International Journal of Engineering & Technology, 7 (4.19) (2018) 884-888

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Events Tagging in Twitter Using Twitter Latent Dirichlet Allocation

Ghaidaa A. Al-Sultany*[1], Hiba J. Aleqabie[2]

[1]Department of Information Network, [2]Department of Software
[1, 2]College of Information Technology, University of Babylon, Babel, Iraq
**Corresponding Author E-mail:ghaidaa.bilal@ itnet.uobabylon.edu.iq,*

Abstract

Twitter has become a great platform to publish and carrying news, advisements, events, topics and even daily events in our lives. Twitter Post has limitations on the length and noise. These limitations make that the post is unsuitable for topic modeling due to sparsity. In this paper, Twitter Latent Dirichlet allocation (TLDA) method for topics modeling was applied to overcome the sparsity problem of tweets modeling. Many steps were implemented for event tagging on Twitter. First: construct a dataset by hashtag pooling technique, and then the preprocessing was performed to extract the features. Secondly, find the suitable number of topics through Perplexity criterion, then, the topics are labeled by WordNet lexicon. Finally, events are tagging using Pricewise Mutual Information (PMI) criterion. The dataset is constructed about various topics including the American elections, Football world cup 2018, and a natural phenomenon and many others; the number of tweets is 63458. This study shows good results in training tweets dataset.

Keywords: *Twitter, TLDA, PMI, and Perplexity.*

1. Introduction

Micro-blogging platforms stand like Twitter have witnessed a rapid and impressive expansion and creating a new way of communication among individuals[1]. The availability of these micro-blogging services had pushed forward the explosion of social data. Twitter is considering nowadays as a significant source of news, live events, affection, and thoughts; it is a productive environment for research and studies, full of different events[2].
The topic modeling's problem is addressed. This undertaking raises various difficulties due to the short, noisy and unstructured language of tweets. The LDA's applications to tweets give incoherent topics[1]. TwitterLDA made a couple of identical alteration.1.) each tweet should map only to one topic. Instead of being mapped to multiple topics.2.)the tokens were categorized as background and topic. The first category includes symbols and stopwords emotions and slang words. The second category is the words represent a topic[2]. Proposing twitter Lda (TLDA) that overcome the problems. It needs to be enhanced with integrity techniques such as pooling to handle tweets corpus[3]. Pooling schemes are presented to group tweets and merge the related tweets into one document that will feed the TLDA model as a training set to work better able to discover topics efficiently[4][5]. Pooling techniques are diverse: basic, user, trend, geo-temporal, hashtag and mixed schemes. Out of that pooling, an integration of similar tweets was done and assign them into a single document; these documents will train using TLDA[3].
The primary goal of the proposed system is to choose the best words in the resulting topic to represent the tag of the topic. That means reducing the number of words in the topic Through Symantec and PMI.
Also, intended to enhance the aggregation process of the tweets. Classify them based on query terms (keywords and hashtags). Then assigning generic label represent the whole story about each topic, create a tag for that topic that indicates an event for each. We examine an evaluation use perplexity criteria to find suitable no. of topics.
The organization of the paper as follows: as Section II Twitter brief description about twitter. Section III semantic lexicon. Section IV depicts the previously related researches on twitterLDA. Section V set for the material and methods. Section VI presents experiential results and discussion. Finally, conclusion and the future work in section VII.

2. Twitter

Twitter is a microblogging utility, posting tweets with 140-characters limit. Posts with noisy content may include relevant information. The plurality (85 %) of trending topics is dealing with news or daily news, Twitter is considered as a news beat for reporters[6].
On Twitter, people tend to follow each other .the concept of "follow," and "follower" is presented. Unlike other social networks services, the following activity does not demand any exchange. Users may follow each other and might be followed by some others. A follower on Twitter implies that the user gets everyone's messages (called tweets) from those users that follow him. Tweets responding had developed into a common application that is the "RT" represents retweet, '#'with word represents a hashtag, and '@' with a name represent used id. The retweet techniques permit users to share tweets without asking to share it[7].

