All About Microtext –

Philosophy and a Survey of Current Results

Jeffrey Ellen – Research Scientist

SPAWAR Systems Center Pacific – US Navy
jeffrey.ellen@navy.mil

8 Aug 2011

The Exponential Growth of Text Capture

- Driven by internet growth & Moore’s law (cheap computers)
  - Electronic encoding first reserved for lengthy and important documents like drafts of books and contracts
  - Expanded to essays, newswire articles, etc
  - Wiki, micro-blog (i.e. Twitter), SMS, voicemail transcription
- ‘Barrier for entry’ of text encoding continues to be lowered:
  - cost required to encode
  - accessibility to encoded work
  - knowledge required to operate encoding technology
  - **As cost goes down, so does formality and rigor**
- **Brief, informal communication has always existed, just not previously available for academic study and analysis**

Defining ‘Microtext’

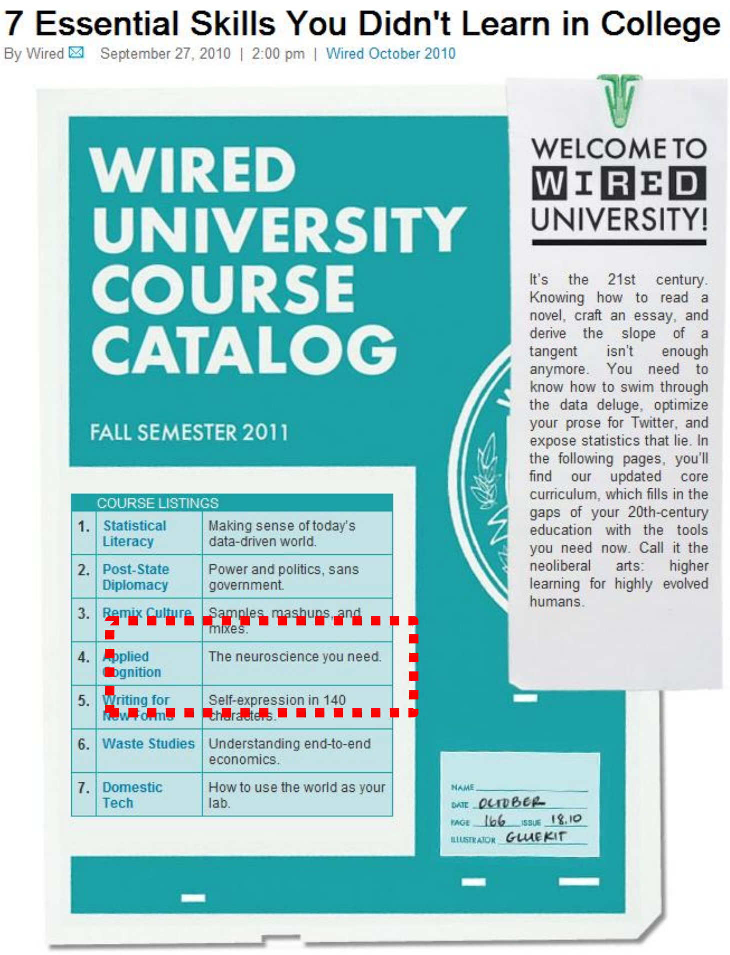
- Microtext : Text :: Dialect : Language ?
- It’s not exactly email: Dalli, Xia, and Wilks (2004) presented a summary of the “unique characteristics of email”:
  - Short messages between **2-800** words.
  - Unconventional grammar & style (frequently).
  - A cross between informal and traditional.
  - **Threading** characteristics
- 800 words too broad for microtext. ~700 words on an AAAI formatted page, ~70 words on this slide.
  - O’Connor, et al (2010) collected tweets and found avg. 11 words
- Threading very common in microtext, and definitely a unique issue, but not required.

Approved for public release; distribution is unlimited.

Why delineate ‘Microtext’?

