Towards a more efficient and personalised advertisement content in on-line social networks *

[Natural Language Processing tesis resume] †

Patxi Galán-García
DeustoTech Computing - S3Lab, University of Deusto
Avenida de las Universidades 24
Bilbao, Spain
patxigg@deusto.es

Carlos Laorden Gómez
DeustoTech Computing - S3Lab, University of Deusto
Avenida de las Universidades 24
Bilbao, Spain
claorden@deusto.es

Pablo García Bringas
DeustoTech Computing - S3Lab, University of Deusto
Avenida de las Universidades 24
Bilbao, Spain
pablo.garcia.bringas@deusto.es

ABSTRACT

Knowing what potential clients want, is the most important issue for companies. The current situation of social communication is generating a lot of information about users, such as favourite sites, food, politic tendencies, hopes or needs. These information has an incalculable value for marketing interests but, unfortunately, it is not trivial to process it. One way to obtain this information without disturbing the users is to store their searches, used words on the Internet and clicked items, to construct specific profiles. But, the problem with these techniques is that they do not retrieve current information about the targeted user because they gather information that may or may not be up to date. In light of this background, we propose a methodology to obtain up to date information about user's interests, likes and needs by analysing users conversations in social networks and instant messaging systems to generate personalised and interesting advertisement with better impact and higher success rates.

*Resume about future disertation of Patxi Galán-García.

†A full version of this paper is available as *Towards a more efficient and personalised advertisement content in on-line social networks* Using LaTeX2$_\epsilon$ and BibTeX at `http://paginaspersonales.deusto.es/patxigg/`

1. INTRODUCTION

Nowadays, who does not have, at least, one or more profiles in one of the many existing on-line social networks? If we are looking for old or new friends, we have Facebook. If we want to publish our opinion about something, we have Twitter. If we are alone and looking for a relationship, we have eDarling. The list of topics for social networks is too long and diverse.

This situation of social communication generates a lot of information. Social network users, sometimes for ignorance or innocence, provide too much information about themselves, their favourite sites, favourite food, their politic tendencies, their hopes and needs. All this information is usually on the Internet for anyone to see. This knowledge has an incalculable value for marketing interests but, unfortunately, it is not trivial to process it.

For all companies, knowing what potential clients want, is the most important issue. But, sometimes, companies do not have the possibility to know these opinions, because there are too many clients and too many ways of communication.

For each product, advertisers need to know what the final user expects, and if it has met the expectations. This situation, sometimes results in bad marketing campaigns and lost of large amounts of money by companies due to the lack of acknowledgement of their potential clients' needs.

This type of business is oriented only to sold products. Nowadays, this model works correctly, but there is another model that focuses in the relationships with the clients that is starting to obtain good results.

At present, companies that want to compete in local and World trades, need strong relationships with their clients. This type of relations, has derived from companies that are concerned on users needs to design and produce what the users wants. This step is followed by feedback systems to obtain information about clients' experiences to improve the product or service, or to generate new products or services. Thus, the company gives a good image, the clients are happy and these clients start talking about this company in their social networks. At last, these conversations could became a trending topic on the Internet, producing high media impact

and good publicity at low cost.

With this arguments, it seems clear that the next step in marketing evolution goes hand in hand with new technologies and social networks. Moreover, in the last years, marketing has had more growth than in the last 50 years [1].

The low cost of using these technologies to spread these advertises (ads), has derived in a massive and out of control sending of these ads. This marketing model has generated a lot of discomfort among users, getting to the point of considering this type of publicity like undesirable, or spam.

This situation can be resolved with advertisement customisation systems. If users could receive ads about their hobbies, concerns, ideas or likes, the effectiveness would increase considerably. In this way, advertising campaigns that use to be considered as spam for targeting wrong users, would be received by receptive users. This paradigm offers a new, and hopeful, situation for advertisement companies, where final users will not be bothered or saturated with unnecessary and unwanted information, but with better and more interesting offers that will have more impact on them.

This marketing change is based on the adaptation of advertisement to each user, depending on preferences and likes. The problem is that, by default, most users are unwilling to give that information about themselves, so it is necessary to find other systems to obtain the information.