3. Semantic Lexican [8]

WordNet is one of a natural processing resources, which consider as a significant one that researches may depend on.it is a vast

lexical database of English. Nouns, verbs, adjectives, and adverbs are gathered into sets of equivalent intellectual words (synsets), each term a meaningful idea.
WordNet cover many semantic relations, some of them :
• Synonymy is WorldNet's essential relation it utilizes a list of synonyms " synsets" to clarify word's senses. it is a uniform relationship, a word, and its senses
• Antonymy also uniforms relation between words, mostly in dealing with regulating the adjectives and adverbs concepts.
• Hyponymy (sub-name) and hypernymy (super-name), are transitive relations between synsets. Since there is typically only one hypernym, this semantic relation arranges the meanings of nouns into a hierarchical organization.
• Meronymy (part-name) and/ holonymy (whole-name), are complex semantic relations. WordNet discriminates component parts, substantive parts, and member parts.
• Troponymy (manner-name) is for verbs what hyponymy is for nouns.
• Entailment relations between verbs.
All of these semantics relations can clarify by "pointer" through words and its synsets. There are over 116,000 pointers represent those semantics. Relational theories of lexical semantics conclude that each word can be explained in term of other words.

4. Literature Survey

Many related types of research in twitter pooling process [4] propose a method for improving the learning of topics by using different pooling schemes for tweet streaming without modifying the standard Lda. Moreover, create automatic label for the hashtag. While [9] proposed several pooling techniques to overcome the problems of data twitter gathering which increase the coherence of topics. [5]collecting tweets based on a conversation pooling technique then use two topic modeling techniques to train the tweets, the LDA, and Author-Topic model (ATM), The conversation pooling technique shows better performance than any other pooling technique. [3] Suggest an approach for pooling operation by merging information retrieval and LDA topic modeling, similar tweets were collected, and then improve the topic coherence by clustering.
When using LDA to Twitter content, it suffers from sparse and noisy due to the nature of Twitter, such as posts are short messages, mixed of URLs, tags, times, and ids, and use informal language with misspelling, acronyms, and nonstandard abbreviations. Due to these limitations, TwitterLDA was proposed in[10]. [11] Propos a method for exploring topic modeling by considering the Twitter-Lda for a collection of discrete data. Then evaluate this technique from the perspective of classifications.[10]show experimentally comparative research between Twitter with a traditional news medium, the New York Times, applying topic modeling technique and discuss the links between the tweet and retweets and their types.
In labeling topic modeling[12] propose a method for automatic labeling for topics of Lda based on Wikipedia title articles, but different annotation generated for each topic and use (ML) for solving the problem. [13] proposed a novel approach for extracting conceptual label for topics based on WorldNet.
In this research hashtag pooling techniques were applied in addition to the default techniques for streaming tweets. The perplexity criteria iteration of the model to find the suitable number of the topics; parameter to the TLDA. Create labels for the topic of TLDA, using WorldNet lexicon. Tagging event using PMI scores was maintained.

5. Methods and Material

The proposed system consists of six stages. First of all constructing the dataset, applying twitter preprocessing on this dataset, predicting the number of topics, train the data via TLDA, labeling the topics, and finally tagging events. See Fig.1.

5.1. Twitter Dataset

Twitter dataset acquisition is the challenge. Because of the limitation that Twitter Terms of Services (TOS) forced on to redistribute the tweets dataset. For that resonant, no sufficient dataset was found that could be suitable for event detection or tagging. All the previous researches construct the tweets dataset through streaming API Twitter TOS allow to share the Tweet ID and User ID, their labels only. In this research constructing the tweet dataset via streaming API. Two different techniques were used, the keyword base and hashtag base.