- Different domains require different algorithms, preprocessing steps, tools, feature extractors.
  - Image processing vs. speech recognition vs. NLP text processing
- Microtext is sufficiently different from any precursor to necessitate unique study
  - Initial experimental results by numerous different parties indicate that traditional long-text techniques do not translate well to microtext

‘Mainstream’ Culture shares intuition



7 Essential Skills You Didn't Learn in College

By Wired ✉ September 27, 2010 | 2:00 pm | Wired October 2010

WIRED UNIVERSITY COURSE CATALOG

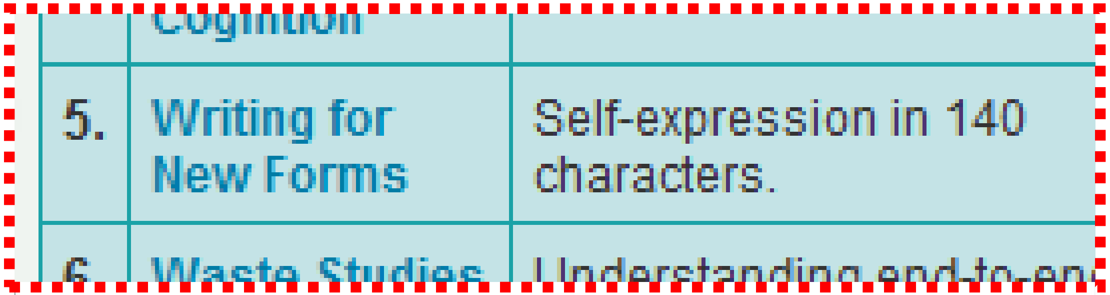
FALL SEMESTER 2011

COURSE LISTINGS		
1.	Statistical Literacy	Making sense of today's data-driven world.
2.	Post-State Diplomacy	Power and politics, sans government.
3.	Remix Culture	Samples, mashups, and mixes.
4.	Applied Cognition	The neuroscience you need.
5.	Writing for New Forms	Self-expression in 140 characters.
6.	Waste Studies	Understanding end-to-end economics.
7.	Domestic Tech	How to use the world as your lab.

WELCOME TO WIRED UNIVERSITY!

It's the 21st century. Knowing how to read a novel, craft an essay, and derive the slope of a tangent isn't enough anymore. You need to know how to swim through the data deluge, optimize your prose for Twitter, and expose statistics that lie. In the following pages, you'll find our updated core curriculum, which fills in the gaps of your 20th-century education with the tools you need now. Call it the neoliberal arts: higher learning for highly evolved humans.

NAME
DATE OCTOBER
PAGE 166 ISSUE 18.10
ILLUSTRATOR GLUEKIT



	Cognition	
5.	Writing for New Forms	Self-expression in 140 characters.
6.	Waste Studies	Understanding end-to-enc

“Why take this course?
You can write a cogent essay, but can you write it in 140 characters or less?
What you’ll learn:
How to adapt your message to multiple formats and audiences—human and machine.”

“Writing today also means mastering metatext, the cues and context that determine how, where, and if your words get read. “

- Wired Magazine, October 2010

A working definition of ‘Microtext’

- Individual author contributions are very brief, consisting of as little as a single word, and almost always less than a paragraph. Frequently the contribution is a single sentence or less.

- The grammar used by the authors is generally informal and unstructured, relative to the pertinent domain. The tone is conversational, and frequently unedited therefore errors and abbreviations are more common.

