Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 296 · 2016

Informatyka i Ekonometria 6

Wiesław Wolny

University of Economics w Katowice
Faculty of Informatics and Communication
Department of Informatics
wieslaw.wolny@ue.katowice.pl

SENTIMENT ANALYSIS OF TWITTER DATA USING EMOTICONS AND EMOJI IDEOGRAMS

Summary: Twitter is an online social networking service where worldwide users publish their opinions on a variety of topics, discuss current issues, complain, and express positive or negative sentiment for products they use in daily life. Therefore, Twitter is a rich source of data for opinion mining and sentiment analysis. However, sentiment analysis for Twitter messages (tweets) is regarded as a challenging problem because tweets are short and informal. This paper focuses on this problem by the analyzing of symbols called emotion tokens, including emotion symbols (e.g. emoticons and emoji ideograms). According to observation, these emotion tokens are commonly used. They directly express one's emotions regardless of his/her language, hence they have become a useful signal for sentiment analysis on multilingual tweets. The paper describes the approach to performing sentiment analysis, that is able to determine positive, negative and neutral sentiments for a tested topic.

Keywords: Twitter, sentiment analysis, symbol analysis, SAS.

Introduction

Microblogging websites such as Twitter (www.twitter.com) have evolved to become a great source of various kinds of information. This is due to the nature of microblogs on which people post real time messages regarding their opinions on a variety of topics, discuss current issues, complain, and express positive or negative sentiment for products they use in daily life.

As the audience of microblogging platforms and social networks grows every day, data from these sources can be used in opinion mining and sentiment

analysis tasks. For example, manufacturing companies may be interested in the following questions:

- What do people think about our product (service, company, etc.)?
- How positive (or negative) are people about our product?
- What would people prefer our product to be like?

Political parties may be interested to know if people support their program or not. Social organizations may ask people's opinion on current debates. All this information can be obtained from social networks, as their users post everyday what they like/dislike, and their opinions on many aspects of their life.

Opinions and its related concepts such as sentiments and emotions are the subjects of study of sentiment analysis and opinion mining. The inception and rapid growth of the field coincide with those of the social media on the Web, e.g., reviews, forum discussions, blogs, microblogs, Twitter, and social networks. Most NLP based methods perform without particular success in social media. Almost all forms of social media are very noisy and full of all kinds of spelling, grammatical, and punctuation errors.

This article proposes method of providing sentiment analysis using such data as Twitter hashtags, e.g. #happy, #fail, emoticons, e.g. :-), :-(, :-| and emoji characters, e.g. 😄😂😉✌💪to identify positive, negative and neutral tweets.

1. Related work

Sentiment analysis is a growing area of the Natural Language Processing task at many levels of granularity. Starting from being a document level classification task [1], [2] it has been handled at the sentence level [3], [4] and more recently at the phrase level [5], [6] or even polarity of words and phrases (e.g., [7], [8]).

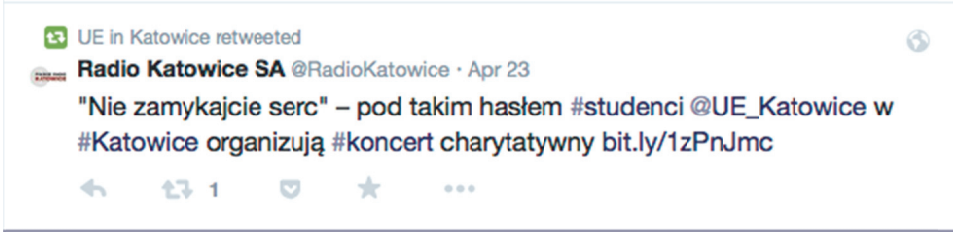
However, the informal and specialized language that is used in tweets, as well as the nature of the microblogging domain make sentiment analysis in Twitter a very different task. With the growing population of blogs and social networks, opinion mining and sentiment analysis have become a field of interest for many researches. A very broad overview of the existing work was presented in [9]. J. Read in [10] used emoticons such as ":-)" and ":-(" to form a training set for the sentiment classification. For this purpose, the authors collected texts containing emoticons from Usenet newsgroups. The dataset was divided into "positive" (texts with happy emoticons) and "negative" (texts with sad or angry emoticons) samples.

Researchers have also begun to investigate various ways of automatically collecting training data. Several researchers have relied on emoticons to define training data [11], [12]. Barbarosa and Feng [13] exploited existing Twitter sentiment sites for collecting training data. Davidov, Tsur, and Rappoport [14] also used hashtags to create training data, but they limited their experiments to sentiment/non-sentiment classification, rather than 3-way polarity classification, as in this article.