5.2. Preprocessing

The next step is Tweet pre-processing after data collection. There are several stages in preprocessing:
- Tokenization its turn the texts into segmented words, digits, letters called tokens. Eliminate blank spaces and ppunctuations.
- Stop-words Removal: exclude the fewer value words value like is, are, in, with, and so forth.
- Url removal: exclude all the URLs from the tweets post if any.
- Mention removal: exclude all the mention that tweet post may contain, i.e., @name

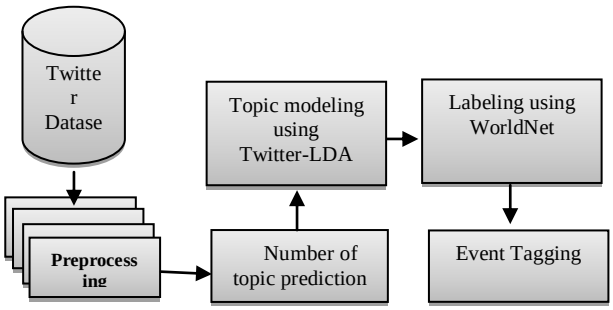

Fig. 1: The Proposed System

RT removal: many tweets are retweeted, so it is redundant and removes it.
- Remove the non-alphabetic words: remove all non-aalphabetic words and the non-English words that may appear in the tweet.
- Stemming: is the way toward distinguishing the root/stem, by expelling various additions of the words.
- Text Transformation: mapping process is applied to finding the frequency of the words and tagging the tokens using Part Of Speech(POS), that assigns a label for each token (word). Selecting words of label nouns, choosing the only noun to precede processing.

5.3. Number of Topics Prediction

Perplexity is a statistical criterion of how quite well a probability model foresees a sample. It shows the strength of the model by computing the inverse log-likelihood of unseen documents, as in (1). Lower perplexity indicates good model[14].

$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log\ p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\} \quad (1)$$

Where, w_d: words in document d;
N_d: Length of document d

A series of TLDA executions were accomplished. Each execution, with a different value of K; a number of topics. For the values (10, 20, 30, 40, 50). Every single examination the perplexity was obtained. Creating set of values. Searching for the lowest value of this set, the lower the value, the better, the topic number which has the lowest perplexity score generally signifies the optimal number of topics. The results were illustrated in figure 4.

5.4. TLDA Topic Model

Lda is a technique for an unsupervised method that detects latent topics in a vast number of documents. Using the principle of "bag of words "which map each word into an id and convert each document into a vector if words frequent.
Every document is characterized by the probability distribution over some topics, while every topic is characterized by a probability distribution over some words[5]. The standard LDA may not function in a right way when trains tweet dataset due to the short posts. To beat this trouble, some researchers suggest aggregating all the tweets as a single document. An effective variant of the standard LDA called Twitter LDA[15] .
An assumption was made that there are T topics in Twitter, each denoted by a word distribution. Let φ^t be the

1. Draw $\phi^B \sim \text{Dir}(\beta)$,$\pi \sim \text{Dir}(\gamma)$
2. For each topic t=1,…,T,
   (a) draw $\phi^t \sim \text{Dir}(\beta)$
3. For each user u=1,…,U,
   (a) draw $\theta^u \sim \text{Dir}(\alpha)$
   (b) for each tweet s=1,…,N_u
   i. draw $z_{u,s} \sim \text{Multi}(\theta^u)$
   ii. for each words n=1,…,$N_{u,s}$
   A. draw $y_{u,s,n} \sim \text{Multi}(\pi)$
   B. draw $w_{u,s,n}$ Multi(ϕ^B)
   If $y_{u,s,n}$ =0 and $w_{u,s,n} \sim$ Multi($\phi^{z_{u,s}}$) if $y_{u,s,n}$=1

Fig. 2: the generation process of topics [15]

Word distribution for topic t and φ^B the word distribution for background words. Let θ^u denote the topic distribution of user u. Let π denote a Bernoulli distribution that governs the choice between background words and topic words[15]. The generation process of tweets is described in Fig. 2. Sample The TLDA topics shown in Fig 3.

5.5. Labeling using WordNet

Creating labels are performed by the word co-occurrence matrix of the top-ranked tokens in each topic. For each topic result from TLDA
For each topic:
1. Retrieve all tweets that top 10 tokens were seen.
2. For each token, find its synsets using WordNet synsets relation.
3. Compute the similarity between tokens by WordNet similarity relation.
4. Get the half top maximum similar tokens.
5. Utilize WordNet to extract the definition of those results from 3, using WordNet definition relation.
6. For each definition try to find the common and shared meaning.
7. Assign the definition as a label for that topic.