One way to obtain this information without disturbing the users is to store their searches, used words on the Internet and clicked items to construct specific profiles. The profiles are then used to develop specific marketing campaigns. This methods are less intrusive and unpleasant, and firms like Google or Facebook have already taken this approach and the results are obvious.

The problem with these techniques is that they do not retrieve current information about the targeted user. They are based on elements that are static and can not be updated frequently.

In light of this background, we propose a methodology to obtain the topics of users' conversations in social networks and instant messaging systems to generate more efficient, personalised and directed advertises.

The remainder of this paper is organised as follows. Section 2 reviews some natural language processing techniques, advertising in social networks and topic extraction approaches. Section 3 introduces the research idea and main objective of the future dissertation. Finally, section 5 concludes and points the next steps to be taken.

2. STATE OF THE ART

The area of this research is natural language processing or NLP. According with [2], NLP is: *a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.*

This approach is used to generate our own definition: *NLP is an area of research and application that investigates new ways to communicate between computers and humans. These methods allow to understand and manipulate natural language text or speech to use in real world.*

NLP has a lot of disciplines and applications [3]. The most known discipline is translation for being used every day to translate lots of Web pages. The most important inconvenience of this type of sub-area is that it, usually, tries to find the literal meanings of the words in each language. But, human communication does not depend only on the meanings of words. Both the sense and context, have a big importance in the communication.

The analysis of those relations, commonly refer to as topic analysis, started to appear at late 90's [4] [5] [6], but the most important grow was experienced in the beginning of the year 2000. In this period of time these techniques became a subarea of the Information Management discipline [7], [8], [9], [10], [11], [12], [13], [14].

Until 2000, the most important approaches for senses detection were based on Machine Learning (ML) and semantic analysis techniques. Since then, several NLP approaches have been used in conjunction with them to improve the results.

At this point, it is necessary to mention that, nowadays, there is no perfect automatic process to detect senses due to the many issues regarding the disambiguation problem.

Nowadays, for the Information Society (IS), NLP is one of the most important problems because the communication between humans and machines is not trivial and it is very important for technological development.

2.1 NLP stages

NLP analysis has 4 different stages, morphological analysis, syntactic analysis, semantic analysis and pragmatic analysis.

Morphological analysis is the way to determine the form, class or grammatical category of each word in the sentence. This analysis, is able to detect the relation between the minimum units that form the sentence [15]. This analysis is related with lexical analysis [16]. The words in common dictionaries have one or more lexical entries depending on how many meanings or how many different grammatical categories they have. These entries represent grammatical categories, morphological information, syntactic irregularities and meaning representation. Normally, lexical analysis only contains the root of the words with regular forms, and morphological analysis determines if the gender, number and verb are adequate with the rest of the sentence.

Syntactic analysis [17] labels each syntactic term in the sentence and analyses how words are combined to build correctly grammatical sentences. The result of this labelling is one structure that corresponds to syntactic categories for each generated lexical unit in the sentence. These grammatical units are formed by rule-sets like:

- PP = Prepositional Phrase
- NP = Noun Phrase
- VP = Verbal Phrase
- Det = Determinant

Semantic analysis [18] examines sentences and its codification to get the meaning and sense according to the context. This step of the NLP analysis, is the most difficult because the techniques to resolve it, despite they have experimented a big evolution, are not good enough. Disambiguation is the problem of this part of the analysis. One word, according with the context of the sentence, can have one or more meanings and interpretations. To solve this problem, generating one structure for each dependent and independent relation is very important because this structure will give the possible meaning according to the context. On the one hand, the meaning that does not depend on the main sentence's context refers to the real meaning of the word. This approach does not take into account the main sentence's sense or the sense acquired by the surrounding words. For this reason, this type of senses are ignored. On the other hand, the meaning that depends on the main sentence's context refers to the influence of the rest of the words and the main sentence's context in the analysed word and context. Lexical analysis provides the semantic component for each word giving particular formalization, and semantic analysis. Then processes this results to obtain a representation of the meaning of the word into the sentence.

Finally, **pragmatic analysis** [19] gives more information about the senses, in accordance with the previous analysis and the context where the terms the appear. This analysis is one of the most complex because it tries to contribute with more significant information about word senses, according to speech and participants information. Moreover, this analysis gives information about the relations between the acts that form the context and the entities within the sentence.

