Manuel Rebol

Automatic Classification of Business Intent on Social Platforms

Bachelor's Thesis

Graz University of Technology

Institute for Interactive Systems and Data Science
Head: Univ.-Prof. Dipl-Inf. Dr. Stefanie Lindstaedt

Supervisor: Dipl.-Ing. Dr.techn. Roman Kern

Graz, Mai 2017

This document is set in Palatino, compiled with pdfLaTeX2e and Biber.

The LaTeX template from Karl Voit is based on KOMA script and can be found online: https://github.com/novoid/LaTeX-KOMA-template

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, ______________________________ ____________________________________

Date Signature

Eidesstattliche Erklärung[1]

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am ______________________________ ____________________________________

Datum Unterschrift

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Abstract

People spend hours on social media and similar web platforms each day. They express a lot of their feelings and desires in the texts which they post online. Data analysts always try to find clever ways to get use of this information.

The aim of this thesis is to first detect business intent in the different types of information users post on the internet. In a second step, the identified business intent is grouped into the two classes: buyers and sellers. This supports the idea of linking the two groups.

Machine learning algorithms are used for classification. All the necessary data, which is needed to train the classifiers is retrieved and preprocessed using a Python tool which was developed. The data was taken from the web platforms Twitter and HolidayCheck.

Results show that classification works accurately when focusing on a specific platform and domain. On Twitter 96 % of test data is classified correctly whereas on HolidayCheck the degree of accuracy reaches 67 %. When considering cross-platform multiclass classification, the scores drop to 50 %. Although individual scores increase up to 95 % when performing binary classification, the findings suggest that features need to be improved further in order to achieve acceptable accuracy for cross-platform multiclass classification.

The challenge for future work is to fully link buyers and sellers automatically. This would create business opportunities without the need for parties to know about each other beforehand.

Contents

Abstract v

1 Introduction 1

2 Background 5

2.1 Business intent detection via keyword selection 5

2.2 Business intent is often linked to web spam 6

2.3 Business intent based on search queries 6

2.4 Business intent has two main phases 7

2.5 Additional ideas and conclusion on previous work 7

3 Methods 9

3.1 Concepts 9

3.2 Implementation 11

3.2.1 Phase 1: Data retrieval 12

3.2.2 Phase 2: Data preparation 12

3.2.3 Phase 3: Machine Learning 14

4 Results 17

4.1 Feature selection 17

4.2 Classification results 17

5 Discussion 21

5.1 Lessons learned 21

5.2 Limitations 22

6 Conclusion 25

7 Future Work 27

Bibliography 29

List of Figures

1.1 Classification into three categories 2
1.2 Linking buyer and seller 3

3.1 The Knowledge Discovery Pipeline 10
3.2 The classification pipeline 11
3.3 Data Set: Number of instances in each class 13
3.4 Feature selection: Number of links 14
3.5 Feature selection: Posts per day 15

List of Tables

2.1 Keyword occurance in Tweets with business intent 6

4.1 Data Set: Importance of features 18
4.2 Machine Learing: Comparing the classfication results within a platform 18
4.3 Machine Learing: Comparing the classfication results throughout different platforms 20

1 Introduction

The aim of this bachelor thesis is to **detect business intent on different web platforms**. The detection has to be done domain independent which allows to apply the results in several business sectors. Additionally, the algorithms should be able to process information from any platform on which users post information. These two requirements make cross-platform recognition of business intent possible.

First of all, the term *Business Intent* needs to be specified further. In this work business intent refers to the desire of a person or company to buy or sell products/services. It could not only be a single, but also multiple products/services. This definition allows the classification of any information created by an user of a web platform into three main categories visualized in Figure 1.1. Firstly, the category *Buyer* refers to content which states that someone is interested in acquiring a product/service. Secondly, the category *Seller* represents the information in which something is offered or sold. Content in these two categories is referred to as business intent. Thirdly, all other information can be classified as having *No Business Intent*. When considering general platforms such as social media services, the third category usually is the largest.

One application of classification is the possibility of linking buyers and sellers. Businesses between two parties could be made throughout different platforms without knowing about content of each other party beforehand. Figure 1.2 shows how the linking could be done. An advantage of this approach is that no active search needs to be performed. For example, a Twitter user which complains about his sound system could be linked to a user which recommends sound systems on a technical forum. Of course

