International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 8, Issue 2, Feb - Mar 2019; 11-20

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

WHAT ARE YOU EXPRESSING? SENTIMENTAL ANALYSIS ON SOCIAL NETWORK

Adil Khan, Izhar Khan, Tahreem Akhtar, Arshi Fahim & Babar Ushmani
Research Scholar, Al Falah University, Haryana, India

ABSTRACT

In this paper, we focus on the emotions of the human, which they are trying to express by posting messages on the social network. Social media is filled with the user-generated microblogs and processing these blogs is very challenging. We have processed the human language in such a manner that our system can understand the emotions of the human that they are trying to express either using text or emoticons. From our research and experimental results on two real-life datasets, the system will able to understand the human sentiments after analyzing their write-ups available on their social profile.

KEYWORDS: *Microblogging, Emotion, Hashtagged, Gammon, Part of Speech, Twitter, Social Network*

Article History

Received: *21 Feb 2019* **|** ***Revised:*** *04 Mar 2019* **|** ***Accepted:*** *19 Mar 2019*

INTRODUCTION

The way people communicate and receive. The information has undergone a radical transformation in recent years with the invention of social networks. In the publication of a web content report of February 2007, Shel Holtz defines social networks as "covering (a) the wide range of channels used by the users of the network that generate their own content.” [1] Citizen journalism. Popularized through blogs, wikis, vlogs, and podcasting, these are examples of social networks. According to Kevin Allen, co-creator of the web content report, "When good things happen in good companies, in the fight against them, online criticism can be the difference between fat, fire, and four alarms." [2] Speaking of social networks and bloggers Stephen Baker and Heather Green of Business Week, they say that corporations "cannot be together because they are the most explosive explosion in the world of information, from the Internet itself." [3] Nowadays, the use of social networks as a communication tool. It is not part of most organizations. Emergency plans, but pay attention to this "explosive rupture" [4] are increasingly important for the ability of the organization to survive. Just like people the collection and creation of information begin to change, emergency communicators must take the initiative to re-evaluate. The way they disseminate the information they speak to voters and respond to public opinion or become a public symbol of national dissatisfaction.

Twitter – Microblogging Social Network

Microblogging sites have evolved to become a diverse source of information. This is due to the nature of microblogs in which people publish in real-time news about your opinions on various topics, discuss current problems, complain and express their opinions. People also write their opinions on the products they use in everyday life. In fact, companies that produce such products have started to check these microblogs to get a general idea for that product. Many times, these companies research and react to user reactions. One of the challenges is to create technology to detect and summarize general feelings. [7]

This is called Twitter and build models for classification."Tweets" in a positive, negative and neutral mood. We create models for two classification tasks: The binary task of classifying feelings as positive and negative classes and the task of tripartite classification Positive, negative and neutral feelings.

Our experiments show this the Unigram model is infact a difficult foundation to achieve. More than 20% based on the chance for both classifications Tasks Our model is based on features that it only uses 100 attributes to achieve precision similar to the unigram A model that uses more than 10,000 functions. Our tree the core-based model exceeds these two models. With a large margin. [8] We also experiment with a combination of models: a combination of unigrams with our characteristics and combination of our qualities with the tree. Core. Both combinations exceed the unigram. Baseline by more than 4% for both classifications Tasks In this article, we present a comprehensive function. Analysis of 100 features that we propose. Our experiments Show which features you need to measure the specific features of Twitter (emoticons, hashtags, etc.) Add value to the classifier, but only marginally. Characteristic that combine the previous polarization of words from them Parts of speech labels are the most important for both Classification tasks. In this way, we see the natural standard. Tools for language processing are useful even in a species that differs from the genre in which they were trained (newswire). Also we also shows that the tree's kernel model performs approximately, as well as the best models based on features, although it does not require a detailed engineering function.[9] We use manually annotated Twitter data thirty experiments the advantage of these data, compared to the previous ones. The data sets used are that tweets are collected. In the streaming method and therefore represent A real sample of real tweets in the language field. Use and content our new data set is available for other researchers. We also present this article two available resources author): 1) dictionary annotated manually for emoticons maps emoticons to their polarization and 2) Dictionary of acronyms collected from the Internet using frequently used translations in English from over 5000. [6]