2.2 Advertising in Social Networks

The advertising world gives each year a lot of money to businesses. In the last 10 years, this area has grown a lot with the social networks. Social networks offer companies the option to show and retrieve information about their users and clients such as their likes, preferences, relations, moods, etc. Besides, recent studies show that Internet users spend 22% of their on-line time in social networks [20].

Depending on the targeted client what type of social network

There are several types of social networks and in order to improve the impact factor of publicity, different marketing strategies should be deployed for users that populated each type [21]. These social networks can be categorised in 4 groups [21]

1. **Pure Social Networks**
   Such as Facebook[1], VKontakte[2], etc., facilitate the communication between the members of the site, connecting them into a network. According to [22] the main demographics of Facebook are 18-45 year old males. Facebook offers extensive profiles where users can provide information about their interests, hobbies, affiliations, etc. Moreover, [23] states that marketing on Facebook must started by creating a concise and trustworthy profile. Pure Social Networks attract advertisers because of their enormous size (according to their own statistical data, Facebook encompasses over 900 million users).
2. **Grouped Social Networks**
   Such as LinkedIn[3], Ravelry[4], etc., connect users based on their present or past affiliations to a company or a professional circle. These sites can aggregate people based on their hobbies, professions and crafts. Grouped Social Networks offer one significant advantage over marketing in the other types of social networks, the advertising of the products to more uniform consumer groups sharing the same expert knowledge yields much faster results.
3. **Content Social Networks**
   Such as Flickr[5], YouTube[6], etc., offer their users the possibility to publish multimedia content in the form of photos and video clips, providing also a possibility to comment on all aspects of the content. The recently added features of subscribing to the content published by a certain provider places these sites into the same categories as blogs and social network sites. Many marketing professionals do not think about Content Social Networks as potential advertising markets. However, it is possible to build and strengthen a brand by posting a number of videos featuring the firm's name with for example how-to instructions or tips.
4. **Broadcast Social Networks**
   Such as Twitter[7] or LiveJournal[8], allow users to publish content, which may be of interests to a number of subscribers and followers of this network. In order to have two-way communication on this network the users must subscribe to each other's content. These networks may be used for broadcasting messages which do not require a direct feedback. Messages with new product features and/or updates, product offers, etc. can be 'tweeted' to a large (millions) number of users. In this way, broadcasting networks thus become a suitable replacement to mailing list.

In order to exploit the possibilities of social networks, Social Media Marketing (SMM) combines the business marketing aims on Internet with social media systems like blogs, content aggregators, content sharing sites, social networks, microblogging sites and many others.

SMM can be understood like a set of techniques and activities for firms to generate relevant and quality traffic in the

[1]http://www.facebook.com
[2]http://vk.com/
[3]http://www.linkedin.com
[4]https://www.ravelry.com
[5]http://www.flickr.com
[6]http://www.youtube.com
[7]http://www.twitter.com
[8]http://www.livejournal.com

most important social networks (such as Facebook, Twitter, Linkedin, etc.).

The firms know the potential of social networks for publicity. Due to this knowledge, some firms have tried to start new spin-offs to create new business models around the social networks [24], and many of these new companies are based on advertisement [25].

But, social networks have to be taken seriously and used with care. Sharing information can sometimes generate negative opinions and bad advertising. The important thing is not what the company says, the most important thing is what users are saying about the company, company products or the customer service team. These type of comments need to be followed and managed, because they form the company's on-line reputation. This reputation has the same impact as the one in the physical world, but is more difficult to manage because it is easy to write bad comments in social networks and share them with millions of users.

These type of situations, need a special and planned strategy. On-line reputation is closely linked to marketing in social networks [26] because to start a new marketing campaign, it is necessary to know how what is the company's reputation on the Internet and social networks, by monitoring, for example, user comments, opinions, experiences, etc.

The main objectives to generate a successful SMM campaign are: (i) increase the relevant traffic, (ii) increase the number of affiliated users, (iii) increase the number of sales, (iv) place the brand in specific areas of the Media Society through social networks and (v) increase the contributions to users. This type of marketing, requires an inversion and, if the firm can drive correctly drive its social reputation and users comments, the result of campaigns will be successful, with a good return of the inversion.