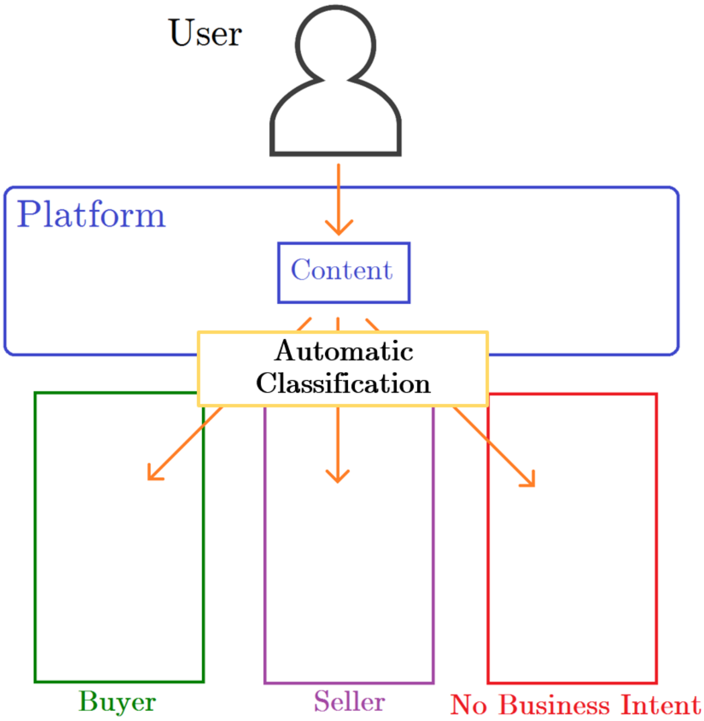


Figure 1.1: Users post content on a certain platform. This content can be classified into three different categories: Buyer(green), Seller(purple) and No Business Intent(red).

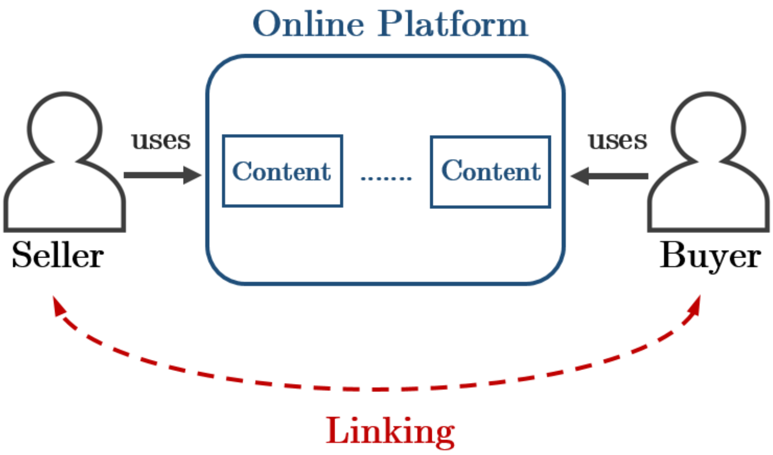


Figure 1.2: Buyer and seller use an online platform without knowing each other. With classification of the content which both user have created, it would be possible to link them. The classification is done by machine learning algorithms. As a consequence, neither buyer nor seller needs to perform search or any other activity to find each other.

there exist many other use cases in which the results could be applied.

The automatic classification is done through machine learning. First, data needs to be collected from different platforms and domains. Second, an algorithm needs to learn about the problem using a training dataset. Afterwards the classification needs to be performed on an independent testing dataset. Finally, results from various algorithms are evaluated and compared.

Previous work, especially Hollerit et al. [1], has focused on detecting business intent on online platforms. However it has not yet been discovered if it is possible to identity business intent throughout several platforms. This is the goal of this thesis. Furthermore, this thesis focuses on classification in the German language which is not commonly discovered in this context.

The thesis consists of the following parts. After the introduction, information about existing work in this research topic is given. Then, the concepts of this work are explained followed by some details about the implementation. Results of the methods used are evaluated and afterwards discussed. At the end the final conclusion is given and ideas about future work are suggested. In the next chapter related work is summarized.

2 Background

In this chapter related work on similar research topics will be discussed. The paper on which this thesis is built up on will be mentioned first, followed by ideas of other work which was used to create this thesis.

2.1 Business intent detection via keyword selection

The paper *Towards Linking Buyers and Sellers: Detecting Commercial Intent on Twitter* [1] provides information about how business intent can be gathered from micro-blogging platforms like Twitter. A textual approach is presented on which posts on Twitter (”Tweets”) can be classified into ones with and without business intent. Additional investigations also distinguish between buying and selling intent.

Results show that a few keywords are sufficient to identify most of the content containing business intent. The keywords *buy* and *cheap* already account for approximately 60 percent of all Tweets expressing business intent. The exact numbers can be seen in Table 2.1.

The conclusion drawn is that when performing search on the right keywords most of the relevant information can be found. Although the results are found in language English, it could be supposed that this behavior is also similar in other languages. The work on this thesis has shown that this is the case also in the German language.

Keyword	Number of Tweets with		
	Business Intent	Buying Intent	Selling Intent
cheap	38	11	27
buy	34	31	3
sell	16	2	14
purchase	13	12	1
bidding	11	8	3
auction	4	2	2
find	2	2	0
retail	2	1	1

Table 2.1: 120 Tweets with business intent were annotated. For each keyword contained in the Tweets a corresponding row exists in the table. Each row indicates the number of Tweets a certain keyword is used in. The second column states the sum of all occurrences with business intent. The third and fourth columns focuses on buying and selling intent separately.

