A STUDY OF FEATURE EXRACTION TECHNIQUES FOR CLASSIFYING TOPICS AND SENTIMENTS FROM NEWS POSTS

Wafa Zubair Abdullah Al-Dyani

814383

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)

UNIVERSITI UTARA MALAYSIA

2014

Permission to Use

I'm presenting this thesis in fulfilment of the requirement for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to Universiti Utara Malaysia and to me for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences

UUM College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

ABSTRAK,

Banyak saluran berita mempunyai laman Facebook sendiri yang jawatan berita telah dikeluarkan di harian. Oleh yang demikian, posting berita ini mengandungi pendapat duniawi tentang peristiwa-peristiwa sosial yang mungkin berubah dari masa ke masa disebabkan oleh faktor-faktor luaran serta boleh menggunakan monitor untuk peristiwa-peristiwa penting berlaku seluruh dunia. Hasilnya, banyak teks perlombongan penyelidikan telah dijalankan dalam bidang analisis sentiment sebagaimana satu tugas yang paling mencabar adalah untuk mengesan dan mengeluarkan ciri-ciri utama dari siaran berita yang tiba secara berterusan lebih masa termuka dalam menghasilkan dataset tidak seimbang. Walau bagaimanapun, mengekstrak ciri-ciri ini adalah satu tugas yang mencabar kerana sifat-sifat yang kompleks di post, juga posting tentang topik tertentu mungkin berkembang atau hilang kerja lebih masa. Oleh itu, kajian ini telah membangunkan satu analisis perbandingan mengenai ciri-ciri kaedah pengekstrakan yang mempunyai pelbagai ciri-ciri pengekstrakan teknik (TF-IDF, TF, b, IG, chisquare) dengan tiga ciri n-gram berbeza (Unigram, Bigram, Trigram), dan menggunakan SVM sebagai Pengelas. Tujuan kajian ini adalah untuk mencari yang optimum ciri pengekstrakan teknik (FET) yang dapat mencapai hasil ketepatan optimum untuk topik dan sentimen klasifikasi. Sehubungan dengan itu, analisis ini adalah dijalankan ke atas tiga saluran berita datasets. Keputusan eksperimen bagi topik klasifikasi telah menunjukkan bahawa chisquare dengan unigram telah terbukti menjadi FET yang terbaik berbanding kaedah lain. Selain itu, untuk mengatasi masalah tidak seimbang data, kajian ini telah digabungkan FET ini dengan teknologi OverSampling. Keputusan penilaian telah menunjukkan peningkatan dalam prestasi di Pengelas dan telah mencapai ketepatan yang lebih tinggi pada 93.37%, 92.89% dan 91.92 BBC, Al-Arabiya dan Al-Jazeera, masing-masing, berbanding dengan apa yang telah diperolehi pada dataset asal. Begitu juga, gabungan yang sama telah digunakan untuk pengkelasan sentimen dan memperolehi ketepatan perakaman pada kadar 81.87%, 70.01%, 77.36%. Walau bagaimanapun, ujian yang diiktiraf optimum TFT jawatan dipilih secara rawak berita tersembunyi telah menunjukkan ketepatan perakaman yang agak rendah bagi kedua-dua topik dan sentimen klasifikasi akibat dari beberapa perubahan topik dan sentimen dari masa ke masa.

Kata kunci: Teks perlombongan, klasifikasi teks, analisis sentimen duniawi, teknik pengekstrakan ciri, saluran berita, acara sosial, data yang tidak seimbang.