Some successful examples of Social Marketing, using social networks, are: *Pastillas contra el dolor ajeno*[9] in Spain, *Mercado Libre*[10] in South America, *Ikea*[11][12] in Denmark and Spain and *Dell*[13] around the World. But not all is success. There are some cases that are just the opposite, sometimes with damaging results to the reputation, like: *Dominos Pizza*[14] and *ComCast*[15], both in USA.

It seems clear that many aspects exist that can improve the impact factor of the publicity. One of them is the influence that some members have in some communities (positive externalities) but the most effective one is the Word-of-Mouth [27], which quickly spreads and reaches the audience with good perspectives. In this way, it is important to target the adequate audience, but it is maybe more important to adapt the content to users' context, interests, moods or seasons.

[9]http://www.msf.es/pastillascontraeldolorajeno/
[10]http://www.youtube.com/watch?v=NeBhs47UBJQ
[11]http://www.youtube.com/watch?v=YMJD53fxihU
[12]http://www.youtube.com/user/IKEAESIberica
[13]http://www.dell.com/
[14]http://www.youtube.com/watch?v=xaNuE3DsJHM
[15]http://www.youtube.com/watch?v=CvVp7b5gzqU

2.3 Topic Extractions

There are several approaches to build topic detection systems for text documents, but all of these methods employ the co-occurrence of the text within the documents. Some of these approaches are:

- Latent Semantic Analysis (LSA) or Latent Semantic Indexing (LSI)
- probabilistic Latent Semantic Analysis (pLSA) or probabilistic Latent Semantic Indexing (pLSI)
- Latent Dirichlet Allocation (LDA)

These methods use the situation of some terms within the documents to give important information about the relation between one text and the document. The difference between these methods is how they use probabilistic assumptions models. LSA represents each document as a linear projection taking into account the frecuency of its terms [28]. pLSA, pLSI [29] and LDA [30] are models of generative probabilistic mixture that consider each document as a combination of topics. LSA and LSI checks the vectorial representations of documents (normally only in frequency terms). These methods use Singular Value Decomposition (SVD) model because is a typical method to create the documents mapping. Using this model, the information retrieval (IR) task has improved. This is because this model assumes the idea of 'the documents that shared co-occurrence terms or words, should have the same representation although does not share the exactly words to build the sentence' [29]. LSA has the same idea, but performs the process reducing the noise [29]. LSA have been a great help for detect synonymy between words that belongs the same topic, and many of his applications have resulted in improvement of terms processing [29].

LDA, pLSA and pLSI are generative mixture models. These models assumes that the outs (documents in this case) are produced according to probabilistic rule-sets. In these models the topic is a words distribution, where the words appear simultaneously with higher probability when the interlocutor is speaking about the topic. LDA, pLSA and pLSI assumes that each document has formed by more than one topic. These models share two multinomial distributions:

- document-topic distribution
- topic-word distribution

These type of models assume that the documents are built by random words about some topics. For each word, the method choose one topic according to document-topic distribution and, after, choose a random word of topic-word distribution. These models use the idea of bag-of-words that ignores the sequence of words in the sentence and document [30]. Almost all of the literature of this area is focused on the relations between these models and the *corporas'* semantic and how effectively and efficiently is to build these models using statistical inferences. For example, with pLSA,

Hoffman used one modification of Expectation Maximization (EM) called tempered EM (TEM) to create estimations for document-topic and topic-word multinomial distributions. Other example is with LDA taking a Bayesian approach and Dirichlet method [29]. This approach puts a previous conjugation before each document-topic multinomial distribution to use estimation methods like variation words inference [30] and Gibbs sampling [31].

3. TOWARDS A MORE EFFICIENT ADVERTISEMENT CONTENT

The challenges that advertisement systems will face in the future mark the boundaries where the scientific community will have to focus its efforts. Therefore, bearing in mind this situation we have established the fundamental hypothesis of the future dissertation:

It is possible to generate a methodology capable of improving the detection of topics in short texts and conversations to improve targeted marketing in instant messaging and social network systems.

This hypothesis aims to show the possibility of adapting marketing strategies to maximise the impact of advertising campaigns. Besides, we think that users could be favoured with this new approach by reducing the amount of unsolicited offers by receiving ads of real interest.