2.2 Business intent is often linked to web spam

In some cases it is quite hard to distinguish between business intent and web spam. The reason for this phenomena is that both categories have quite similar characteristics. However, Benczúr et al. [2] showed that when selecting the right features differentiation is possible and even spam detection algorithms can be improved. Nevertheless, it has to be taken into account that there exists always spam on web platforms and it has quite similar characteristics as business intent. Therefore, when linking possible buyers and sellers security concerns regarding spam have to be taken into account.

2.3 Business intent based on search queries

Another interesting approach of analyzing business intent is presented by Ashkan and Clarke [3] and Ashkan et al. [4]. In their work, the detection of business intent is based on the ad click behavior after a search query was performed. Results showed that users who search for terms like *phone*, *car*, *hotels*, *airlines* and *games* usually have business intent, whereas the terms *news*, *mail* and *school* do typically not indicate business activity. Taking this into account, the selection of possible business domains for this thesis becomes easier. Tourism and car branch are businesses in which buying and selling intention might be high because of the query topics car, hotels and airlines.

2.4 Business intent has two main phases

A very interesting research paper is written by Zhao et al. [5]. This paper mentions different phases in which business activities are performed by users. The two main ones are *Research* and *Commit*. On the one hand, the phases do not play a big rule in detecting business intent, because it can be done regardless of the situation the user is in. On the other hand, when linking buyers and sellers this becomes important, as if a buyer has already committed a purchase, it will not be interested in buying anymore.

2.5 Additional ideas and conclusion on previous work

Another idea which was introduced by Zhao et al. [6] is to detect business intent based on demographic information. Unfortunately, it is very difficult to get the required information on the different web platforms. One more interesting paper [7] found that it is also possible to detect trends using business intent detection. Therefore, it could be possible to link new business. In addition, a review platform might also be used in order to find users with business intentions. Such an approach is described by Kasper and Vela [8]. Previous work by Want et al. [9] has also shown that it is possible to classify all the business relevant Tweets in six categories.

To summarize, it should be said that there already has been done a lot of research in this field. There exists at least one or two observations in every paper which were essential to create this work. However, one of the things which was not investigated previously is the topic business intent in context of the German language. It is quite hard to find information in this area. Nevertheless, it is always important to consider previous work to avoid making mistakes and use efficient methods. The exact methods which were used to create this thesis will be shown in the next chapter.

3 Methods

In the first part of this chapter the general concepts behind this work is described focusing on the most important facts. In the second part, these concepts are explained more in detail.

The most abstract view splits the work into three main stages. First, data is retrieved from a platform. Second, important features are extracted and the data is labeled. Third, the labeled data is put into a machine learning algorithm to make future estimations.

However, this processes will be now explained in a more detailed view.

3.1 Concepts

In this section the underlying concepts of this thesis are explained. First, the selection of the platform form which the data is take needs to be defined. Then, data needs to be prepared in the knowledge discovery process. Finally, the best algorithm for machine learning needs to be selected.

At the begin, the *Knowledge Discovery Pipeline* (Figure 3.1) based on [10] can be used to explain the first steps. Data of selected platforms in gathered. In order to have approximately the same amount of instances with and without business intent preselection is needed. The concept used for preselection is keyword search.

Then, data which was created within a time horizon of greater than one year is taken in order get time independent results. Each instance of data is labeled using the three relevant classes: Buyer, Seller, No Business Intent. Features are chosen from the attributes. Only attributes which can be

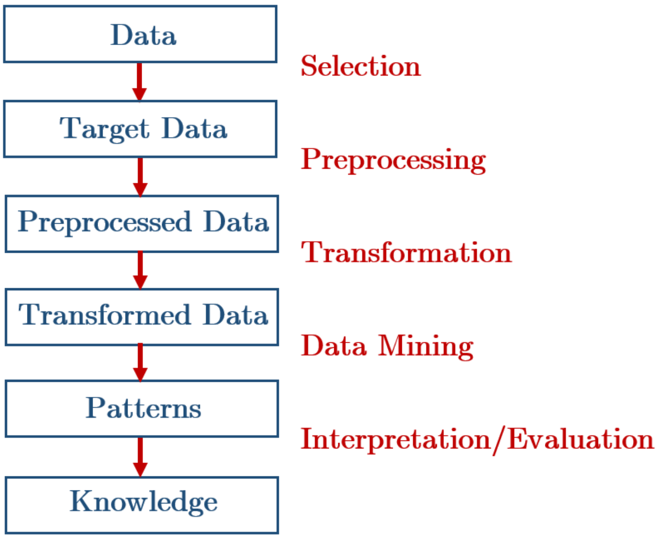


Figure 3.1: The Knowledge Discovery Pipeline [10]: In the first step specific data is selected resulting in target data. After preprocessing, the output is called preprocessed data. The transformations are applied generating transformed data. Later data mining is done which produces patterns. At the end these patterns are interpreted and evaluated resulting in knowledge.

obtained form every platform and text based attributes are selected to be features. This supports the idea of platform independent classification.