ABSTRACT

Recently, many news channels have their own Facebook pages in which news posts have been released in a daily basis. Consequently, these news posts contain temporal opinions about social events that may change over time due to external factors as well as may use as a monitor to the significant events happened around the world. As a result, many text mining researches have been conducted in the area of Temporal Sentiment Analysis, which one of its most challenging tasks is to detect and extract the key features from news posts that arrive continuously overtime. However, extracting these features is a challenging task due to post's complex properties, also posts about a specific topic may grow or vanish overtime leading in producing imbalanced datasets. Thus, this study has developed a comparative analysis on feature extraction Techniques which has examined various feature extraction techniques (TF-IDF, TF, BTO, IG, Chi-square) with three different n-gram features (Unigram, Bigram, Trigram), and using SVM as a classifier. The aim of this study is to discover the optimal Feature Extraction Technique (FET) that could achieve optimum accuracy results for both topic and sentiment classification. Accordingly, this analysis is conducted on three news channels' datasets. The experimental results for topic classification have shown that Chi-square with unigram have proven to be the best FET compared to other techniques. Furthermore, to overcome the problem of imbalanced data, this study has combined the best FET with OverSampling technology. The evaluation results have shown an improvement in classifier's performance and has achieved a higher accuracy at 93.37%, 92.89%, and 91.92 for BBC, Al-Arabiya, and Al-Jazeera, respectively, compared to what have been obtained on original datasets. Similarly, same combination (Chi-square+Unigram) has been used for sentiment classification and obtained accuracies at rates of 81.87%, 70.01%, 77.36%. However, testing the recognized optimal FET on unseen randomly selected news posts has shown a relatively very low accuracies for both topic and sentiment classification due to the changes of topics and sentiments over time.

Keywords: Text mining, Text classification, Temporal Sentiment analysis, Feature extraction techniques, News channels, Social events, Imbalanced data.

Acknowledgement

All praise is to Allah, who by His grace and blessings I have completed my thesis. Though only my name appears on the cover of this dissertation, a great many people have contributed to its production. I owe my gratitude to all those people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

My deepest gratitude is to my advisor and supervisor, Dr. Farzana Kabir Ahmad. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, and at the same time the guidance to recover when my steps faltered. Dr. Farzana taught me how to question thoughts and express ideas. His patience and support helped me overcome many crisis situations and finish this dissertation. I hope that one day I would become as good an advisor to my students as she has been to me.

Many friends have helped me stay sane through these difficult years. Their support and care helped me overcome setbacks and stay focused on my graduate study. I greatly value their friendship and I deeply appreciate their belief in me.

Most importantly, none of this would have been possible without the love and patience of my family. My family to whom this dissertation is dedicated to, has been a constant source of love, concern, support and strength all these years. I would like to express my heart-felt gratitude to my family especially to my most important person in my life, my beloved mother who has been a mother and father to me throughout my life and whose without her prays I might not be able to gain what I have achieved until now, so all thanks and gratitude to my dear mother.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations

A	Accuracy
API	Application Programming Interface
BOW	Bag Of Words
ENN	Edited Nearest Neighbour
ETD	Emerging Trend Detection
FE	Feature Extraction
GIBC	General Inquire Based Classifier
HMM	Hidden Markov Model
HTTP	Hyper Text Transfer Protocol
IG	Information Gain
IMDB	Internet Movie DataBase
KDD	Knowledge Data Discovery
KNN	K Nearest Neighbourhood
LDA	Latent Divichlet Allocation
ME	Maximum Entropy
MI	Mutual Information
ML	Machine Learning
MLT	Machine Learning Techniques
NB	Navie Bayes
NFS	New Feature Selection
NLP	Natural Language Processing
OM	Opinion Mining
PMI	Pointwise Mutual Information
POS	Part Of Speech
RBC	Rule Based Classifier
RMSE	Root Mean Square Error
ROS	Random Over Sampling
SA	Sentiment Analysis
SBC	Static Based Classifier
SC	Sentiment Classification
SMOTE	Synthetic Minority Oversampling Technique
SNs	Social Network sites
SVM	Support Vector Machine
SWN	SentiWordNet
TD	Trend Detection
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
TH	ThresHold
TOM	Temporal Opinino Mining
UGC	User Generated Content
WMP	Weighted Mean Precision
WMR	Weighted Mean Recall

CHAPTER ONE

INTRODUCTION

1.1 Overview of Study

Recent years have gained a great attention in the text mining and temporal sentiment analysis research field due to the large amount of opinion data generated in Social Networks sites (SNs) such as Facebook and Twitter. Facebook is the most famous and common SNs among Internet users for expressing their feelings, opinions, emotions and thoughts. Furthermore, Facebook has shown a tremendous increase in usage as it offers a valuable source for real time news and act as an opinions platform [1,2]. Hence, large number of news channels committee have created their own pages on Facebook, to allow news reader to post their opinion and thought on daily news items. The key idea at this point is to gain deep insight about what news readers think and feel towards various events.