Taking as a reference the fundamental hypothesis we have established the following specific goals:

- Processing conversations. Retrieve conversations from social networks and instant messaging (IM) systems, process them, and obtain the morphological, syntactic, semantic and pragmatic analysis.
- Obtaining topics from conversations. Infer users' interests, moods and needs by analysing the dialogues extracted from the previously processed conversations.
- Generating personalised advertisement campaigns. Use the topics extracted from the analysed conversations to offer the users more interesting offers.
- Sharing the generating campaigns on the Internet. Use the tools provided by social networks to spread the designed campaigns.

Figure 1 shows a graphic scheme representing the flow that the final system will follow, using as a reference the main objectives of the future dissertation.

Next we present in more detail each of the objectives.

3.1 Processing conversations

This research aims to improve NLP techniques for short texts analysis, like conversations in IM or chats, providing new tools to obtain the structure and meaning of sentences in conversations that have a high degree of ambiguity, polarity, irony, syntax or grammatical mistakes. To accomplish this objective we need to process and filter a lot of sources to

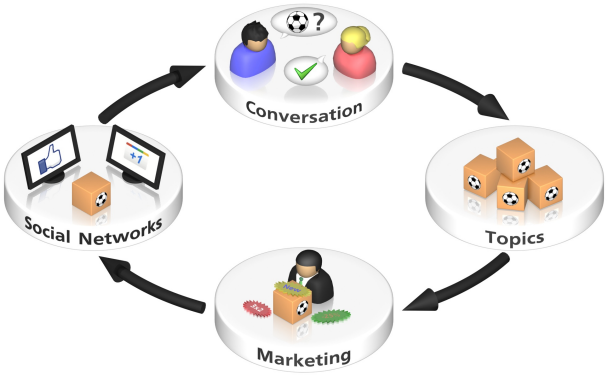

Figure 1: Customised advertisement in on-line social networks by analysing users' conversations.

obtain information, but not all of this information is structured and correctly written. IM systems generate a lot of information, but there are some problems:

- Colloquial language. Altered languages that are adopted by common use.
- Abbreviations. Use of contractions or abbreviated words.
- Misspelling. Terms representing no correctly spelled word of the same language at all.
- More than one interlocutor. It is important to detect all the users that take part in the conversation.
- Split text. Sentences can be fragmented in several subsentences (sentences fragmented by the introduction of new lines or the interaction of another user).

Taking into account these problems, processing IM conversations is more difficult and tedious, and requires a more complex process, but the effort is really worthy. Nevertheless, in this type of communication channel, users say what they really want, so this source gives important and valuable information about likes and needs.

3.2 Obtain topics from conversations

After processing the conversations, we want to extract the topics from them in order to learn users' interests, needs and moods. To achieve this goal, we will design a methodology to analyse the conversations taking into account some natural language problems, such as term ambiguity, and extract the knowledge we are looking for.

Our methodology will be formed by two main approaches. The former will receive as input previously labelled conversations to train ML algorithms that will build the classifiers able to then classify new conversations. Because our interest lies in finding what users want in real time, those labelled conversations must reflect the changes of topics that conversations usually suffer. In this way, each labelled instance could/should include several topics depending on what the users are talking about through the conversation. After

we acquire our knowledge base, we will make use of commonly used ML algorithms to create our models, or classifiers. Once we have the classifiers, we will then classify new unlabelled conversations, taking into account that this classification has to be able to show the changes of topics through the dialogue, to validate the accuracy of our method. Summarising, we have a document classification problem, in which, instead of documents, we are classifying pieces of conversations. As this classification will analyse previously acquired, and complete, conversations, we will refer to this approach as our off-line method.

The second approach will make use of the previously acquired knowledge base only this time, instead of classifying complete conversations, it will be applied to extract the topics from sentences, or group of sentences, in real time. The idea is to adapt our method to social network chats or IM systems, to analyse conversations on the fly. In this way, as this approach will try to infer users' interests as they talk, we will refer to this approach as our on-line method.

Thus, if we apply our off-line method to generate marketing campaigns, we would be getting a mid-term advertising programme, after recollecting conversations from users and then providing the offers that where presumably desired some time ago. This type of advertising is the one that seems to be taking form nowadays. On the other hand, if we apply our on-line method to generate offers for the users, we would be using the most current information about their interests and needs, improving the impact of the advertisement and improving the predisposition of the user.