After the generation of datasets for different platforms machine learning algorithms are applied. The results of these algorithms are used to decide which features should be used. Additionally, it can also be determined which type of machine learning algorithm work best. Finally and most importantly, the results of machine learning show if it is possible to link buyers and sellers from different platforms which have business intent.

In contrast to this section which focuses on the underlying concepts, the section below explains the technologies used for implementation.

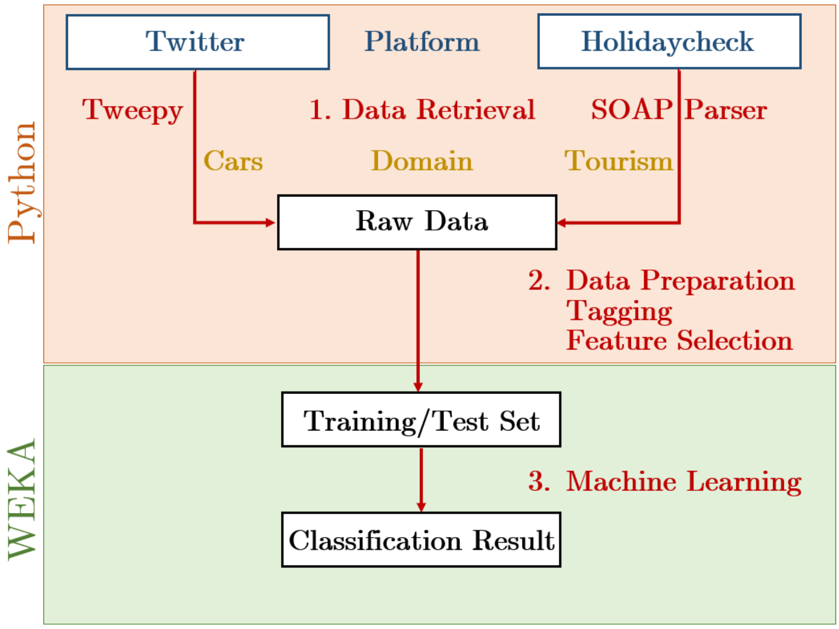


Figure 3.2: The platforms Twitter and Holidaycheck are used to collect data from. In the first step data is retrieved in Python using Tweepy and Beautiful SOAP Parser. Data of the car business and tourism domain are obtained via keyword search. In the second step the raw data is processed further, annotated and features are selected. After this step training and test sets are created. Finally, the tool Weka is used to apply different classification algorithms on the data. After evaluation final results are produced.

3.2 Implementation

The whole implementation can be separated into three parts which are shown in Figure 3.2. The first two parts are developed in the programming language Python. The reason why Python was selected is that it provides great plugins for machine learning operations. In the third part the machine learning toolkit Weka is used to run classification.

3.2.1 Phase 1: Data retrieval

In the first step information from different domains and platforms is collected. The two platforms considered are:

- Twitter
- HolidayCheck

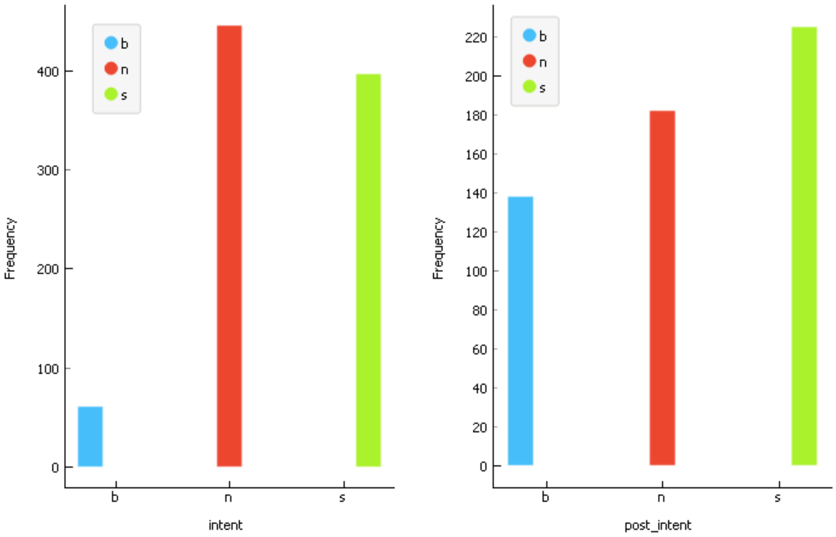
On the platform Twitter, the Python plugin Tweepy is used to connect with the Twitter Application Programming Interface (API) using an app connection token which was created beforehand. Keyword search is applied to obtain Tweets from the car domain. In order to get Tweets with buying intent the keywords *Suche*, *Kaufe* and *Ankauf* are used in combination with all the different car brands. Similarly, keywords *Verkaufe* and *Biete* were used to find Tweets with selling intent. Because Twitter API only crawls Tweets via search from the last few days, manual search is used too in order to get Tweets from a period up to three years. The raw data containing as much attributes as possible is stored in a Comma Separated Value (CSV) file.