Generally, news posts can be used as a monitor mechanism to detect the significant events which have been happening around the world. Furthermore, some events may grow up or vanish over time due to external factors such as change of time, evolution of recent events, or emergence of new events. As a result, such events may affect the overall opinions and consequently change correlated sentiments. Hence, in order to analyze these changes, a new field of sentiment analysis has been emerged in this area which is called Temporal Opinion Mining (TOM). TOM is defined as "a process of detecting and monitoring possible changes to particular opinions and their correlated sentiments over a given period of time and can be seen as a continuation of opinion mining" [3]. The main idea of TOM is to find the opinions average on a specific topic at different times. This analysis leads to

The contents of the thesis is for internal user only

REFERENCES

[1] J. Akaichi, Z. Dhouioui, and M. J. Lopez-Huertas Perez, "Text mining facebook status updates for sentiment classification," in *System Theory, Control and Computing (ICSTCC), 2013 17th International Conference*, 2013, pp. 640–645.

[2] E. N. Neumann, *The spiral of silence Public opinion–our social skin*. Chicago: University of Chicago Press, 1993.

[3] E. Bjørkelund and T. Burnett, "Temporal Opinion Mining," no. June, p. 122, 2012.

[4] T. Fukuhara, H. Nakagawa, and T. Nishida, "Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events.," in *ICWSM*, 2007.

[5] R. Chakraborty, "D OMAIN K EYWORD E XTRACTION T ECHNIQUE : A N EW W EIGHTING M ETHOD," pp. 109–118, 2013.

[6] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, and T. By, "Sentiment Analysis on Social Media," *2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min.*, pp. 919–926, 2012.

[7] H. Bacan, I. S. Pandzic, and D. Gulija, "Automated news item categorization," in *Proceedings of the 19th Annual Conference of The Japanese Society for Artificial Intelligence*, 2005, pp. 251–256.

[8] J. Zhang, Y. Kawai, S. Nakajima, Y. Matsumoto, and K. Tanaka, "Sentiment bias detection in support of news credibility judgment," in *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, 2011, pp. 1–10.

[9] J. Allan, "Introduction to topic detection and tracking," in *Topic detection and tracking*, Springer, 2002, pp. 1–16.

[10] D. Clarke, P. Lane, and P. Hender, "Developing robust models for favourability analysis," in *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, 2011, pp. 44–52.

[11] P. D. Turney, "Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews," no. July, pp. 417–424, 2002.

[12] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Syst.*, vol. 24, no. 7, pp. 1024–1032, Oct. 2011.

[13] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 251–258, 2011.

[14] Z. Fu, X. Sun, J. Shu, and L. Zhou, "Plain Text Zero Knowledge Watermarking Detection Based on Asymmetric Encryption," vol. 48, no. Cia, pp. 126–134, 2014.

[15] L.-W. Ku, Y.-T. Liang, and H.-H. Chen, "Opinion Extraction, Summarization and Tracking in News and Blog Corpora.," in *AAAI spring symposium: Computational approaches to analyzing weblogs*, 2006, vol. 100107.

[16] I. P. Cvijikj and F. Michahelles, "Monitoring trends on Facebook," *Proc. - IEEE 9th Int. Conf. Dependable, Auton. Secur. Comput. DASC 2011*, pp. 895–902, 2011.

[17] M. Cataldi, U. Torino, L. Di Caro, U. Torino, C. Schifanella, and U. Torino, "a4-Cataldi," 2010.

[18] J. Weng, Y. Yao, E. Leonardi, F. Lee, and B. Lee, "Event Detection in Twitter Event Detection in Twitter ∗ ," 2011.

[19] G. Burnside, D. Milioris, and P. Jacquet, "One Day in Twitter: Topic Detection Via Joint Complexity," *Www*, 2014.

[20] S. Greener and A. Rospigliosi, *ePub - European Conference on Social Media: ECSM*, vol. 7. Academic Conferences Limited, 2014.

[21] D. Richter, P. D. D. K. Riemer, and J. vom Brocke, "Internet social networking," *Wirtschaftsinformatik*, vol. 53, no. 2, pp. 89–103, 2011.

[22] S. Setty, R. Jadi, S. Shaikh, C. Mattikalli, and U. Mudenagudi, "Classification of facebook news feeds and sentiment analysis," in *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*, 2014, pp. 18–23.