3.3 Generating personalised advertising campaigns and sharing them on the Internet

Finally, the last step for this research work will be to generate a methodology that makes use of the previously extracted topics in order to generate advertising campaigns following the interests and needs of the users. The objective is to obtain a road map for marketing advertisers to improve the outcome of their programmed campaigns.

4. FIRST STEPS IN RESEARCH

As stated in the previous section, the future dissertation will tackle the problems of conversation processing, topic analysis, directed campaigns generation and sharing of those campaigns. Next, we introduce the first steps taken in this research work.

4.1 Conversation processing

In this step, we analyse the structure and codification of the conversations.

To solve the text structure problem, we have developed a methodology to organise, correct misspelling words and translate from Short Message Service (SMS) language. The system recovers the input data, from now on "Conversational Unit" (CU), to analyse it. This step of recovery has two parts. The first part is that the person can write one sentence in one or more lines, like Figure 2. The second part is that the person can write one sentence or question, and then, wait before write another question or sentence.

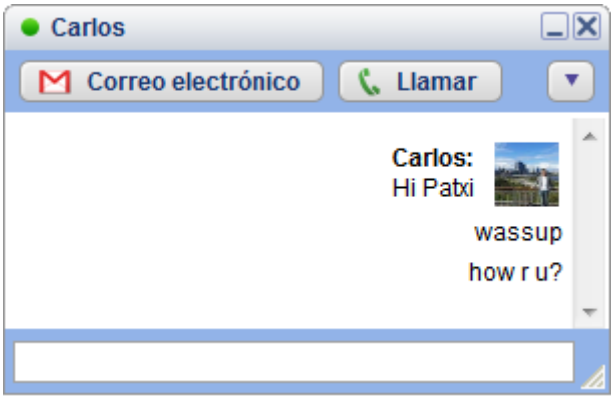

Figure 2: Chat conversation

This system of data recovery starts working when the person stops writing. It has a time counter to wait. When the counter is finished, if the person does not write again, the CU is saved and the analysis system starts to working.

When we have the CU, we apply some methods for clean it, like remove repeat words and letters and remove special characters. After this clean, the system replaces the "emoticons", the misspelled words will pass by the Levenshtein [32] distance for revise it, and if it is a mistake or a term of SMS language, is corrected. The SMS sentences in English are like *"Hi Patxi, wassup, how r u?"* This sentence well written in English is *"Hi Patxi, What's up, How are you?"*.

Moreover, not all terms are simple nouns, few terms are personal nouns, and the system has methods to recognize it. For this step, the system uses `www.planetamama.com.ar` personal nouns database. This website has more than 10 years running, its personal nouns database has more than 3.000 values and it has a big collaborator community.

With this methodology we have achieved good results on the analysis of disorganized texts [33].

4.2 Topic analysis

It must be noted that in our research, the validation is quite difficult because, sometimes, the topics can be interpreted with different meanings. In addition, we need a lot of labelled conversations to validate results. In this way, we are currently labelling conversations with different formats like AVE [34].

For the first experiments, we have obtained some information from specific and significant articles of the Spanish Wikipedia[16] and from the Royal Spanish Academy (RAE[17]).

To manage the different types of syntactic analysis, we have used the OpenSource tool *FreeLing* [35]. This tool allows us to obtain the different meanings, senses, values and genres of the words in the sentence. Also it can be configured for a lot of languages.

[16]http://es.wikipedia.org
[17]http://www.rae.es

Finally, we have used the indexing tool *lucene* [36] to store the obtained knowledge, divided by the sentences' topic and context.

4.2.1 Methodology to obtain the knowledge

The sources to obtain the necessary knowledge have been Wikipedia, because this Web site has a huge amount of articles, periodical updates and colloquial expressions and the RAE, because this site has the accepted definitions for words in the Spanish language.

For us, a valid knowledge about one topic is information or articles about this topic cleaned of special characters like icons, references, symbols, etc, that is, we only want text.