HolidayCheck.at forum represents the second platform. In this case the Python BeatifulSoap Parser plugin downloads the data, because the API is only provided for commercial users. Similar to how it is done on the first platform, keyword search is performed to enforce the collection of business intent. However, only the keyword *Suche* is used, because no domain-specific keywords are needed. The time horizon of collected data from the HolidayCheck.at forum is about ten years. The data again is stored in a CSV file which is processed further in the second step.

3.2.2 Phase 2: Data preparation

The raw data from phase one is filtered to remove Tweets and forum posts containing bad language. Furthermore, data and time information is unified and text is transformed into a Weka readable format. Therefore, special characters are deleted and Hypertext Markup Language (HTML) tags are excluded.

After all the attributes which can be directly gathered from the platform are in the dataset, the additional attribute *intent* is added manually which



(a)Twitter Classes (b)HolidayCheck Classes

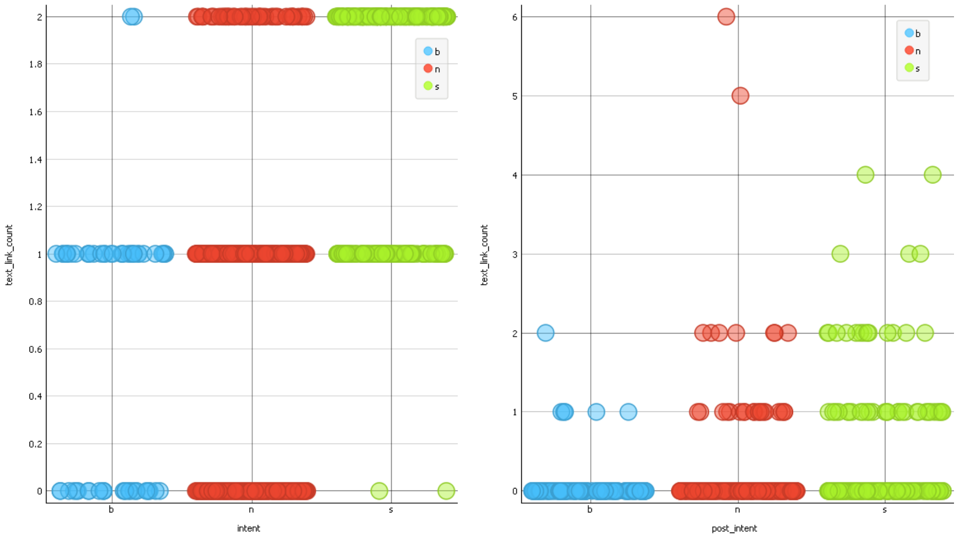
Figure 3.3: Each bar represents one class: **B**uyer (blue), **N**o business intent (red), **S**eller (green). The height of each bar indicates the number of instances in each class.

indicates whether a data instance has buying, selling or no business intent. The results of tagging the Twitter and the HolidayCheck dataset is shown in Figures 3.3a and 3.3b.

After all instances have been tagged, features need to be selected in order to get best results in machine learning phase. The first feature introduced is word frequency. This feature is generated by applying the StringToWordVector filter on the transformed text attribute using the Weka tool.

The second feature represents the number of links in the text attribute. Tweets contain up to two links (Figure 3.4a), whereas HolidayCheck forum posts have up to six links (Figure 3.4b). In general it could be said that content with selling intent has a higher number of links. Posts with no business intent lie somewhere in the middle.

The third feature reflects the activity level of the user which created the content. This feature shows that users on Twitter in general post more frequently (Figure 3.5a) than users on tourism forums (Figure 3.5b). In



(a)Twitter Link Count

(b)HolidayCheck Link Count

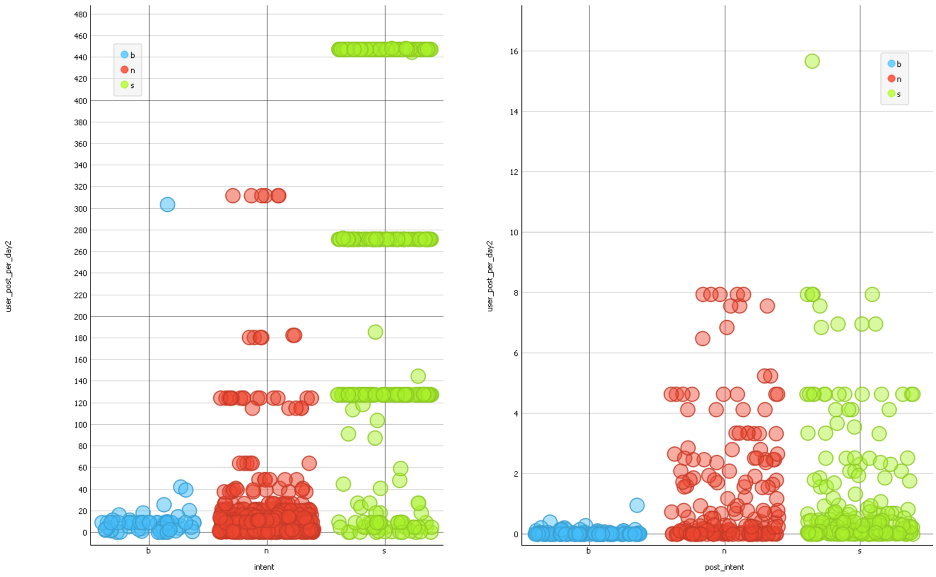
Figure 3.4: The three different classes on the x-axis: **B**uy (blue), **N**o Business Intent (red) and **S**ell (green). The y-axis represents the number of links. Each data point refers to one instance in the dataset.