[23] J. K. Ahkter and S. Soria, "Sentiment analysis: Facebook status messages," *Unpubl. master's thesis, Stanford, CA*, 2010.

[24] A. E.-D. A. Hamouda and F. E. El-taher, "Sentiment Analyzer for Arabic Comments System," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 3, 2013.

[25] N. Bansal and N. Koudas, "BlogScope: a system for online analysis of high volume text streams," *Proc. 33rd Int. Conf. Very large data bases*, pp. 1410–1413, 2007.

[26] H. Choi and H. Varian, "Predicting the present with google trends," *Econ. Rec.*, vol. 88, no. s1, pp. 2–9, 2012.

[27] T. Nguyen, D. Phung, B. Adams, and S. Venkatesh, "Event extraction using behaviors of sentiment signals and burst structure in social media," *Knowl. Inf. Syst.*, vol. 37, no. February, pp. 279–304, 2013.

[28] S. Bai, Y. Ning, S. Yuan, and T. Zhu, "Predicting Reader ' s Emotion," pp. 16–27, 2013.

[29] B. Thomas, "Exploration of Robust Features for Multiclass Emotion Classification," pp. 1704–1709, 2014.

[30] J. Zhang, Y. Kawai, and T. Kumamoto, "Extracting Similar and Opposite News Websites Based on Sentiment Analysis," in *Proc. of International Conference on Industrial and Intelligent Information (ICIII 2012)*, 2012, pp. 24–29.

[31] L. U. Ye and R. Xu, "E Motion Prediction of News Articles From Reader ' S Perspective Based on Multi-Label Classi Fication," pp. 15–17, 2012.

[32] C. G. Patil, "Use of Porter Stemming Algorithm and SVM for Emotion Extraction from News Headlines," vol. 2, no. 7, pp. 9–13.

[33] G. Li and F. Liu, "A clustering-based approach on sentiment analysis," *Proc. 2010 IEEE Int. Conf. Intell. Syst. Knowl. Eng. ISKE 2010*, pp. 331–337, 2010.

[34] J. Kamps, M. J. Marx, R. J. Mokken, and M. De Rijke, "Using wordnet to measure semantic orientations of adjectives," 2004.

[35] J. Staiano and M. Guerini, "DepecheMood: a Lexicon for emotion analysis from crowd-annotated news," *arXiv Prepr. arXiv1405.1605*, 2014.

[36] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Machine Learning: ECML 2004*, Springer, 2004, pp. 39–50.

[37] F. Chang, J. Guo, W. Xu, and K. Yao, "A Feature Selection Method to Handle Imbalanced Data in Text Classification.," *J. Digit. Inf. Manag.*, vol. 13, no. 3, p. 169, 2015.

[38] A. Zughrat, M. Mahfouf, Y. Y. Yang, and S. Thornton, "Support Vector Machines for Class Imbalance Rail Data Classification with Bootstrapping-based Over-Sampling and Under-Sampling," in *19th World Congress of the International Federation of Automatic Control. Cape Town*, 2014.

[39] P. G. Preethi and V. Uma, "Temporal Sentiment Analysis and Causal Rules Extraction from Tweets for Event Prediction," *Procedia Comput. Sci.*, vol. 48, pp. 84–89, 2015.

[40] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news," *Proc. Seventh Int. Conf. Lang. Resour. Eval.*, pp. 2216–2220, 2010.

[41] S.-M. Kim and E. Hovy, "Extracting opinions, opinion holders, and topics expressed in online news media text," in *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, 2006, pp. 1–8.

[42] R. H. W. Pinheiro, G. D. C. Cavalcanti, R. F. Correa, and T. I. Ren, "A global-ranking local feature selection method for text categorization," *Expert Syst. Appl.*, vol. 39, no. 17, pp. 12851–12857, 2012.

[43] "Top News on Facebook | Fan Page List." [Online]. Available: http://fanpagelist.com/category/news/view/list/sort/fans/page1. [Accessed: 06-Dec-2015].

[44] "BBC World News achieves major distribution milestone, reaching more than 330m households worldwide," 2012.

[45] "About BBC World News TV," 2011.

[46] "Media Use in the Middle East 2013 | Northwestern University in Qatar." [Online]. Available: http://menamediasurvey.northwestern.edu/. [Accessed: 09-Dec-2015].

