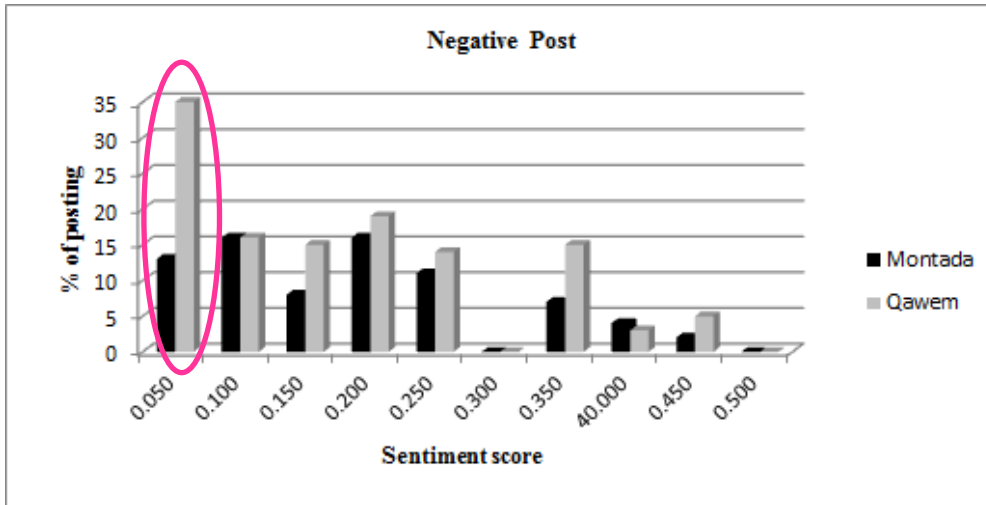


← Top high frequency words in Montada

Top high frequency words in Qawem →

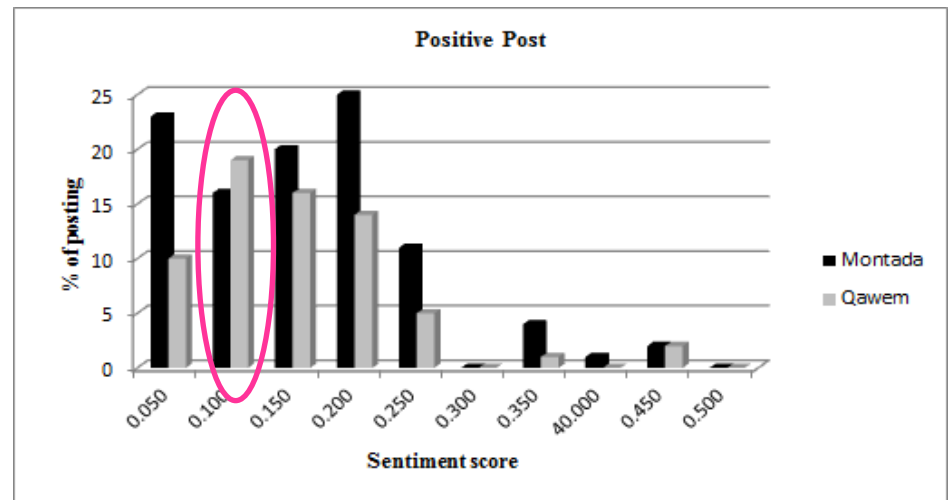


# Results (2)



← Negative scores of sentiment analysis

Positive scores of sentiment analysis →



# Conclusion

---

- Two web forums “Montada” and “Qawem” were chosen because their contents related to radicalisation.
- System was developed by using Python programming language, SentiWordNet, WordNet and NLTK.
- From the overall results, Qawem has the contents that related to radial Islamic ideology more than Montada.
- SentiWordNet could be used for analyse contents derived from web forums.

- Word Sense Ambiguity will be research in more details.
- Annotators who are native speaker of Arabic will be asked to provides ratings of sentences, in order to compare these with the results generated via this experiment.

Any Question?

Thank you!

# Humans annotation

<b>Negative</b>	<b>Neutral</b>	<b>Positive</b>	<b>No meaning</b>
HA 1 HA 5	HA 4	HA 2	HA 3



# POS tagging labels

POS Meaning	POS Tag	SentiWordNet Tag
Verb	VB, VBD, VBG, VBN, VBP, VBZ	v
Noun (s)	NN, NNS, NNP, NNPS	n
Adverb (s)	RB, RBR, RBS	r
Adjective (s)	JJ, JJR, JJS	a

## Stopwords

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves',  
'you', 'your', 'yours', 'yourself', 'yourselves', 'he', 'him',  
'his', 'himself', 'she', 'her', 'hers', 'herself', 'it', 'its', 'itself',  
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which',  
'who', 'whom', 'this', 'that', 'these', 'those', 'am', 'is', 'are',  
'was', 'were', 'be', 'been',...]

# Example of POS tagging

We will clean all the zones of unclean Americans and Suduis.\*

('We', 'PRP')	('will', 'MD')	('clean', 'VB'),
('all', 'DT')	('the', 'DT')	('zones', 'NNS')
('of', 'IN')	('unclean', 'JJ')	('Americans', 'NNPS')
('and', 'CC')	('Suduis', 'NNPS')	

\*These are not views expressed or implied by the author or the University of Northumbria at Newcastle.