2. Data description and collection

Twitter has its own conventions that renders it distinct from other textual data. The Twitter messages are called tweets. There are some particular features that can be used to compose a tweet (Figure 1).

Figure 1. Example of a tweet



The first pieces of information ”UE in Katowice retweeted” means the tweet was forwarded from a previous post. “@RadioKatowice” and ”@UE_Katowice” are a twitter name of Radio Katowice and University of Economics in Katowice. Using twitter names in tweet sends information the user about mention of them. #Katowice, #studenci and #koncert are a tags provided by the user for this message, so-called hashtags and “bit.ly/1zPnJmc” is a link to some external source. Length of tweets is limited, therefore long links are often shortened using special websites like bitly.com.

Users of Twitter use the “@” symbol to refer to other users. Referring to other users in this manner automatically alerts them. Users usually use hashtags to mark topics. This is primarily done to increase the visibility of their tweets. These symbols gives easy way to identify Twitter user names and topics and thus allows searching and filtering of information on any subject.

Twitter messages have many unique attributes, which differentiates twitter analysis from other fields of research. First is length. The maximum length of a Twitter message is 140 characters. Average length of tweet is 14 words [15].

This is different from the domains of other research, which were mostly focused on reviews which consisted of multiple sentences. The second attribute is availability of data. With the Twitter API or other tools, it is much easier to collect millions of tweets for training.

2.1. Emoticons

There are two fundamental data mining tasks that can be considered in conjunction with Twitter data: text analysis and symbol analysis. Due to the nature of this microblogging service (quick and short messages), people use acronyms, make spelling mistakes, use emoticons and other characters that express special meanings. Emoticons are metacommunicative pictorial representation of a facial expression pictorially represented using punctuation and letters or pictures; they express the user's mood.

The use of emoticons can be tracked back to the 19th century. The first documented person to have used the emoticons :-) and :-(on the Internet was Scott Fahlman from Carnegie Mellon University in a message dated 19 September 1982.

Some emoticons as a characters are included in the Unicode standard – three in the Miscellaneous Symbols block, and over sixty in the Emoticons block [16].

Emoticons can be categorized as:

- Happy emoticons : :-) :) :D :o) :] :3 :c) :> =] 8), etc.
- Sad emoticons: >:[:-(:(:-c :c :-< :っC :< :-[:[:{, etc.
- Neutral emoticons: >:\ >:/ :-/ :-. :/ :\ =/ =\ :L =L :S >.<, etc.

More symbols and meanings like angry, crying, surprise can be found on Wikipedia site [17], which can be used to determine their emotional state. The top 20 of emoticons collected from 96 269 892 tweets is presented in [18].

2.2. Emoji ideograms

Emoji were originally used in Japanese electronic messages and spreading outside of Japan. The characters are used much like emoticons, although a wider range is provided. The rise of popularity of emoji is due to its being incorporated into sets of characters available in mobile phones. Apple in IOS, Android and other mobile operating systems included some emoji character sets. Emoji characters are also included in the Unicode standard [19]. Emoji can be categorized into same categories as emoticons. Emoji can be even translated to English using http://emojitranslate.com/.

2.3. Data collection

The main problem is how to extract the rich information that is available on Twitter and how can it be used to draw meaningful insights. To achieve this, first we need to build an accurate sentiment analyzer for tweets, which is what this solution aims to achieve. As a software to data analyze can be used SAS Text Miner, SAS Visual Analytics or other tools. The challenge remains to fetch customized Tweets and clean data before any text or symbol mining. SAS Visual Analytics allows direct import of Twitter data, but to use SAS Text Miner and other tools, data have to be downloaded and converted.

Twitter allows developers to collect data via Twitter REST API [20] and The Streaming API [21]. Twitter has numerous regulations and rate limits imposed on its API, and for this reason it requires that all users must register an account and provide authentication details when they query the API. This registration requires users to provide an email address and telephone number for verification, once the user account is verified the user will be issued with the authentication detail which allows access to the API.

Unfortunately Twitter API exports data only in JSON format, which need to be translated to readable for databases or analytical software format. A combination of Twitter API, scripts for converting JSON to CSV [22], SAS Macro [23] or Excel Macro [24] can be used to extract information from twitter and create an input dataset for the analysis. The entire process of data acquisition can be fully automated by scheduling the run of Visual Basic for Applications (VBA) or SAS macros. Since opinions have targets, further pre-processing and filtering of collected data can be done using @twitter_names and #hashtags as a targets in the way described in [20]. This method is more precise and provides better result than other text mining approaches.

3. Sentiment analysis

Sentiment analysis which is also known as opinion mining, focuses on discovering patterns in the text that can be analyzed to classify the sentiment in that text. The term sentiment analysis probably first appeared in [25], and the term opinion mining first appeared in [26]. However, the research on sentiments and opinions appeared earlier.