5.6. Tagging events

Tagging events from topics are performed by pointwise mutual information (PMI) which tends to evaluate the quality of inferred topics based on the top ten words of each topic. PMI has a famous tendency to give unnecessary scores of relatedness to word pairs that involve low-frequency words; PMI is defined as in (2):

$$PMI = \log \frac{p(a,b)}{p(a)p(b)} \quad (2)$$

Where P (a) and p (b)) are the probability that word occurs in a text window of a given size while p(a, b) denotes the probability both a and b appear together [12].Through this step an attempt to determine the tag word in filtering process through :

Topic #0:	Topic #3:	Topic #6:	Topic #8:
0.076 * "audience"	0.062 * "work"	0.092 * "leader"	0.103*"woman" 0.068*"election"
0.048 * "semi"	0.060 * "team"	0.042 * "moon"	0.033*"today"
0.039 * "point"	0.057 * "wolf"	0.033 * "dear"	0.032*"post" 0.029*"criticism"
0.038 * "press"	0.035 * "try"	0.020 * "blood"	0.027*"role" 0.024*"opportunity"
0.031 * "show"	0.033 * "night"	0.020 * "harry"	0.020*"hope" 0.017*"voting"
0.024 * "week"	0.028 * "trip"	0.019*"compassion"	0.0016*"offer"
0.018 * "picture"	0.027 * "year"	0.017 * "need"	
0.015 * "sense"	0.025 * "seats"	0.017 * "tell"	
0.014 * "interview"	0.025 * "places"	0.016 * "luck"	
0.012 * "answer"	0.020 *"interest"	0.011 * "hours"	

Fig 3. Samples of TLDA Topics

For each topic
1) Any word that has many synonyms in the topic is considered as the desired word.
2) The co-occurrence of every word with other words is paired and calculates their frequent appearance in the tweets retrieved. The word that has a frequent appearance in the high-revived Tweets are taken.

6. Experiential, Discussion and Evaluation

6.1. Dataset

The dataset was collected via Twitter API. The Twitter API platform offers options for streaming real-time Tweets. Each option provides a varying number of filters. This step is performed by a python package (Tweepy). Tweepy tries to make authentication (OAuth). To begin the process, we need to register our client application with Twitter. Create a new application and once complete, consumer token and secret should be taken[2].
This dataset consists of two types of tweets: keyword base and hashtag base query. The construction based on the query term to search for tweets to collect them together. Each query collects its tweet in a single document and labels the tweets by its query term. The dataset was constructed in 2018, about various topics including the American elections, Football world cup 2018, a natural phenomenon and many others; The number of tweets collected is 63458.

6.2 Preprocessing

The tweets were saved in JSON file format. The tweet's text would be extracted only to be processed. These tweets were preprocessed by first removing the punctuations, stop words, numbers, mentions, URLs links and hashtags. Filter out the word with length less than

three characters. Apply Stem operation on those tokens using PorterStemmer. Using the POS to classify the tokens and labels them into classes noun, verb, adjective…etc. The final step in preprocessing is, checking each token if it is an English word or not, by searching for token's entry in English dictionary. Exclude all the meaningless words.

6.3 Experiments

The run of the TLDA initially starts at number of topic =10. Perplexity was using to estimate the number of topics, according to Fig.3 the best value for number of topics was 27.Rerun the system by 27 to achieve stability. As well perplexity considered as the evaluation of the topic model TLDA. See Fig.4
The top-ranked tokens for each topic. Top-ranked was based on their probability distribution.
For each topic, Labels were established through the use of the WordNet lexicon, by finding the similarity between all the tokens. Acquire the most similar tokens and get the common definition between them. This definition refers to the topic and will help to indicate the tags for events later. For this research, according to Fig.3, the topic 0 has the following tokens 'audience', 'point', 'show', 'semi' , 'press' ,'result' ,'sense' ,'week' ,'something',' hand'. The similarity was computed to find that the words 'audience', 'show', 'press', are related to each other, and the common definition was entertaining. The results of all the topic are illustrated in Table 1.
PMI has been utilized for discovering collocations and relationships between words, i.e., Checking of events and co-events of words in a content corpus can be utilized to rough the probabilities p(x) p(x, y). Table 1 shows counts of pairs of words for each topic, and the PMI scores.
The tagging is performed by computing the PMI for both single occurrences and mutual co-occurrences of noun phrases. PMI is computed twice, PMI I for the singular appearance of the token with the rest of the words on each tweet. Moreover, PMI II for the mutual appearance of the noun phrases in tweets. Find the subscriber of the pairs by PMI I and II and compare the value and the best is a tag for that topic. Finally, for all the topics retrieve the tweets for both the tokens in the topics and the tags that indicate events to confirm the results. The higher PMI the well-combined pairs due to the probability of co-event are little less than the probability of each occurrence of words. On the contrary, if the occurrence of words is less than the co-occurrence of the words, leads to fewer PMI scores.