[47] T. Johnson and S. Fahmy, "Who is winning the hearts and minds of the Arab public?," *Int. Commun. Res. J.*, vol. 45, no. 1–2, pp. 24–48, 2010.

[48] "Major Events in 2014, What Happened in 2014." [Online]. Available: http://www.mapsofworld.com/events/year-2014/. [Accessed: 09-Dec-2015].

[49] "The Biggest News Stories of 2014 - ABC News." [Online]. Available: http://abcnews.go.com/International/biggest-news-stories-2014/story?id=27466867. [Accessed: 09-Dec-2015].

[50] "The 10 Biggest International Stories of 2014 - The Atlantic." [Online]. Available: http://www.theatlantic.com/international/archive/2014/12/the-10-biggest-international-stories-of-2014/383935/. [Accessed: 09-Dec-2015].

[51] "2014 Year in Review | Facebook Newsroom." [Online]. Available: http://newsroom.fb.com/news/2014/12/2014-year-in-review/. [Accessed: 09-Dec-2015].

[52] "Facebook's most talked-about topics of 2014 - CBS News." [Online]. Available: http://www.cbsnews.com/news/facebooks-most-talked-about-topics-of-2014/. [Accessed: 09-Dec-2015].

[53] "Twitter's top tweets and retweets of 2014: Ellen, World Cup score big - TODAY.com." [Online]. Available: http://www.today.com/money/twitters-top-tweets-retweets-2014-ellen-world-cup-score-big-1D80349123. [Accessed: 09-Dec-2015].

[54] "Twitter And Facebook Launch Their 2014 'Year In Review' With Top Content, Trends & More." [Online]. Available: http://marketingland.com/twitter-facebook-launch-2014-year-review-top-content-trends-110643. [Accessed: 09-Dec-2015].

[55] "Facebook," 2011. [Online]. Available: http://www.facebook.com/.

[56] "Facebook," 2011. [Online]. Available: http://www.facebook.com/press/info.php?statistics.

[57] "Harpsocial," p. http://www.harpsocial.com/2011/04/social–medias–sh, 2011.

[58] "Twitter," p. http://ww.twitter.com/, 2011.

[59] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," *VAST 10 - IEEE Conf. Vis. Anal. Sci. Technol. 2010, Proc.*, pp. 115–122, 2010.

[60] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 11, pp. 2169–2188, 2009.

[61] A. Shrivastava and B. Pant, "Opinion extraction and classification of real time Facebook Status," *Glob. J. Comput. Sci. Technol.*, vol. 12, no. 8, 2012.

[62] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 591–600.

[63] M. R. Morris, J. Teevan, and K. Panovich, "What Do People Ask Their Social Networks, and Why?," *Chi*, vol. 69, p. 1739, 2010.

[64] A. Kothari, W. Magdy, K. Darwish, A. Mourad, and A. Taei, "Detecting Comments on News Articles in Microblogs."

[65] C. Lin, Y. He, and R. Everson, "Sentence subjectivity detection with weakly-supervised learning," pp. 1153–1161, 2011.

[66] S. Amer-Yahia, S. Anjum, A. Ghenai, A. Siddique, S. Abbar, S. Madden, A. Marcus, and M. El-Haddad, "Maqsa: a system for social analytics on news," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 653–656.

[67] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.

[68] B. Liu, "Sentiment analysis and subjectivity," *Handb. Nat. Lang. Process.*, vol. 2, pp. 627–666, 2010.

[69] B. Liu, "Sentiment Analysis and Opinion Mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012.

[70] A. Montoyo, P. MartíNez-Barco, and A. Balahur, "Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments," *Decis. Support Syst.*, vol. 53, no. 4, pp. 675–679, 2012.

[71] M. Sadegh, R. Ibrahim, and Z. A. Othman, “Opinion mining and sentiment analysis: A survey,” *Int. J. Comput. Technol.*, vol. 2, no. 3, pp. 171–178, 2012.

[72] P. Case and G. D. V, “Opinion Mining and Classification of User Reviews in Social Media,” vol. 7782, pp. 37–41, 2014.

[73] M. Tsytsarau and T. Palpanas, “Survey on mining subjective data on the web,” *Data Min. Knowl. Discov.*, vol. 24, no. 3, pp. 478–514, 2012.