Previous Work and Surveys

The analysis of feelings is a growing field of natural language. Processing with tests that go from classification to document level. We focused on polarity words and phrases. Given the limitations of character in tweets classify the sentiment of messages on Twitter is more like the analysis of feelings at the sentence level. [11] However, an informal and specialized language is used in tweets, and the very nature of microblogging. The domain makes the analysis of moods on Twitter is very different. [12]

It was not until last year that a series of documents appeared looking at feelings and rumors about Twitter. Other researchers have begun to study the use of part-speech functions, but the results stay in the mix Common microblogging features[5](for example emoticons) are also common, but there were few of them. Research on the usefulness of existing feelings resources developed in data other than microblogging. Researchers also began to explore various forms automatically collect training data. Several researchers trust emoticons to define your emotion or feelings. We use existing Twitter sites to collect training data. Also, uses of hashtags to create training data, but limit their experiments to classify feelings / non-feelings instead. Classification of three polar polarities, like us. [13]

The feeling analysis was treated as natural language processing on many levels of detail. From being a document level classification task was considered at the prayer level, and more recently level of opinion microblogging data, such as Twitter, in which users publish. [14]

Reactions in real time and opinions on "everything", poses new and different challenges. Some first and last results in the analysis of moods. The data on Twitter comes from Go et al. (2009), (Berminghamand Smeaton, 2010) and Pak and Paroubek (2010).

Go et al. [15] (2009) uses distance learning to acquire feelings data they use tweets that end with positive emoticons as ":)" ":-)" as positive and negative emoticons as ":(" ":-(" as negative creates models using Naive Bayes, MaxEnt and Support Vector Machines (SVM) and report exceeded SVM results. Other classifiers in terms of feature space, they try the Unigram model, Bigram with functions of the speech part (POS). [16] They notice it the Unigram model surpasses all other models. In particular, the Bigrams and POS functions do not help. Pak and Paroubek (2010) collect data after a similar paradigm of learning. However, different classification tasks: subjective versus objective. They collect for subjective data tweets that end with emoticons in the same way. In the case of objective data, they are tracked. Twitter accounts from popular newspapers such as "Time of India"," Washington Messages "etc. Report that POS and bigrams help. [17]

Both of these Approaches, however, are based mainly on ngrams In addition, the data they use for training and Search tests collect evidence, and therefore are [18]On the contrary, we present features that can be achieved significant increase compared to the initial value for the one-piece. Additional we are investigating a different method of data representation. And report significant improvement in Unigram models. Another contribution of this work. [20] It is that we report results in manually annotated data. It does not suffer from known prejudices. Our Data is a random sample of broadcast tweets as opposed to Data collected through detailed consultations. Size Our hand-tagged data allows us to cross-check experiments and check the variance in Performance classification through pleats. Another significant effort in the classification of feelings. [21] On Twitter, the data comes from Barbosa and Feng (2010). They use polarization predictions from three websites as loud labels for model training and use 1000 manually. Tweets marked for tuning and another 1000 manually. Tweets marked for the test. They, however Not to mention how they collect their test data.[19] They suggest using the syntax functions of such tweets as Retweet, hashtags, link, punctuation, and exclamation. Brands combined with features such as pre-polarity. Words and POS words. We are expanding your approach using the actual value of the previous polarization and by combining the previous polarity with POS. Our results show that features that improve our performance most classifiers are functions that connect the previous one’s Polarity of words with their parts of speech. Syntax functions help tweet, but only marginally. [10]Gammon (2004) performs an analysis of sentiments in feedback data from the Global Support Services survey. One of the goals of his work is job analysis. 31 with language features such as POS markers. There Extensive analysis of features and selection of features and Prove that the features of abstract linguistic analysis. It contributes to the accuracy of the classifier. In this role, we carried out a comprehensive analysis of the features and showed it uses only 100 abstract linguistic features as well as a hard-baseline. [23]