Liu stated that “Sentiment analysis is the field of study that analyses people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, and their attributes. It

represents a large problem space. There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc." [27, p. 7]. Sentiment analysis has grown to be one of the most active research fields in natural language processing. It is also widely studied in data mining, Web mining, and text mining. In fact, it has spread from computer science to management sciences and social sciences due to its importance to business and society.

Sentiment analysis is predominantly implemented in software which can autonomously extract emotions and opinions from a text. It has many real world applications it allows companies to analyze how their products or brand is being perceived by their consumers, politicians may be interested in knowing how people will vote in elections, etc. It is difficult to classify sentiment analysis as one specific field of study as it incorporates many different areas such as linguistics, Natural Language Processing (NLP), and Machine Learning or Artificial Intelligence. As the majority of the sentiment that is uploaded to the internet is of an unstructured nature it is a difficult task for computers to process it and extract meaningful information from it. Some of the most effective machine learning algorithms, e.g., support vector machines, naïve Bayes and conditional random fields, produce no human understandable results.

Emotions are closely related to sentiments. Emotions can be defined as a subjective feelings and thoughts. People's emotions have been categorized into some distinct categories. However, there is still not a set of agreed basic emotions among researchers. Based on [28], people have six primary emotions, i.e., love, joy, surprise, anger, sadness, and fear, which can be sub-divided into many secondary and tertiary emotions. Each emotion can also have different intensities. Emotions in virtual communication differ in a variety of ways from those in face-to-face interactions due to the characteristics of computer mediated communication. Computer mediated communication may lack many of the auditory and visual cues normally associated with the emotional aspects of interactions. While text-based communication eliminates audio and visual cues, there are other methods for adding emotion. Emoticons, or emotional icons, can be used to display various types of emotions.

For purposes of this work, sentiment can be defined as a personal positive, neutral or negative opinion. Classification is done in supervised learning using lexicon-based approach. The sentiment lexicon contains a list of sentiment emoticons and emoji ideograms. Opinions can be gathered by searching Twitter posts using Twitter API. Each tweet can labelled, using emoticons and emoji icons, as

positive, negative, neutral or junk. The "junk" label means that the tweet cannot be understood. In order to use this method an assumption must be made, this assumption is that the emoticon in the tweet represents the overall sentiment contained in that tweet. This assumption is quite reasonable as the maximum length of a tweet is 140 characters so in the majority of cases the emoticon will correctly represent the overall sentiment of that tweet. This kind of evaluation is commonly known as the document-level sentiment classification because it considers the whole document as a basic information unit.

Model can be developed on a sample of data; this can be used to classify sentiments of the tweet. Manual classification will be done on a sample of tweets. Accuracy of model can be tested against validating sample. Tweets assigned manually will be divided into 2 parts – 80% of data should be taken in Model sample and 20% of data should be taken as validating sample. Results obtained will be compared with the manually assigned classification.

Conclusions

Microblogging like twitter nowadays became one of the major types of the communication. The large amount of information contained in these web-sites makes them an attractive source of data for opinion mining and sentiment analysis. Most text based methods of analysis may not be useful for sentiment analysis in these domains. To make a significant progress, we still need novel ideas. Using twitter names and hashtags to collect training data can provide better results. Also adding symbol analysis using emoticons and emoji characters can significantly increase the precision of recognizing of emotions. The most successful algorithms will be probably integration of natural language processing methods and symbol analysis.

References

[1] S. Das, M. Chen, *Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards*, "Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)" 2001, Vol. 35.

[2] P.D. Turney, *Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews* [in:] *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, PA 2002, pp. 417-424.

[3] M. Hu, B. Liu, *Mining and Summarizing Customer Reviews* [in:] *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'04, ACM, New York, NY 2004, pp. 168-177.

[4] S. Kim, E. Hovy, *Determining the Sentiment of Opinions* [in:] *COLING '04 Proceedings of the 20th international conference on Computational Linguistics*, Geneva 2004.

[5] A. Agarwal, F. Biadsy, K. McKeown, *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, Athens 2009, pp. 24-32.

[6] T. Wilson, J. Wiebe, P. Hoffmann, *Recognizing Contextual Polarity in Phrase--Level Sentiment Analysis* [in:] *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, MIT Press, Cambridge, MA 2005, pp. 399-433.

[7] A. Esuli, F. Sebastiani, *Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining*, "Proceedings of LREC" 2006, Vol. 6.

[8] V. Hatzivassiloglou, K.R. McKeown, *Predicting the Semantic Orientation of Adjectives* [in:] *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Madrid 1997, pp. 174-181.