[74] L.-W. Ku, L.-Y. Lee, T.-H. Wu, and H.-H. Chen, “Major topic detection and its application to opinion summarization,” in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, pp. 627–628.

[75] S.-M. Kim and E. Hovy, “Automatic detection of opinion bearing words and sentences,” in *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2005, pp. 61–66.

[76] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, “Sentiment analysis in the news,” *arXiv Prepr. arXiv1309.6202*, 2013.

[77] G. Vinodhini and R. Chandrasekaran, “Sentiment Analysis and Opinion Mining: A Survey,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 2, no. 6, pp. 282–292, 2012.

[78] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.

[79] N. Isa, M. Puteh, R. Mohamad, and H. Raja, “Sentiment Classification of Malay Newspaper Using Immune Network (SCIN),” vol. III, 2013.

[80] G. Jaganadh, “Opinion mining and Sentiment analysis CSI communication,” 2012.

[81] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, “Introduction to WordNet: An On-line Lexical Database *,” *Int. J. Lexicogr.*, vol. 3, no. 4, pp. 235–244, Jan. 1990.

[82] S. Mohammad, C. Dunne, and B. Dorr, “Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, 2009, pp. 599–608.

[83] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” *Proc. Conf. Empir. Methods Nat. Lang. Process. July 6-7, 2002, Philadephia, Pennsylvania, USA*, pp. 79–86, 2002.

[84] H. Tang, S. Tan, and X. Cheng, "A survey on sentiment detection of reviews," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10760–10773, 2009.

[85] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.," in *LREC*, 2010, vol. 10, pp. 2200–2204.

[86] D. Das, A. K. Kolya, A. Ekbal, and S. Bandyopadhyay, "Temporal analysis of sentiment events–a visual realization and tracking," in *Computational Linguistics and Intelligent Text Processing*, Springer, 2011, pp. 417–428.

[87] G. Mishne and M. De Rijke, "MoodViews: Tools for Blog Mood Analysis.," in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 153–154.

[88] G. B. Tran, M. Alrifai, I. A. Intelligence, and N. Language, "Predicting Relevant News Events for Timeline Summaries," *Www*, pp. 91–92, 2013.

[89] D. Bhattacharya and S. Ram, "Sharing news articles using 140 characters: A diffusion analysis on twitter," *Proc. 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2012*, pp. 966–971, 2012.

[90] "Top 10 Arabic YouTube Channels in the Middle East | IstiZada." [Online]. Available: http://istizada.com/blog/top-10-arabic-youtube-channels/. [Accessed: 09-Dec-2015].

[91] "Al_Jazeera_English," 2015. [Online]. Available: https://en.wikipedia.org/wiki/Al_Jazeera_English. [Accessed: 06-Dec-2015].

[92] "Al_Arabiya," 2015. [Online]. Available: https://en.wikipedia.org/wiki/Al_Arabiya#cite_note-cablegatesearch1-17. [Accessed: 06-Dec-2015].

[93] "Media Use in the Middle East:An Eight-Nation Survey - NU-Q." [Online]. Available: http://www.scribd.com/doc/137906439/Media-Use-in-the-Middle-East-An-Eight-Nation-Survey-NU-Q. [Accessed: 09-Dec-2015].

[94] J. Kleinnijenhuis, F. Schultz, D. Oegema, and W. van Atteveldt, "Financial news and market panics in the age of high-frequency sentiment trading algorithms," *Journalism*, p. 1464884912468375, 2013.

[95] L.-C. Yu, J.-L. Wu, P.-C. Chang, and H.-S. Chu, "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news," *Knowledge-Based Syst.*, vol. 41, pp. 89–97, Mar. 2013.

[96] J. Teevan, D. Ramage, and M. R. Morris, "# TwitterSearch: a comparison of microblog search and web search," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 35–44.

[97] "Billion-Dollar Weather and Climate Disasters: Overview | National Climatic Data Center (NCDC)." [Online]. Available: http://www.ncdc.noaa.gov/billions/. [Accessed: 03-Apr-2015].

[98] Wikipedia, "Diseases and disorders," 2015. [Online]. Available: http://en.wikipedia.org/wiki/Disease. [Accessed: 17-Apr-2015].