Data

Twitter is a social website and microblogging. [22] A service that allows users to publish messages in real time, called tweets. Tweets are short, limited messages up to 140 characters long. Because of the nature of this Microblogging service (short and short messages). People use acronyms, they make mistakes in spelling, and they use Emoticons and other characters that express promotions. Meaning there is a shortly related terminology. With tweets. Emoticons: its facial expressions. Represented graphically according to punctuation. And letters; they express the user's mood. Goal: Twitter users use the "@" symbol to refer to other Users on the microblog. Referring to other users in this way, it automatically warns them. Hashtags: Users often use hashtags to mark topics. This is mainly done to increase your visibility tweets we bought 11,875 Twitter annotations manually. Data (tweets) from a commercial source. They have He submitted part of his publicly available data. Information on how to get the data, see Acknowledgments Section at the end of the article. [14] They gathered Data archiving real-time flow. No language location or any other type of restriction Taken during the transmission process. In fact, the collection consists of tweets in foreign languages. They use Google Translate to convert it to English. Before the annotation process. Each tweet is labeled the shooter as a positive, negative, neutral man or trash the "garbage" label means that the tweet cannot understand the shooter. Manual Analysis of a random sample of tagged tweets. As "trash" he suggested that many of these tweets were those that have not been translated well using Google translate we eliminate tweets with junk tags. For experiments. This leaves us with an imbalance. A sample of 8,753 tweets. [17] We use layered sampling. To obtain a balanced set of 5127 tweets (1709 Tweets of each of the positive, negative and neutral classes). [25]

Hashtagged (“ # ”)

The hashtag data set is a subset of Twitter in Edinburgh body the edimburski body contains 97 million tweets. Receive in two months. To create a hashtag. [24]Data set, we first filtered duplicate tweets, not English Tweets and tweets that do not contain hashtags. We examined the remaining part (about 4 million) Distribution of hashtags and identification of what we hope. Frequent sets of labels that indicate positive, negative, and neutral messages. These hashtags are for selection Tweets that will be used for development and training. Table 2 contains the 15 most-used hashtags in Edinburgh body In addition to the very popular hashtags that are part of it. Twitter community (e.g. #followfriday, #musicmonday), we find hashtags that seem to point to a message. Polarization: #fail, #omgthatsotrue, #iloveitwhen, etc. To select the final set of messages that will be included in The HASH data set identifies all hashtags that appear at least 1000 times in the corps in Edinburgh. [28] We choose them Better hashtags that we think will be the most useful. To identify positive, negative and neutral tweets. These hashtags are shown in Table 3. Messages with these hashtags they were included in the final data set and polarization of each of them the message is defined by its hashtag.

Emoticon

The Emoticon data set was created by Go, Bhayani, and Huang for a project at Stanford University, collecting tweets with positive ":)" and negative ":(". Containing both positive and negative emoticons. In addition, many tweets have been manually tagged for use for evaluation, but for our experiments, we use them only training data. This set contains 381,381 tweets, 230,811 positive and 150,570 negatives. Interestingly, most of these messages do not contain any hashtags. [27]

Isieve

iSieve data contain about 4,000 tweets. It was like that Compiled and commented manually by iSieve Corporation. The data in this collection has been selected for specific topics, and the label of each tweet reflects your feelings (positive, negative or neutral) towards tweet. We use this Data set for evaluation only [26].

Resources and Pre-Processing of Data

We use three different corpora of Twitter messages in our experiments. For development and training, we use the hashtagged data set (HASH), which we compile from the Edinburgh Twitter corpus1, and the emoticon data set (EMOT) from http://twittersentiment and appspot.com. For evaluation, we use a manually annotated data set produced by the iSieve

Corporation2 (ISIEVE). The number of Twitter messages and the distribution across classes is given in Table 1.