[9] B. Pang, L. Lee, *Opinion Mining and Sentiment Analysis*, "Foundations and Trends in Information Retrieval" 2008, Vol. 2(1-2), pp. 1-135.

[10] J. Read, *Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification* [in:] *Proceedings of the ACL Student Research Workshop (ACLstudent '05)*, Association for Computational Linguistics, Stroudsburg, PA 2005, pp. 43-48.

[11] A. Pak, P. Paroubek, *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*, "LREC" 2010, Vol. 10.

[12] A. Bifet, E. Frank, *Sentiment Knowledge Discovery in Twitter Streaming Data*, Discovery Science, Springer, Berlin-Heidelberg 2010.

[13] L. Barbosa, J. Feng, *Robust Sentiment Detection on Twitter from Biased and Noisy Data* [in:] *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, Association for Computational Linguistics, Bejjing 2010, pp. 36-44.

[14] D. Davidov, O. Tsur, A. Rappoport, *Enhanced Sentiment Learning Using Twitter Hashtags and Smileys* [in:] *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, Association for Computational Linguistics, Bejjing 2010, pp. 241-249.

[15] A. Go, R. Bhayani, L. Huang, *Twitter Sentiment Classification Using Distant Supervision*, CS224N Project Report, Stanford 2009, pp. 1-12.

[16] Unicode Miscellaneous Symbols, http://www.unicode.org/charts/PDF/U2600.pdf (accessed: May 2015).

[17] List of emoticons, Wikipedia, http://en.wikipedia.org/wiki/List_of_emoticons (accessed: May 2015).

[18] N. Berry, *DataGenetics*, http://www.datagenetics.com/blog/october52012/index.html (accessed: May 2015).

[19] *Unicode Emoji, Draft Unicode Technical Report #51*, http://www.unicode.org/reports/tr51/ (accessed: May 2015).

[20] Twitter REST API, *The Search API*, https://dev.twitter.com/rest/public/search (accessed: May 2015).

[21] Twitter, *The Streaming APIs*, https://dev.twitter.com/streaming/overview (accessed: May 2015).

[22] S. Falko, *How to Import Twitter Tweets in SAS DATA Step Using OAuth 2 Authentication Style*, http://blogs.sas.com/content/sascom/2013/12/12/how-to-import-twitter-tweets-in-sas-data-step-using-oauth-2-authentication-style (accessed: May 2015).

[23] S. Garla, G. Chakraborty, *%GetTweet: A New SAS® Macro to Fetch and Summarize Tweets*, Paper 324-2011, Oklahoma State University, Stillwater, OK 2001.

[24] *Twitter Text Mining Using SAS*, Social Media Analytics, http://www.analytics-tools.com/2012/06/social-media-analytics-twitter-text.html (accessed: May 2015).

[25] T. Nasukawa, J. Yi, *Sentiment Analysis: Capturing Favorability Using Natural Language Processing* [in:] *Proceedings of the K-CAP-03, 2nd International Conference on Knowledge Capture*, Sanibel Island, FL 2003, pp. 70-77.

[26] K. Dave, S. Lawrence, D.M. Pennock, *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews* [in:] *Proceedings of International Conference on World Wide Web*, Budapest 2003, pp. 519-528.

[27] B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, Williston, VT 2012.

[28] W.G. Parrott (ed.), *Emotions in Social Psychology*: *Essential Readings*, Key Reading in Social Psychology, Psychology Press, Philadelphia, PA 2001.

ANALIZA WYDŹWIĘKU DANYCH Z TWITTERA Z WYKORZYSTANIEM EMOTIKONÓW I EMOJI

Streszczenie: Twitter jest ogólnoświatowym serwisem, w którym użytkownicy publikują swoje opinie na różne tematy, dyskutują na temat bieżących wydarzeń oraz wyrażają pozytywne bądź negatywne opinie o produktach, których używają w codziennym życiu. Z tego powodu Twitter jest potężnym źródłem danych do badania opinii i analizy wydźwięku. Jednak analiza wydźwięku komunikatów na Twiterze (tweetów) uważana jest za problem, będący zarazem wyzwaniem, z powodu niewielkiej objętości tekstu tweetów i często nieformalnego charakteru ich języka. Artykuł skupia się na analizie symboli znanych jako emotikony i emoji. Zgodnie z przeprowadzonymi badaniami, symbole te są powszechnie używane w komunikacji za pomocą Twittera. Wyrażają one bezpośrednio emocje niezależnie od języka, dlatego mogą być używane w wielojęzycznych tekstach. W artykule przedstawiono podejście do analizy wydźwięku umożliwiającej określenie pozytywnego, negatywnego lub neutralnego wydźwięku badanych tekstów.

Słowa kluczowe: Twitter, analiza wydźwięku, analiza symboli, SAS.