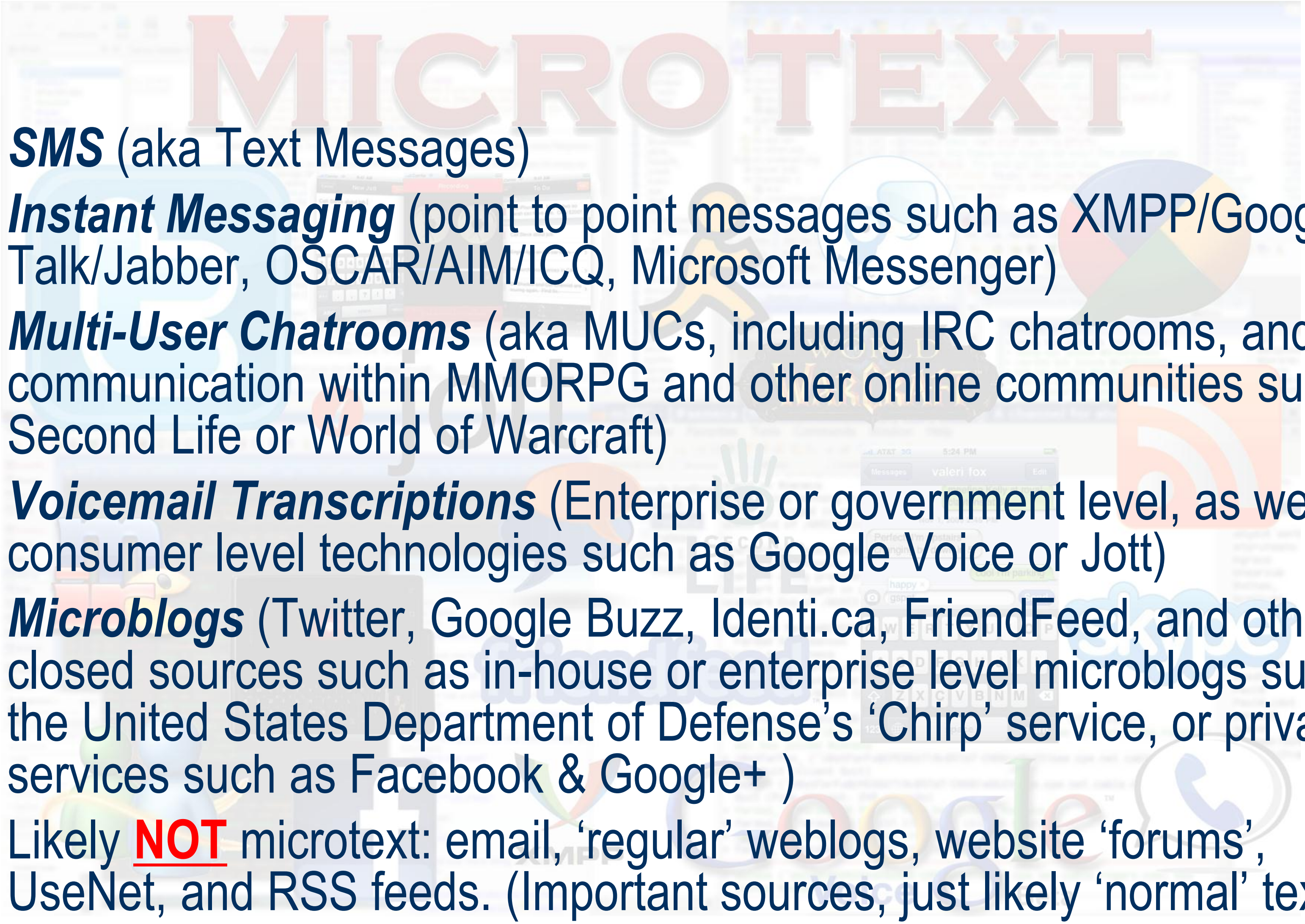
- The text is ‘semi-structured’ by traditional NLP definitions since it contains some meta-data in proportion to some free-text. At a minimum, all microtext has a minute-level timestamp and a source attribution (author).