Table 1: Corpus Statistics

	Positive	Negative	Neutral	Total
HASH	31,861 (14%)	64,850 (29%)	125,859 (57%)	222,570
EMOT	230.81 (61%)	150,570 (39%)	-	381,381
ISIEVE	1,520 (38%)	200 (5%)	2,295 (57%)	4,015

Table 2: Top Positive, Negative and Neutral Hashtags used to Create the HASH Data Set

Positive	#iloveitwhen, #thingsilike, #bestfeeling, #bestfeelingever, #omgthatssotrue, #imthankfulfor, #thingsilove, #success
Negative	#fail, #epicfail, #nevertrust, #worst, #worse, #worstlies, #imtiredof, #itsnotokay, #worstfeeling, #notcute, #somethingaintright, #somethingsnotright, #ihate
Neutral	#job, #tweetajob, #omgfacts, #news, #listeningto, #lastfm, #hiring, #cnn

Table 3: Most Frequent Hashtags in the Edinburgh Corpus

Hashtag	Frequency	Synonyms
#followfriday	226,530	#ff
#nowplaying	209,970	
#job	136,734	#tweetajob
#fb	106,814	#facebook
#mm	78,585	#minimilitia
#tinychat	56,376	
#tcot	42,110	
#quote	33,554	
#letsbehonest	32,732	#tobehonest
#omgfacts	30,042	
#fail	23,007	#epicfail
#factsaboutme	19,167	
#news	17,190	
#random	17,180	
#shoutout	16,446	

We pre-process all the tweets as follows: a) replace all the emoticons with their sentiment polarity by looking up the emoticon dictionary, b) replace all URLs with a tag ||U||, c) replace targets (e.g. "@John") with tag ||T||, d) replace all negations (e.g. not, no, never, don't, cannot) by tag "NOT", and e) replace a sequence of repeated characters by three characters, for example, convert cooooooooool to coool.

Table 4: Example Acrynoms and their Expansion in the Acronym Dictionary

Acronym	English Expansion
gr8, gr8t	great
lol	laughing out loud
rolf	rolling on the floor
bff	best friend forever

We present some preliminary statistics about the data in Table 3. We use the Stanford tokenizer to tokenize the tweets. We use a stop word dictionary3 to identify stop words. All the other words which are found in WordNet are counted as English words. We use the standard tag set defined by the Penn Treebank for identifying punctuation. We record the occurrence of three standard twitter tags: emoticons, URLs and targets. The remaining tokens are either non-English words (like coool, zzz etc.) or other symbols.

Table 5: Part of the Dictionary of Emoticons

Emoticon	Polarity
:-) :) :o) :] :3 :c)	Positive
:D C:	Extremely-Positive
:-(:(:c :[	Negative
D8 D; D= DX v.v	Extremely-Negative
:|	Neutral

Pre-processing of data consists of three stages: 1), atomization, 2) normalization and 3) speech marking (POS). Emoticons and abbreviations (eg OMG WTF, BRB) are identified as part of the tokens and treated as individual tokens. In the case of the normalization process, the presence of abbreviations is saved in a tweet, and the abbreviations are replaced by their actual meaning (for example, BRB -> is backward). We have also identified as informal amplifiers of all letters (see eg "I love this show!!! and repetitions of characters (p. Eg, I have a loan !! happyyyyyy"), save its presence in a tweet. Capitalized words are repeated characters are replaced. Finally, the presence of special Twitter tokens (eg #hashtags, user tags and URL) is followed by substitutes, which are given to indicate the type of token, and we hope that this normalization improves the performance of POS Tagger, which is the last step ahead.

Features

In our classification experiments, we use different characteristics. We use the unigram and bigram for the baseline. We also include features typically used in the analysis of feelings, namely, traits that represent information about feelings Lexicon and POS functions. Finally, we add functions to Capturesome of the most specific languages in microblogging.