After obtaining this knowledge, we then used FreeLing to analyse all the knowledge to retrieve and store the syntactic information about the sentences. Articles and information are divided in sentences that are then analysed syntactically and semantically. From these sentences, we remove *stopwords* [37], apply an *stemming* [38] process to obtain the root of the terms, obtain the sentence's sense and context, and keep all this information in the knowledge system. Finally, we use all this information to extract the sentences' topic.

4.2.2 Managing syntactic aspects of the sentence

To make syntactic and disambiguation analysis of sentences, we use the FreeLing tool. Our system stores the sentences' words in *'syntactic containers'*. This container has all syntactical and semantical information about its principal element concerning to the context of the analysed sentence. Disambiguating the word and obtaining the sense have an important value to retrieve the topic of the sentence. If it is possible to obtain a sense, the *'syntactic containers'* have a specific field to store it.

4.2.3 Methodology to retrieve the sentence topic

Normally, every sentence has a topic. Sometimes, this topic is related with sentences that appear previously in the text and, other times, the analysed sentence has its own and independent topic. We have tried to obtain the sentences' topics using the tool *lucene*, by using two different knowledge bases, or lucene indexes. The first index contains information about topics and the second one has information about word senses. We decided to separate the information about senses and topics to give more importance to topics because we are looking for users' interests and needs, which can be inferred from topics. Hence, senses are only used to help in the topic choice.

Once we have all the conversations stored in lucene, when we want to analyse a new sentence, we query the system and get as a result a list of topics. Each of these topics are representing different sentences which are similar to the one that is being analysed. When the resultant sentences surpass a certain similarity threshold, we can imply that they are similar to one under study and that they share the same topic.

4.3 Future experiments

Although these experiments have shown an improvement detecting topics in short texts, the knowledge is too small and we only classify under a few general topics. Now, we are experimenting with other methods to improve the results. For example, we are indexing the information and relations found in DBpedia[18] to increase our knowledge base and, we are manually labelling many conversations to validate the results.

Besides, as a first approximation, we have developed a chatterbot, called *Negobot* [33], to infer if the subject involved in a conversation has paedophile tendencies. This prototype uses our topic extraction system and domain specific knowledge to detect these type of behaviours in social networks and IM systems.

5. CONCLUSIONS

Digital live is becoming an important part for most of the World inhabitants. This situation generates a lot of information about likes, preferences or points of view from users. Some firms try to use this information to improve their situation in the market by reaching a wider audience with a more interesting message, but they do not always achieve their goal.

Currently, advertisement systems do not differ between users that want publicity and users that do not want it. This situation generates a lot of unwanted ads and, in fact, bad image for the publicised company.

In light of this background, the main aim of this research is to study the context and sense of conversations on chats or IM services like Facebook, Tuenti, Google+ or WhatsApp to propose a new paradigm of targeted advertisement.

Our system will try to resolve this situation by analysing users' conversations in social networks and IM services. To achieve this goal, we propose two approaches, an off-line method able to generate better marketing campaigns in mid term, and an on-line system able to offer, in real time, what the user wants.

Summarising, this research work, and future dissertation, is directed towards improving current marketing campaigns by adding to the equation update information about users' interests and needs. This new and fresh information should increase the impact of the offered advertisement and improve the success of the message, leading to viral marketing, Word of Mouth and, hence, better outcomes.

6. REFERENCES

[1] S. Zyman and A. Brott, *The end of advertising as we know it.* Wiley, 2002.

[2] E. Liddy, "Natural language processing," 2001.

[3] G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, no. 1, pp. 51–89, 2003.

[4] S. Argamon, M. Koppel, and G. Avneri, "Routing documents according to style," 1998.

[5] E. Spertus, "Smokey: Automatic recognition of hostile messages," 1997, p. 1058?1065.

[6] B. Kessler, G. Numberg, and H. Schütze, "Automatic detection of text genre," 1997, p. 32?38.

[18] http://dbpedia.org

[7] N. Kobayashi, T. Inui, and K. Inui, "Dictionary-Based acquisition of the lexical knowledge for p/n analysis." *SIG SLUD*, vol. 33, p. 45?50, 2001.

[8] A. Rauber and A. Müller-Kögler, "Integrating automatic genre analysis into digital libraries," 2001, p. 1?10.