A working definition of ‘Microtext’

- Individual author contributions are very brief, consisting of as little as a single word, and almost always less than a paragraph. Frequently the contribution is a single sentence or less.
  Quantify experimentally through results or through measurements
- The grammar used by the authors is generally informal and unstructured, relative to the pertinent domain. The tone is conversational, and frequently unedited therefore errors and abbreviations are more common.
  Flesh-Kincaid Grade level, Flesh readability but for longer texts
- The text is ‘semi-structured’ by traditional NLP definitions since it contains some meta-data in proportion to some free-text. At a minimum, all microtext has a minute-level timestamp and a source attribution (author).
  Determine if other meta-data is required (audience, url, #tag)

Examples of Microtext

Examples of Microtext



- ***SMS*** (aka Text Messages)
- ***Instant Messaging*** (point to point messages such as XMPP/Google Talk/Jabber, OSCAR/AIM/ICQ, Microsoft Messenger)
- ***Multi-User Chatrooms*** (aka MUCs, including IRC chatrooms, and communication within MMORPG and other online communities such as Second Life or World of Warcraft)
- ***Voicemail Transcriptions*** (Enterprise or government level, as well as consumer level technologies such as Google Voice or Jott)
- ***Microblogs*** (Twitter, Google Buzz, Identi.ca, FriendFeed, and other closed sources such as in-house or enterprise level microblogs such as the United States Department of Defense’s ‘Chirp’ service, or private services such as Facebook & Google+)
- Likely **NOT** microtext: email, ‘regular’ weblogs, website ‘forums’, UseNet, and RSS feeds. (Important sources, just likely ‘normal’ text)

Topic ID / Individual Summarization

- Topic Detection within IRC chatrooms. Modified TF/IDF approach with **temporal augmentation** (Adams 2008)
- Ranganath, Jurafsky, and McFarland (2009) were able to achieve 71.5% accuracy on a system designed to detect a speaker's intent to flirt using a spoken corpus of speed-dates. **Transcriptions included interruptions, pauses, laughter, backchannel utterances**. (Examples include 'Uh-huh, Yeah, Wow, Excuse Me, Um, Uh).
- Ritter, Cherry, and Dolan (2010) model Twitter conversations using an unsupervised learning. In their collection of 1.3 million tweets, they note that Twitter postings tend to be "highly ungrammatical, and filled with spelling errors". They also note that 69% of the conversations in their data had a length of two. **Modified LDA overcame difficulties encountered by named entity recognizers and noun-phrase chunkers.**

Clustering / Mass Sumarization

- Zeitgeist or ‘Trending Topics’ currently being pursued by multiple companies, all closed & proprietary
- TweetMotif (O’Connor, 2010) extends collocations to tweets: starting with one term, “... groups them by statistically unlikely phrases that co-occur”. **Discards duplicates by “messages whose sets of trigrams have a pairwise Jaccard similarity exceeding 65%.”**
- TWinner (Abrol, 2010) to attempt to cluster tweets by physical location, and then utilize this information to “improve the quality of web search and predicting whether the user is looking for news or not.” TWinner also defines a ‘Frequency-Population ratio), which is a ratio of the number of tweets per geographic location, normalizing with respect to population density.

Classification

- Phan (2008) proposes a “general framework for building classifiers that deal with short and sparse text & Web segments by making the most of hidden topics”. The approach leverages a ‘universal dataset’ to augment the short and sparse text collected. *Same limitation as long NLP.*
- Dela Rosa and Ellen (2009) have completed a series of experiments on classification of military chat posts.
  - Evaluated algorithms including SVMs, k-Nearest Neighbour, Rocchio, and Naive Bayes.
  - Evaluated various feature selection methodologies: **Mutual Information (MI) and Information Gain (IG) were found to perform relatively poorly.**
  - **K-NN and SVM were found to be the most suitable in a binary and four-way classification task.**

Sentiment Analysis

- Go and Bhayani (2010) perform sentiment analysis of Twitter messages. They are able to **leverage emoticons as noisy labels**, a technique first presented by Read (2005). **Attempted to perform clustering to assist with the analysis, and found that it unexpectedly hurt results.**
- Wilson, Wiebe, and Hoffmann (2005) examine contextual polarity (aka semantic orientation) of phrases in great detail. The stated goal of this work is to provide insight into phrase-level sentiment analysis. Some microtext is not much more than a phrase in length, so this type of research is definitely applicable.