N-Gram Features

539 n-gram functions to identify a set of useful n-grams, we first remove keywords. Then we detected elemental refusals by attaching a word to a word that precedes or follows Deadline for refusal. This turned out to be useful in the previous one Work (Pak and Paroubek 2010). Finally, all unigrams and Bigrams are identified in the training data and classified by your information gain, measured using Chi-square. For our experiments, we use the best 1000 n-grams in a bag of words moda. [3]

Lexicon Features

Characteristics of the lexicon the words detailed the MPQA lexicon of subjectivity (Wilson, Wiebe and Hoffmann 2009) are marked with an earlier polarization: Positive, negative or neutral. We create three features. Based on the presence of any word in the dictionary

Part-of-Speech Features

Characteristics of parts of speech for every tweet we have a numerical characteristic. Verbs, adverbs, adjectives, nouns and every other part speak Microblogging functions.

Micro-Blogging Features

We create binary features that capture the presence of positives, Negative and neutral emoticons, and abbreviations as well the presence of amplifiers (e.g. all uppercase letters and character repeats). For emoticons and shortcuts, we use a shortcut Internet Lingo Dictionary (Wasden 2006) and several internet Jargon dictionaries available online.

Experiment

Our goal of these experiments is twofold. We want first evaluate whether our training data with labels come from Hashtags and emoticons are useful for training affection classifiers. On Twitter. Secondly, we want to evaluate their effectiveness. Sectional characteristics for sentiment analysis in Data from Twitter. How useful is the lexicon of developed feelings? To the formal text with short and informal tweets? How much is it Do we use the specificity of the domain? In our first set of experiments, we use HASH and EMOT data sets. We started with a random 10% sample collection HASH data to use as a validation set. This set of validation is it is used to select the n-gram characteristic and to adjust the parameters. The rest of the HASH data is for training. AND we trained the classifier, we collected 22,274 tweets of training. Data and use this data to train AdaBoost.MH (Schapire and Singer 2000) models with 500 rebar rounds.

Because the EMOT data set has no neutral data and ours the experiments include 3-way classification, which is not included. In the initial experiments. Instead, we check if it is it is useful to use EMOT data to extend the HASH and data improve the classification of feelings. 19,000 messages from the EMOT data set, divided equally between positive and negative, they are randomly selected and added to the HASH and data Experiments are repeated. To be aware of the upper-performance limit we can expect from models trained by HASH and if Inclusion of EMOT data may first cause improvement Check the model results in the validation set. Figure 1 shows the average F value for the n-gram baseline and for all Characteristics of HASH and HASH + EMOT data. In this data, by adding EMOT data for training, leads to Improvements, especially when all functions are used. Returning to the test data, we evaluate the trained models. In HASH and HASH + EMOT data in ISIEVE data set. Figure 2 shows the average F measurement for the baseline and four combinations of features: n-grams and lexicon. Features (n-gram + lex), n-grams and functions of discourse parts (n-gram + POS), n-gram, lexical and microblogging functions (n-grams + lex + twit), and finally all functions set Figure 3 shows the accuracy of the same experiments. Interestingly, the best performance in the evaluation data. It comes from the use of n-grams together with a lexicon. Characteristics and characteristics of microblogging. In this part of speech, the functions really give a drop in performance. If this is due to the accuracy of the POS tagger Tweets or POS tags are less useful in microblogging the data will require further investigation. In addition, although it contains EMOT data for training gives good performance improvement in the absence of microblogging functions when there are microblogging functions Incorporated, improvements are falling or disappearing. Best results in the assessment data, n-grams, lexical are derived. And Twitter functions only trained with marked data.