Table 1: Assigning label, Number of pairs, Top co-occurrences and PMI

	Most similar token (semantically)	Similarity	Label	No. of pairs	Top co-occurrence	PMI
Topic 0	('audience', show') ('audience', 'semi') ('audience', 'press')	0.266 0.285 0.266	entertaining	4363573	('semi', 'audience')	14.4397
Topic 3	('trip', 'try') ('work', 'trip') ('work', 'team')	0.556 0.556 0.485	activities	4395025	('madrid', 'cristiano')	14.37167
Topic 6	('leader', 'dear') ('need', 'grace') ('leader', 'tell')	0.705 0.727 0.666	sanctification	400888	('motors', 'chevy')	13.7866
Topic 8	('election', 'role') ('election', 'voting') ('election', 'criticism')	0.588 0.352 0.307	elections	4646299	('optimism','election')	15.8145

Secondly, the pairs are constructed due to the mutual occurrences of tokens in each tweet. Also, the PMI scores were computed for all those mutual pairs to figure out the relations between them. Moreover, a comparison was made to check which of those pair scores more. Table 1 shows the top score of PMI for each topic.
In topic 0 the first pair of tokens are ('semi', 'audience') with score 14.439 refer to the tweet ('Tensed 1st semi final for anxious audience #FRABEL #worldcup #ItsComingHome #RedTogether') that is an event of tense of the audience due to the final and semifinal match of Russia 2018. In topic3 the pair is ('madrid', 'cristiano') the tweet (Real Madrid and Juventus close to a greeing Cristiano Ronaldo deal https://t.co/QjHnxCm15Y #football #news #sport https://t.co/U856HbwaFU)which indicate an event of a deal.

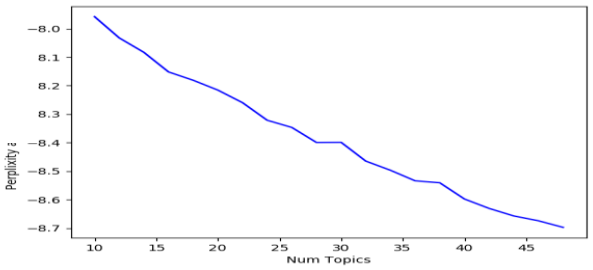

Fig. 4: Number of topics Vs. Perplexity

6.4. Discussion

To illustrate the idea, as an example Topic #0 that consists of tokens each with its probability The topic 0 has the following tokens 'audience', 'point', 'show', 'semi' , 'press' ,'result' ,'sense' ,'week' ,'something',' hand'.
Firstly, the pairs are [('audience', 'point'), ('point',' show'), ("show, 'semi'), ('semi', 'press '), ('press', 'result'), ('result', 'sense'), ('sense', 'week'), ('week', 'something'), ('something', 'hand')].
The PMI will be computed for all those pairs to figure out the relations between the tokens . To compute PMI, according to equation (2) the probability of both words are required as well as probability of each word. Counting of occurrences and co-occurrences of tokens in a corpus can be used as probabilities p(x) and p(x, y) respectively.

For topic 6 the pair is ('motors', 'chevy') and the tweet (General Motors is sending Mexican made a model of Chevy Cruze to U.S. car dealers-tax free across border. Make in U.S.A.or pay big border tax!) that refer to free of charge motors. Finally, topic 8 has the pair ('optimism', 'election') of the tweet ("Small business optimism soars after Trump election' https://t.co/WjBaTp824U') that refer to the event of an election.