order to make the features comparable throughout different platforms, their values have to be normalized.

After feature selection, the data is prepared to be accepted as input for the machine learning toolkit Weka. The data types are matched and an arff-file which is used in phase 3 is created.

3.2.3 Phase 3: Machine Learning

Finally, in phase 3 the input file created in phase 2 is selected in Weka. Several classification algorithms are tested using training and test data sets. Classification is applied on the generated data sets which include data instances from different domains and platforms. The algorithms applied perform binary classification if only two output classes are present. Whenever classification is performed on all three classes then the algorithms apply multiclass classification. Results of how well the algorithms perform



(a)Twitter Post Frequency (b)HolidayCheck Post Frequency

Figure 3.5: Every instance of the three classes **B**uy (blue), **N**o Business Intent (red) and **S**ell (green) is represented by a point. On the y-axis the average number of posts per day is shown.

are visualized and compared. The observations are discussed in the next section.

4 Results

The results of the applied method are presented in this chapter. At first, the feature selection process is reviewed, then final classification scores are calculated.

4.1 Feature selection

The selection of features is based on how useful they are for classification. This is evaluated using the Information Gain Attribute Evaluator in Weka. The results can be seen in Table 4.1. They are determined for each data set independently.

Results suggest that classification in the Twitter data set will work better than in the HolidayCheck data set because the features seem to be more meaningful. However, the detailed results of the classification algorithms are presented in the next section.

4.2 Classification results

The Weka tool from The University of Waikato is used to compute classification results using different algorithms.

In the first step the platforms Twitter and HolidayCheck were trained and tested independently (Table 4.2). As already estimated using the feature evaluator, the Twitter data set shows a high percentage of correctly classified instances. Although the HolidayCheck data set reports less accuracy, the percentage can be raised using more platform specific features.

Rank	Feature data set Twitter	Feature data set HolidayCheck
1	User posts per day	User posts per day
2	Text link count	_Suche
3	_Verkaufe	_Danke
4	_Biete	_Jemand
5	_Hier	_Du
6	_Suche	_Dir

Table 4.1: Twitter and HolidayCheck data sets are evaluated using the Information Gain Attribute Evaluator in Weka. The rank shows how important the information each feature is. The six most important features are listed for each data set. The features are presented in column two and three. Features starting with ”_” count word frequency in the text.

Data			Correctly Classified Instances			
Train	Test	Classes	Naive Bayes	Logistic Regression	J48	Random Classifier
Twitter 33 %	Twitter 67 %	B, S, N	91 %	**96 %**	95 %	44 %
Holiday 33 %	Holiday 67 %	B, S, N	66 %	**67 %**	47 %	35 %

Table 4.2: Classification on the Twitter data set in the car business and HolidayCheck (Holiday) data set in the tourism domain is done separately. 33 percent of instances are used for training, the rest is used for testing. The three classes used are: Buying Intent (B), Selling Intent (S) and No Business Intent (N). Results of the probability based Naive Bayes, Simple Logistic Regression and the decision tree based J48 algorithms are shown in columns four to six. The percentage values in those cells refer to the number of correctly classified instances. The best classification results is highlighted in bold. The random classifier uses only the distribution of the test data as information.

In the second step, testing and training data sources are from different platforms and domains. The results can be seen in Table 4.3. Because classification into three classes does not provide satisfying feedback, binary classification is also applied. Although there might be some improvements adjusting the parameters of algorithms, the outcomes already give a good overview about how well the classes can be distinguished. The class *Buying Intent* can be separated very well from all the other classes, whereas *Selling Intent* and *No Business Intent* have a lower percentage of correctly classified instances.

Now, that the quantitative results are presented first ideas may be derived. The findings and all its consequences will be discussed further in the discussion and conclusion chapter.