Information Extraction

- Marom and Zukerman (2009) study a corpus of paired question & response help desk emails to automate the process. They investigate sentence level granularity. **One thing specifically investigated is sentence cluster cohesion, a measure of the similarity of sentences to each other.**
- Gruhl, et. al (2009) explore “statistical NLP techniques to improve named entity annotation in challenging Informal English domains”. **They achieve notably better results through application of SVMs.**

Semi-Structured Data Exploitation

- Kinsella, Passant, and Breslin (2010) examine the occurrences of hyperlinks in online message boards. They observe that not only is the use of hyperlinks increasing, but the hyperlinks themselves often reference “resources with associated structured data”
- Wang (2010) provides another example of utilizing the structure of the data in his research into identifying spammers on Twitter. He utilizes some of the relationship information available from **twitter accounts to construct graphs** and examine some typical directed graph features. Also, Wang makes the interesting choice of **ignoring the NLP aspect of the tweets completely, and instead treating authors’ contributions AS STRINGS OF SYMBOLS, and compares them using Levenshtein distance, ignoring grammar and semantic content completely.**

Leveraging non-linguistic aspects

- Using SMS to interface with other systems like FAQs (Kothari, 2009) or yellow pages (Kopparapu, 2007).
- Mowbray (2010) identifies Twitter spam through API abuse.
- Implications of ‘New Media/Social Networks’ on society:
  - The influence of Twitter (Cha, 2010) (Lee, 2010)
  - Using Twitter to predict elections (Tumasjan, 2010)
  - Using Twitter to predict the stock market, or movie results, or the flu (Ritterman, 2009).
  - These approaches generally relied on specific term matches
- It is really the publicness and ubiquity of the mechanisms that are being exploited, not the microtext.

Tackling linguistic aspects

- Identifying & Normalizing ‘ill-formed’ and ‘out of vocabulary’ words’, specifically in SMS & Twitter messages.
  - (Han & Baldwin, ACL 2011)
- Chat word dictionary (http://chat.reichards.net/)
  - No different than stopwords as agreed upon from a vetted source, such as Cornell’s SMART program
- Dozens of attempts to parse/leverage tags & hashtags

Conclusions

- Explosive # of papers published on NLP and AI techniques as applied to brief, poorly formatted, semi-structured text.
- Most current work is more engineering than science; providing anecdotal or experimental evidence about a single use case.
- **Some discussion and meta-experimentation on the field itself would lead to greater insights, with a higher level of reuse. A first step in that direction is defining terminology, 'Microtext', so that researchers can have a common ground for future discussion.**
- Next step: investigating and more rigorously quantifying the three attributes in the microtext definition.

References

- Abrol, S. and Khan, L. 2010. TWinner: understanding news queries with geo-content using Twitter. In *Proceedings of the 6th Workshop on Geographic information Retrieval* (Zurich, Switzerland, February 18 - 19, 2010). GIR '10. ACM, New York, NY, 1-8
- Adams, P., and Martell, C., 2008. Topic Detection and Extraction in Chat. In *International Conference on Semantic Computing,* IEEE.
- Bullen, R.H. Jr., and Millen, J. K., 1972. Microtext: the design of a microprogrammed finite state search machine for full-text retrieval. In *Proceedings of the AFIPS Joint Computer Conferences*. ACM.
- Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. 2010. Measuring user influence in twitter: the million follower fallacy. In *Proceedings of the 4th International Conference on Weblogs and Social Media*, AAAI, Washington, D.C., 2010.
- Chi, E. 2009 "Information Seeking Can Be Social," Computer, pp. 42-46, March, 2009. IEEE
- Cong, G., et al. (2008). Finding question-answer pairs from online forums. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 467-474, New York, NY, USA. ACM.
- Dalli, A., Xia, Y., and Wilks, Y., 2004. FASIL email summarisation system. In*Proceedings of the 20th international conference on Computational Linguistics* (COLING '04). ACL, Morristown, NJ, USA, , Article 994 .
- Davidov, D., Tsur, O., Rappoport, A. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys, In *Proceedings of the 23rd international conference on Computational Linguistics (COLING), 2010*.
- Flesch, R. (1948); A new readability yardstick, *Journal of Applied Psychology*, Vol. 32, pp. 221–233