6.5. Evaluation

The first evaluation was performed by perplexity showing that the lower the percentage the better the quality of implementation. See Fig 4 and Table 1.
Second, An evaluation was performed by retrieving the tweets.in two ways of retrieval and comparisons were made to prove whether it represent events or not.

1.) Retrieve the tweets for the TLDA's topics' word (tokens) that had a high probability, i.e. top 3 words for each topic.
2.) Retrieve the tweets for the pair words of higher PMI scores, which appear in it.
It was found that tweets retrieved in a second way are better because they depend on the most significant familiar appearance of names and not on distribution only.

7. Conclusion and Future Work

The twitter dataset is considered as the primary challenge of this research due to TOS. Solving this challenge by, streaming API using python libraries, that enables the researchers to construct the raw datasets. The second challenge was the topic modeling, and how to overcome the limitation of LDA when applied to twitter dataset. This limitation was solved by first streaming tweets using hashtag query and second, by applying TwitterLDA that assign each tweet to a single topic.
The overlapping in meaning was encountered. Using WordNet, extracting ordered pairs of topics' words in one document, then compute for similarities for finding the related words, then finding a general definition that refer to these associated words and by these detentions labels were created.
Furthermore, this limitation can be solved by improving the LDA by enriching the tweets using DBpedia, Wikipedia, wikidata, and Wikimedia. The drawback of LDA can be prevented using other topic modeling techniques.

References

[1] A. O. Steinskog, J. F. Therkelsen, and B. Gambäck, “Twitter Topic Modeling by Tweet Aggregation,” Proc. 21st Nord. Conf. Comput. Linguist., no. May, pp. 77–86, 2017.
[2] H. Cai, Y. Yang, X. Li, and Z. Huang, “What are Popular : Exploring Twitter Features for Event Detection , Tracking and Visualization,” MM ’15 Proc. 23rd ACM Int. Conf. Multimed., pp. 89–98, 2015.
[3] X. Zhao, J. Jiang, and W. X. Zhao, “An Empirical Comparison of Topics in Twitter and Traditional Media,” Singapore Manag. Univ. Sch. Inf. Syst. Tech. Pap. Ser., 2011.
[4] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, “Improving LDA topic models for microblogs via tweet pooling and automatic labeling,” Proc. 36th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR ’13, p. 889, 2013.
[5] D. Alvarez-Melis and M. Saveski, “Topic Modeling in Twitter: Aggregating Tweets by Conversations,” $Icwsm16, no. Icwsm, pp. 519–522, 2016.
[6] W. D. Penniman, Social Informatics, vol. 6430. 2010.
[7] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter , a Social Network or a News Media?,” Int. World Wide Web Conf. Comm., pp. 1–10, 2010.
[8] K. Sarkar and R. Law, “A Novel Approach to Document Classification using WordNet,” arXiv1510.02755 [cs], pp. 1–14, 2015.
[9] G. Ifrim, B. Shi, and I. Brigadir, “Event detection in Twitter using aggressive filtering and hierarchical tweet clustering,” CEUR Workshop Proc., vol. 1150, pp. 33–40, 2014.
[10] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, “An overview of topic modeling and its current applications in bioinformatics,” Springerplus, vol. 5, no. 1, 2016.
[11] D. A. Ostrowski, “Using latent dirichlet allocation for topic modelling in twitter,” Proc. 2015 IEEE 9th Int. Conf. Semant. Comput. IEEE ICSC 2015, pp. 493–497, 2015.
[12] X. Wan and T. Wang, “Automatic Labeling of Topic Models Using Text Summaries,” Proc. 54th Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap., pp. 2297–2305, 2016.
[13] C. C. Muşat, Ş. Trăuşan-Matu, J. Velcin, and M.-A. Rizoiu, “Automatic extraction of conceptual labels from topic models,” UPB Sci. Bull. Ser. C Electr. Eng., vol. 74, no. 2, pp. 57–68, 2012.
[14] A. Huang, R. Lehavy, A. Zang, and R. Zheng, “Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach,” Ssrn, 2014.
[15] W. X. Zhao et al., “Topical keyphrase extraction from Twitter,” Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol. 1, pp. 379–388, 2011.