[99] Wikipedia, "Terrorism," 2015. [Online]. Available: http://en.wikipedia.org/wiki/Terrorism. [Accessed: 19-Apr-2015].

[100] Ask.com, "what are the effects of terrorist attacks," 2015. [Online]. Available: http://www.ask.com/. [Accessed: 01-Jan-2015].

[101] Wikipedia, "Conflict_(process)," 2014. [Online]. Available: http://en.wikipedia.org/wiki/Conflict_(process). [Accessed: 22-Dec-2014].

[102] Collinsdictionary, "plane-crash," 2015. [Online]. Available: http://www.collinsdictionary.com/dictionary/english/plane-crash. [Accessed: 01-Jan-2015].

[103] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*, 2008, pp. 1556–1560.

[104] R. Plutchik, "A general psychoevolutionary theory of emotion," *Theor. Emot.*, vol. 1, 1980.

[105] "List of Human Emotions - List of Human Emotions." [Online]. Available: http://www.listofhumanemotions.com/listofhumanemotions. [Accessed: 21-May-2015].

[106] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 37–45.

[107] A. Kontostathis, L. M. Galitsky, W. M. Pottenger, S. Roy, and D. J. Phelps, "A survey of emerging trend detection in textual data mining," in *Survey of Text Mining*, Springer, 2004, pp. 185–224.

[108] "Poll results: Top languages for analytics/data mining programming." [Online]. Available: http://www.kdnuggets.com/2012/08/poll-analytics-data-mining-programming-languages.html. [Accessed: 23-Dec-2015].

[109] "RapidMiner at CeBIT 2010: the Enterprise Edition, Rapid-I and Cloud Mining - Data Mining - Blog.com." [Online]. Available: http://www.data-mining-blog.com/cloud-mining/rapidminer-cebit-2010/. [Accessed: 09-Dec-2015].

[110] "RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series): Markus Hofmann, Ralf Klinkenberg: 9781482205497: Amazon.com: Books." [Online]. Available: http://www.amazon.com/RapidMiner-Analytics-Applications-Knowledge-Discovery/dp/1482205491. [Accessed: 09-Dec-2015].

[111] "KDnuggets Annual Software Poll:RapidMiner and R vie for first place." [Online]. Available: http://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html. [Accessed: 09-Dec-2015].

[112] "Rexer Analytics 5th Annual Data Miner Survey - 2011." [Online]. Available: http://www.rexeranalytics.com/Data-Miner-Survey-Results-2011.html. [Accessed: 09-Dec-2015].

[113] "German Predictive Analytics Startup Rapid-I Rebrands As RapidMiner, Takes $5M From Open Ocean, Earlybird To Tackle The U.S. Market." [Online]. Available: http://techcrunch.com/2013/11/04/german-predictive-analytics-startup-rapid-i-rebrands-as-rapidminer-takes-5m-from-open-ocean-earlybird-to-tackle-the-u-s-market/. [Accessed: 09-Dec-2015].

[114] F. Ben Abdesslem, I. Parris, and T. Henderson, "Reliable online social network data collection," in *Computational Social Networks*, Springer, 2012, pp. 183–210.

[115] B. Rieder, "Studying Facebook via data extraction: the Netvizz application," in *Proceedings of the 5th Annual ACM Web Science Conference*, 2013, pp. 346–355.

[116] C. Troussas, M. Virvou, K. Junshean Espinosa, K. Llaguno, and J. Caro, "Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning," in *Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference on*, 2013, pp. 1–6.

[117] A. Shrivatava and B. Pant, "Opinion Extraction and Classification of Real Time Facebook Status," vol. 12, no. 8, 2012.

[118] R. Rogers, "The end of the virtual," 2009.

[119] C. Cesarano, B. Dorr, A. Picariello, D. Reforgiato, A. Sagoff, and V. Subrahmanian, "Oasys: An opinion analysis system," in *AAAI spring symposium on computational approaches to analyzing weblogs*, 2004.

[120] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988.

[121] "TheySay PreCeive API Demo." [Online]. Available: http://apidemo.theysay.io/. [Accessed: 10-Dec-2015].

[122] "Python NLTK Sentiment Analysis with Text Classification Demo." [Online]. Available: http://text-processing.com/demo/sentiment/. [Accessed: 10-Dec-2015].