Data			Correctly Classified Instances			
Train	Test	Classes	Naive Bayes	Logistic Regression	J48	Random Classifier
Twitter	Holiday	B, S, N	35 %	**40 %**	31 %	35 %
Twitter	Holiday	B, S	46 %	**80 %**	60 %	53 %
Twitter	Holiday	B, N	61 %	62 %	**71 %**	51 %
Twitter	Holiday	S, N	43 %	**45 %**	45 %	51 %
Twitter	Holiday	I, N	34 %	**50 %**	40 %	56 %
Holiday	Twitter	B, S, N	**48 %**	47 %	47 %	44 %
Holiday	Twitter	B, S	89 %	**95 %**	93 %	77 %
Holiday	Twitter	B, N	87 %	**88 %**	88 %	79 %
Holiday	Twitter	S, N	**52 %**	43 %	50 %	50 %
Holiday	Twitter	I, N	48 %	**52 %**	51 %	50 %

Table 4.3: Classification uses the Twitter data set in the car business and HolidayCheck (Holiday) data set in the tourism domain.
The different classes are: Buying Intent (B), Selling Intent (S), No Business Intent (N) and Business Intent (I). The class I contains all instances of the classes B and S. If only two classes are mentioned in the third column, binary classification is conducted.
Results of the probability based Naive Bayes, Simple Logistic Regression and the decision tree based J48 algorithms are shown in columns four to six. The percentage values in those cells refer to the number of correctly classified instances. The best classification result is highlighted in bold. The random classifier uses only the distribution of the test data as information.

5 Discussion

Besides the results that are presented, there is also additional work which will be discussed in this chapter. Similarly, the existing limitations of the findings are also mentioned.

5.1 Lessons learned

Several different platforms have been compared besides Twitter and HolidayCheck. Unfortunately, smaller tourism forums like TripAdvisor and Bergfex do not have an open API and it takes more effort to crawl them efficiently.

The Twitter API has advantages over other social media APIs, because it is easier to use and information can be collected more efficiently. However, it still has some limitations when collecting vast amounts of data in a short amount of time. Unfortunately, services which worked in the past [1] like for example Twapperkeeper do not provide free use anymore.

In contrast to the Twitter API, the Facebook API is more complex to use. The biggest disadvantage compared to other services is that no search can be performed using the Facebook API. Therefore, it takes much more effort to find valuable content. The Facebook API in general is quite hard to use when crawling through sites to find information. However, the best way to get useful content is via open groups created to exchange information on a specific topic. With those groups it is possible to collect information about i.e. the tourism domain using the API in an efficient way.

One thing that should be improved is the reduction of labeling effort. It already takes a lot of time to label a small number of data instances, but when creating larger data sets to be able to make more significant statements, platforms like Amazon Turk [3] should be used to outsource work effort.

In the next section limitations regarding the results of this thesis will be discussed.

5.2 Limitations

Some aspects which always have to be taken into account when evaluating the results are mentioned in this section.

Firstly, because textual attributes are used for classification, it has to be said that these attributes are vulnerable to changes in the use of language. If the language shifts and different words would be used to express business intent, also algorithms have to be adopted.

Secondly, due to the fact that keyword search is used to filter data in the first place, the machine learning algorithms also highly rely on these keywords as it can be seen in the feature selection Table 4.1. This significantly influences the performance of the machine learning algorithms. Nevertheless, it does not limit the overall results because the keyword search filtering can be applied automatically.

Thirdly, this work is based on the language German, this means that results are restricted to German. However, it seems like the findings will at least be similar in other languages, especially in English.

Fourthly, this thesis is composed on a certain time span and therefore automated data selection via API and SOAP parsing is done on the months between November 2016 and April 2017. Special events which happened at that time of course influences the results. Users which were more active in that days affect the findings stronger than those which were inactive in that time period. In order to prevent that, additional information from the past up to ten years was added manually.

Finally, although features are selected as general as possible, it is not guaranteed that they can be conducted form every platform.

Now that all limitations were explained, the final conclusion can be drawn.

6 Conclusion

It is concluded by the results of this work that it is possible to automatically detect business intent on different platforms using machine learning. Findings from Table 4.2 state that this can be done with good accuracy when conducting training and test data from the same platform. However, when applying training and testing phase in different platforms and domains business intent detection becomes more difficult as shown in Table 4.3. Nevertheless, it is still possible to identify content which has buying intent through binary classification.

The results can be produced to a high degree automatically. The Twitter API and the SOAP parser build up the connection and retrieve data fully automatically. The filter through keyword search can be set using parameters. In addition, text correlated attributes are generated and a general language filter is applied to remove offensive content. Furthermore, separated training and test sets can be produced automatically avoiding duplicates in the data sets. Only the preparation for the classification process is done manually, however this step can also be automated easily. Eventually, it is even possible to link buyers and sellers without much effort via the Twitter API. It would take a bit more effort to achieve this using forums or other platforms. However, the linking is not implemented yet.

Different use cases can be generated where the fully automated business intent detection my be used to link buyers and sellers in a certain domain. Instead of querying search engines, people will be informed about existing buyers/sellers using the information they post during their daily life. Similarly to existing recommender systems users would see their ads based on the content their produce.

Fortunately, no content containing web spam was found although the probability of web spam in business intentioned content is very high [2]. Nevertheless, spam detection needs to be done before users are linked because this service loses a lot of trust if security problems exist.