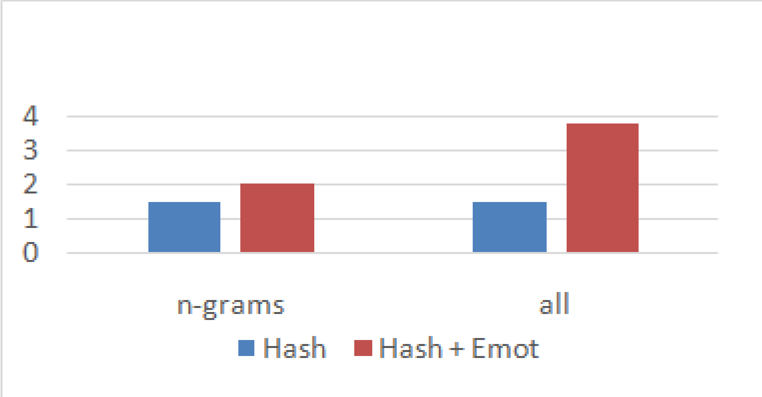

Figure 1: Average F-Measure on the Validation Set over Models Trained on the HASH and HASH+EMOT Data

CONCLUSIONS

Our experiments in analyzing moods on Twitter show that the characteristics of parts of speech may not be useful for the analysis of feelings in the microblogging domain. Further research is needed to determine if POS features are of low quality Due to the tagger results or if POS features they are less useful to analyze feelings in this area. The characteristics of the lexicon of the existing feeling were something useful in combination with microblogging functions, but microblogging properties (ie the presence of amplifiers and positive/negative/neutral emoticons and abbreviations)Apparently the most useful one. Using hashtags to collect training data has been useful, and the use of collected data based on positive and negative results, However, which method gives the best results training data and two training data sources they are complementary and may depend on the type of functions used. Our experiments show that when there are microblogging functions this takes into account the reduction of the benefits of emoticon training data.

REFERENCES

1. *M Hu and B Liu. 2004. Mining and summarizing customer reviews. KDD.*
2. *S M Kim and E Hovy. 2004. Determining the sentiment of opinions. Coling.*
3. *Adam Bermingham and Alan Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage is brevity an advantage? ACM, pages 1833–1836.*
4. *T. Wilson, J. Wiebe, and P. Hoffman. 2005. Recognizing contextual polarity in phrase level sentiment analysis.*
5. *Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of the 20th international conference on Computational Linguistics.*
6. *Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.*
7. *David Haussler. 1999. Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz.*
8. *Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 24–32, March.*

9. *Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.*
10. *C M Whissel. 1989. The dictionary of Affect in Language. Emotion: theory research and experience, Acad press London.*
11. *Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In Proceedings of the 17th European Conference on Machine Learning.*
12. *Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.*
13. *B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts. ACL.*
14. *P. Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. ACL.*
15. *Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pages 423–430.*
16. *ACL. Tumasjan, A.; Sprenger, T. O.; Sandner, P.; and Welpe, I. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of ICWSM.*
17. *Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In Proc. of Coling.*
18. *Bifet, A., and Frank, E. 2010. Sentiment knowledge discovery in twitter streaming data. In Proc. of 13th International Conference on Discovery Science.*
19. *Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of Coling.*
20. *Esuli, A., and Sebastiani, F. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of LREC.*
21. *Schapire, R. E., and Singer, Y. 2000. BoosTexter: A boosting-based system for text categorization. Machine Learning 39(2/3):135–168.*
22. *Jansen, B. J.; Zhang, M.; Sobel, K.; and Chowdury, A. 2009. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology 60(11):2169–2188.*
23. *Kim, S.-M., and Hovy, E. 2004. Determining the sentiment of opinions. In Proceedings of Coling.*
24. *O'Connor, B.; Balasubramanyan, R.; Routledge, B.; and Smith, N. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of ICWSM.*
25. *Pak, A., and Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Proc. of LREC.*
26. *Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2):1–135.*

27. *Hatzivassiloglou, V., and McKeown, K. 1997. Predicting the semantic orientation of adjectives. In Proc. of ACL.*
28. *C. Fellbaum. 1998. Wordnet, an electronic lexical database. MIT Press.*