[123] "DepecheMood - Try Our Online Demo!" [Online]. Available: http://www.depechemood.eu/DepecheMood.html. [Accessed: 10-Dec-2015].

[124] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, 2007, pp. 70–74.

[125] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013.

[126] D. Meyer, K. Hornik, and I. Feinerer, "Text mining infrastructure in R," *J. Stat. Softw.*, vol. 25, no. 5, pp. 1–54, 2008.

[127] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[128] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[129] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, 2002.

[130] Y. Mejova and P. Srinivasan, "Exploring Feature Definition and Selection for Sentiment Classifiers," pp. 546–549, 2011.

[131] S. Li and C. Zong, "A new approach to feature selection for text categorization," in *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*, 2005, pp. 626–630.

[132] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM Sigkdd Explor. Newsl.*, vol. 6, no. 1, pp. 80–89, 2004.

[133] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguist.*, vol. 37, no. 2, pp. 267–307, 2011.

[134] J. S. Kessler and N. Nicolov, "Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations.," in *ICWSM*, 2009.

[135] J. Kleinberg, "Bursty and hierarchical structure in streams," *Data Min. Knowl. Discov.*, vol. 7.4, pp. 373–397.

[136] K. Balog, G. Mishne, and M. De Rijke, "Why are they excited?: identifying and explaining spikes in blog mood levels," in *Proceedings of the Eleventh*

Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations, 2006, pp. 207–210.

[137] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[138] D. Maynard and A. Funk, “Automatic detection of political opinions in Tweets,” pp. 81–92.

[139] B. Ohana and B. Tierney, “Sentiment classification of reviews using SentiWordNet,” in *9th. IT & T Conference*, 2009, p. 13.

[140] P. Chaovalit and L. Zhou, “Movie review mining: A comparison between supervised and unsupervised classification approaches,” in *System Sciences, 2005. HICSS’05. Proceedings of the 38th Annual Hawaii International Conference on*, 2005, p. 112c–112c.

[141] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[142] K. T. Durant and M. D. Smith, “Mining sentiment classification from political web logs,” in *Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006), Philadelphia, PA*, 2006.

[143] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” *CS224N Proj. Report, Stanford*, pp. 1–12, 2009.

[144] R. Zheng, J. Li, H. Chen, and Z. Huang, “A framework for authorship identification of online messages: Writing- style features and classification techniques,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 378–393, 2006.

[145] Q. Ye, Z. Zhang, and R. Law, “Sentiment classification of online reviews to travel destinations by supervised machine learning approaches,” *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6527–6535, 2009.

[146] H. Cui, V. Mittal, and M. Datar, “Comparative experiments on sentiment classification for online product reviews,” in *AAAI*, 2006, vol. 6, pp. 1265–1270.

[147] E. Airoldi, X. Bai, and R. Padman, “Markov blankets and meta-heuristics search: Sentiment extraction from unstructured texts,” in *Advances in Web Mining and Web Usage Analysis*, Springer, 2006, pp. 167–187.

[148] B. Xu, T.-J. Zhao, D.-Q. Zheng, and S.-Y. Wang, “Product features mining based on conditional random fields model,” in *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, 2010, vol. 6, pp. 3353–3357.

[149] D. K. Kirange and R. R. Deshmukh, "Emotion Classification of News Headlines Using Svm," vol. 5, pp. 104–106, 2012.

[150] N. V Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *ACM Sigkdd Explor. Newsl.*, vol. 6, no. 1, pp. 1–6, 2004.

[151] Y. Y. Yang, M. Mahfouf, G. Panoutsos, Q. Zhang, and S. Thornton, "Adaptive neural-fuzzy inference system for classification of rail quality data with bootstrapping-based over-sampling," in *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, 2011, pp. 2205–2212.

[152] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, 2003.

[153] D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and naive bayes," in *ICML*, 1999, vol. 99, pp. 258–267.

[154] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM Sigkdd Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, 2004.

[155] M. Porter, "The Porter stemming algorithm, 2005," *See http//www. tartarus. org/~ martin/PorterStemmer*.

[156] W. B. Frakes, "Information Retrieval: CHAPTER 8: STEMMING ALGORITHMS." [Online]. Available: http://dns.uls.cl/~ej/daa_08/Algoritmos/books/book5/chap08.htm. [Accessed: 17-Dec-2015].