[9] P. Subasic and A. Huettner, "Affect analysis of text using fuzzy semantic typing," *Fuzzy Systems, IEEE Transactions on*, vol. 9, no. 4, p. 483?496, 2001.

[10] M. Dimitrova, A. Finn, N. Kushmerick, and B. Smyth, "Web genre visualization," 2002.

[11] S. D. Durbin, J. N. Richter, and D. Warner, "A system for affective rating of texts," *KDD Wksp. on Operational Text Classification Systems (OTC-3)*, 2003.

[12] H. Liu, H. Lieberman, and T. Selker, "A model of textual affect sensing using real-world knowledge," 2003, p. 125?132.

[13] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," 2003, p. 105?112.

[14] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," 2003, p. 34?36.

[15] J. M. Anglin and G. A. Miller, *Vocabulary development: A morphological analysis.* Wiley-Blackwell, 2000.

[16] T. T. Ballmer and W. Brennenstuhl, *Speech act classification: A study of the lexical analysis of English speech activity verbs.* Springer-Verlag, 1980.

[17] G. Salton and M. Smith, "On the application of syntactic methodologies in automatic text analysis," vol. 23, 1989, p. 137?150.

[18] C. Goddard, *Semantic analysis.* Oxford University Press, 1998.

[19] J. Wilson, *Politically speaking: The pragmatic analysis of political language.* Basil Blackwell, 1990.

[20] "How the world spends its time online -," http://visualeconomics.creditloan.com/how-the-world-spends-its-time-online_2010-06-16/. [Online]. Available: http://visualeconomics.creditloan.com/how-the-world-spends-its-time-online_2010-06-16/

[21] I. Pustylnick, "Advertising in social networks," 2011.

[22] B. Chappell, "Social network analysis report-geographic- demographic and traffic data revealed," *Ignite Social Media Marketing. Retrieved from Ignite on October*, vol. 1, p. 2009, 2009.

[23] S. Holzner, *Facebook marketing: leverage social media to grow your business.* Que Pub, 2009.

[24] P. Sääskilahti, "Monopoly pricing of social goods," 2007.

[25] "TiVo's stealth giveaway - slate magazine." [Online]. Available: http://www.slate.com/articles/business/moneybox/2000/10/tivos_stealth_giveaway.html

[26] C. Fombrun, "Indices of corporate reputation: An analysis of media rankings and social monitors' ratings," *Corporate reputation review*, vol. 1, no. 4, pp. 327–340, 1998.

[27] J. Hartline, V. Mirrokni, and M. Sundararajan, "Optimal marketing strategies over social networks," in *Proceedings of the 17th international conference on World Wide Web.* ACM, 2008, pp. 189–198.

[28] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[29] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 1999, pp. 50–57.

[30] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[31] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of latent semantic analysis*, vol. 427, no. 7, pp. 424–440, 2007.

[32] T. Okuda, E. Tanaka, and T. Kasai, "A method for the correction of garbled words based on the Levenshtein metric," *Computers, IEEE Transactions on*, vol. 100, no. 2, pp. 172–178, 1976.

[33] C. Laorden, P. Galán-García, I. Santos, B. Sanz, J. Gomez-Hidalgo, and P. Bringas, "Negobot: A conversational agent based on game theory for the detection of paedophile behaviour," in *Proceedings of the 5th International Conference on Computational Intelligence in Security for Information Systems (CISIS)*, 2012, in press.

[34] A. Penas, Á. Rodrigo, V. Sama, and F. Verdejo, "Overview of the answer validation exercise 2006," *Evaluation of Multilingual and Multi-modal Information Retrieval*, pp. 257–264, 2007.

[35] X. Carreras, I. Chao, L. Padró, and M. Padró, "Freeling: An open-source suite of language analyzers," in *Proceedings of the 4th LREC*, vol. 4, 2004.

[36] E. Hatcher, O. Gospodnetic, and M. McCandless, "Lucene in action," 2004.

[37] W. Wilbur and K. Sirotkin, "The automatic identification of stop words," *Journal of information science*, vol. 18, no. 1, pp. 45–55, 1992.

[38] J. Lovins and M. I. O. T. C. E. S. LAB., *Development of a stemming algorithm.* MIT Information Processing Group, Electronic Systems Laboratory, 1968.