Finally, it should be mentioned that the error rate during classification would have to be improved further. Certainly, no mistakes should be possible when actively linking people. Therefore, feature selection have to improved further in order to reach better accuracy.

The work which is still to be done in the future will be explained in the next chapter.

7 Future Work

At the end of this thesis a few ideas about what can be done in further steps are suggested.

Firstly, as mentioned in the previous chapter, classification results could be improved through additional features. Furthermore, the machine learning algorithms can deliver better accuracy when their configuration parameters are adjusted. As observed from the results, the largest steps of performance increase is achieved by adding new beneficial features.

Secondly, the linking of users through actively connecting them can be done in future. This step should be taken carefully, because many legal problems arise when deploying such an intrusive technology. However, it can be done in a secure test environment in order to prove functionality.

Thirdly, other languages can be used to receive similar results. Twitter API supports the use of language filters. Therefore, only the keywords have to be adjusted in order to collect content in different languages. With other platforms this step works similar. However, with forums it is a bit more difficult, because they often exists in only one language. Nevertheless, it should be possible to find a forum or similar platform for a given domain in the selected target language.

Additionally, it is possible to collect data without pre-filtering through keyword search. On the downside it has to be mentioned that random posts have a very low probability of containing business intent. Therefore, open access to all the data of a platform would be required in order to allow the algorithms to work efficiently.

Another study that could be conducted is to compare intent on a single platform, but domain independent. For example, Twitter is used to classify data from different domains like car business, real estate, fruit shopping, fitness products, etc.

Of course, additional platforms should be tested as well. Facebook was already used, but other social media services would also be interesting. Instagram has already over 400 million daily active users which also contains high potential business value. However, scanning Instagram for business intent is a far more difficult task because only little text based content is provided by the users. It is suggested, that this would need the introduction of visual algorithms as well in order to solve the task successfully.

Finally, another promising strategy is to only use text features to detect business intent. The advantage would be that no additional features are required. Therefore, this method could be used at every platform where the users post text content without limitations.

To summarize, there exists many ideas about what could be done in the future. The most important ones would be those which make classification even more accurate. However, the higher accuracy needs to be achieved without the loss of generality. After optimization, the algorithms still need to work for every platform and domain in order to be used for multipurpose operations without the need of major modification.

This final chapter suggested the most important ideas about what could be done in the future. It also concludes the work of this thesis.

Bibliography

[1] B. Hollerit, M. Kroell and M. Strohmaier. “Towards Linking Buyers and Sellers: Detecting Commercial Intent on Twitter.” In: *WWW '13 Companion Proceedings of the 22nd International Conference on World Wide Web*. 2013, pp. 629–632 (cit. on pp. 3, 5, 21).

[2] A. Benczur, I. Biro, K. Csalogany and T. Sarlos. “Web Spam Detection via Commercial Intent Analysis.” In: *AIRWeb '07 Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*. 2007, pp. 89–92 (cit. on pp. 6, 25).

[3] A. Ashkan and C. L. A. Clarke. “Term-Based Commercial Intent Analysis.” In: *SIGIR '09 Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009, pp. 800–801 (cit. on pp. 6, 22).

[4] A. Ashkan, C. L. A. Clarke, E. Agichtein and Q. Guo. “Classifying and Characterizing Query Intent.” In: *ECIR '09 Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*. 2009, pp. 578–586 (cit. on p. 6).

[5] H. K. Dai, Z. Nie, L. Wang, L. Zhao, J. R. Wen and Y. Li. “Detecting Online Commercial Intention (OCI).” In: *WWW '06 Proceedings of the 15th international conference on World Wide Web*. 2006, pp. 829–837 (cit. on p. 7).

[6] W. X. Zhao, Y. Guo, Y. He, H. Jiang, Y. Wu and X. Li. “We Know What You Want to Buy: A Demographic-based System for Product Recommendation On Microblogs.” In: *KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014, pp. 1935–1944 (cit. on p. 7).

[7] J. Wang, W. X. Zhao, H. Wei, H. Yan and X. Li. “Mining New Business Opportunities: Identifying Trend related Products by Leveraging Commercial Intents from Microblogs.” In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, pp. 1337–1347 (cit. on p. 7).

[8] W. Kasper and M. Vela. “Monitoring and Summarization of Hotel Reviews.” In: *Information and Communication Technologies in Tourism 2012: Proceedings of the International Conference in Helsingborg, Sweden*. 2012, pp. 25–27 (cit. on p. 7).

[9] J. Wang, G. Cong, W. X. Zhao and X. Li. “Mining User Intents in Twitter: A Semi-Supervised Approach to Inferring Intent Categories for Tweets.” In: *AAAI'15 Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015, pp. 318–324 (cit. on p. 7).

[10] R. Kern and D. Helic. “Knowledge Discovery Process.” In: *Lecture Preprocessing in Course University of Technology Graz: Knowledge Discovery and Data Mining 1*. 2015, p. 2 (cit. on pp. 9, 10).