References (continued)

- Go, A., Bhayani, R., and Huang, L. 2010. Exploiting the Unique Characteristics of Tweets for Sentiment Analysis. *CS224N Project Report*, Stanford.
- Gruhl, D., Nagarajan, M., Pieper, J., Robson, C., and Sheth, A. 2009. Context and Domain Knowledge Enhanced Entity Spotting in Informal Text. In *Proceedings of the 8th international Semantic Web Conference*. 260-276
- Han, B., Baldwin, T. 2011. Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. In *ACL 2011*.
- Kinsella, S., Passant, A., Breslin, J. 2010. Ten Years of Hyperlinks in Online Conversations. In *Proceedings of the Web Science Conference 2010.* WWW2010.
- Lee, C., Kwak, H., Park, H., Moon, S., 2010. Finding influentials based on the temporal order of information adoption in twitter. In *Proceedings of the 19th international conference on World wide web (WWW '10)*. ACM, New York, NY, USA, 1137-1138.
- Laporte, Leo. 2009. [Internet Radio Broadcast] This Week in Google 13. October 24, 2009.
- Kopparapu, S. K., Srivastava, A., and Pande, A. 2007. SMS based natural language interface to yellow pages directory. In *Proceedings of the 4th international Conference on Mobile Technology, Applications, and Systems and the 1st international Symposium on Computer H uman interaction in Mobile Technology* ACM. Mobility '07. ACM, New York, NY, 558-563.
- Kothari, G., Negi, S., Faruquie, T. A., Chakaravarthy, V. T., and Subramaniam, L. V. 2009. SMS based interface for FAQ retrieval. In P*roceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th international Joint Conference on Natural Language Processing of the Afnlp*. Association for Computational Linguistics. Morristown, NJ, 852-860.

References (continued)

- Marom, Y. and Zukerman, I. 2009. An empirical study of corpus-based response automation methods for an e-mail-based help-desk domain. *Computational Linguist.* 35, 4 (Dec. 2009), 597-635
- Mowbray, M. 2010. The Twittering Machine. In *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST 2010)*. INSTICC. 299-304.
- O'Connor, B., Krieger, M., and Ahn, D. 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media.* Washington, DC, May 2010
- Phan, X.-H., et al. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pp. 91-100, New York, NY, USA. ACM.
- Ranganath, R., Jurafsky, D., and McFarland, D. 2009. It's not you, it's me: detecting flirting and its misperception in speed-dates. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*. ACL
- Read, J. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop* ACL
- Ritter, A., Cherry, C. And Dolan, B. 2010 Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, Los Angeles, CA, 172-180.

References (continued)

- Ritterman, J., Osborne, M., and Klein, E. 2009. Using prediction markets and Twitter to predict a swine flu pandemic. In *1st International Workshop on Mining Social Media - 13th Conference of the Spanish Association for Artificial Intelligence, 2009.* AEPIA (Asociación Española de Inteligencia Artificial)
- Rosa, K. D. and Ellen, J. 2009. Text Classification Methodologies Applied to Micro-Text in Military Chat. In *Proceedings of the 2009 international Conference on Machine Learning and Applications* (December 13 - 15, 2009). ICMLA. IEEE Computer Society, Washington, DC, 710-71
- Sharifi, B., et al. (2010). Summarizing Microblogs Automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 685-688, Los Angeles, CA. ACL.
- Tumasjan, A., et al. 2010. Predicting Elections with Twitter: Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Weblogs and Social Media*, AAAI, Washington, D.C., 2010.
- Wang, A. H. 2010. Don't follow me - Spam Detection in Twitter. In *Proceedings of the International Conference on Security and Cryptography (SECRYPT 2010)*. INSTICC. 142-151.
- 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*. ACL, Morristown, NJ, 347-354.